The Dissertation Committee for Stephen P. Sebastian
certifies that this is the approved version of the following dissertation:

Pattern detection in natural images

Committee:

Wilson S. Geisler III, Supervisor

Alan Bovik

Lawrence K. Cormack

Mary Hayhoe

Eyal Seideman

Pattern detection in natural images

by

Stephen P. Sebastian, 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

December 2016

This work is dedicated to my parents, Garry and Sue Sebastian.

Acknowledgments

None of the work presented here would have been possible without the many people who helped me along my academic journey. First, I would like to thank Dr. Zygmunt Pizlo, whose lab I joined as an undergraduate at Purdue. Joining his lab changed the course of my life, and I would not have pursued a PhD without his encouragement and support. His enthusiasm for science was and still is infectious, and his emphasis that we must first and foremost love what we do has stuck with me. Thank you Dr. Pizlo (am I allowed to call you Zyg now?).

I am grateful to my lab mates in the Geisler lab at UT (both present and past) who made coming into the lab each day seem less like work and more like fun (in fact, it's amazing that any work got done at all). Thanks to Melchi, Anthony, Chris, Brian, Jeff, Yoon, and Calen. I owe a special thanks to Johannes Burge, who was a post-doc when I entered the lab, and who is now a professor at the University of Pennsylvania. My first task in the lab was to help Johannes calibrate a camera, which involved many hours spent together in a dark room. Johannes quickly became a collaborator, scientific mentor, and friend. He patiently taught me much of what I know about vision science, and I almost certainly would not have completed graduate school without his support.

One of the best things about working in academia is being able to meet so many wonderful people (one of the worst things is that they all eventually move away for post-doc positions). I am fortunate to have had so many friends at UT. Jake, you are a great friend who is always willing to help with anything (stat questions, moving, finishing a 2 gallon horchata). One day we will beat Sonic 2 together. Leor, your passion for everything in life makes being your friend an absolute joy (and thanks for introducing me to Sally). Thank you also to Ben, Kenneth, Lenny, Akram, Sophie, Milly, Memming, and Giacomo.

Thank you to my committee - Al Bovik, Larry Cormack, Mary Hayhoe, and Eyal Seideman - for your helpful feedback, suggestions, and support. Also, I am very grateful to my advisor, Bill Geisler, who taught me to think in a deeper, more scientific way. Over the years, I knocked on Bill's door unannounced hundreds of times. I honestly cannot recall a time when he was too busy to sit down with me to answer a question or look at a new result. Many of my "quick questions" turned into hour long discussions, and some turned into the work presented in this dissertation. Thank you Bill.

I owe a huge thank you to my family. Thank you to my parents, who always pushed me to be my best, and to pursue whatever I wanted in life. Thank you to my grandparents who, though their hard work and sacrifice made it possible for me to be able to pursue higher education. In particular, I owe a lot to Connie, who taught me that the most important thing in life is to be kind to other people, and to not stress too much. Finally, thank you to my partner in life, Sally. Your encouragement, love, and support helped get

me though graduate school, and for that I am forever grateful.

Pattern detection in natural images

Publication No. _____

Stephen P. Sebastian, Ph.D.
The University of Texas at Austin, 2016

Supervisor: Wilson S. Geisler III

A fundamental visual task is to detect target objects within a background scene. Using relatively simple stimuli, vision science has identified several major factors that affect detection thresholds, such as the luminance of the background, the contrast of the background, the spatial similarity of the background to the target, and uncertainty due to random variations in the properties of the background and in the amplitude of the target. Here I use a new experimental approach together with a theoretical analysis based on signal detection theory, to discover how these factors affect detection in natural scenes. First, I sorted a large collection of natural image backgrounds into multidimensional bins, where each bin corresponds to a narrow range of luminance, contrast and similarity. Detection thresholds were measured by randomly sampling a natural image background from a bin on each trial. In low-uncertainty conditions both the bin and the amplitude of the target were

blocked and in high-uncertainty conditions the bin and amplitude varied randomly on each trial. I found that thresholds increased approximately linearly along all three dimensions and that detection accuracy was unaffected by bin and amplitude uncertainty. The entire set of results was predicted from first principles by a normalized matched template detector, where the dynamic normalizing factor follows directly from the statistical properties of the natural backgrounds. This model assumed that the properties of the background underneath the target were constant across the image, but in natural images this is often not the case. Therefore, in a separate experiment, I measured detection thresholds on backgrounds where the contrast was modulated underneath the target. I found that varying the contrast underneath the target signal had a substantial effect on detectability, and that the pattern of results was predicted by an ideal observer that weighted its response based on an estimate of the local contrast (under the target). This suggests that the human visual system is able to use the varying properties of the background under the target in an near optimal way. Taken together, the results provide a new explanation for some classic laws of psychophysics and their underlying neural mechanisms.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In July of 1966, a group of student researchers at MIT were tasked with building, over the course of that summer, a computer system that could emulate the human visual system. In hindsight, this was quite an optimistic goal for a summer project, and the researchers were, of course, unsuccessful (Papert 1966). The notion that vision could be solved in a summer by students on 1960's era computers seems absurd today. However, the fact that it was once thought possible illustrates just how effortless our visual experience is. Our visual system works so well that we often take for granted the complexity of the diverse set of problems it solves. We are able to estimate the distance and relative size of the objects around us. We are able to track moving objects and recognize a friend's face in a crowd. We are able to guide the movements of our limbs to grasp objects in front of us. Many of the visual tasks we perform every day are so fast and effortless for us that they appear simple. However, even today's most advanced computer algorithms fail to match the performance and flexibility of the human visual system, and 50 years since the summer project at MIT, our understanding of the visual system is still far from complete.

One of the things that makes understanding visual processing so difficult is the variety and complexity of natural tasks and natural scenes. Computer vision systems are often tailored to meet the demands of a specific task and specific stimuli (for example recognizing product defects in a factory line). Computer algorithms designed for more general purposes often fail in unexpected ways (such as categorizing a school bus as an ostrich (Szegedy et al. 2013)). The most advanced deep neural nets, like Google's TensorFlow, are remarkably good at categorizing objects. These neural nets outperform humans at some tasks (Abadi et al. 2016), but fail to categorize objects in the face of even small occlusions (Chandler & Mingolla 2016), something that the human visual system is very good at doing.

Unlike most computer vision systems, the human visual system is able to perform a remarkable range of tasks quickly and often with little error. A major reason for this is because the visual system evolved to solve a certain set of tasks necessary for survival and to deal with all of the visual complexity and variety found in nature. Understanding human visual processing in the natural environment is important because it not only tells us about information processing in the brain, it can also guide models in computer vision applications.

Despite the importance of understanding how the visual system processes visual information in the natural environment, most of our current understanding of the visual system comes from behavioral and physiological experiments that used simple artificial stimuli displayed on simple backgrounds.

Visual receptive fields were first described in the frog retina by Hartline in 1938 using spots of light (Hartline 1938). Lettvin, Maturana & McCulloch (1959) showed the frog retina computes and sends behaviorally relevant information to the frog cortex. In that experiment they used as stimuli physical objects connected to a surface via magnets that were moved by hand. The orientation selectivity of neurons in cat (and later primate) primary visual cortex (V1) were discovered by Hubel and Wiesel in the 1960's using glass slides displayed on a projector screen (Hubel & Wiesel 1962, Hubel & Wiesel 1968). Human performance in many perceptual tasks such as crowding (Levi 2008), visual search (Eckstein 2011), and depth perception (Howard & Rogers 2008) have been measured and modeled using similar artificial stimuli.

Artificial stimuli offer many advantages over more natural stimuli. Their simplicity makes them easy to parameterize. Because of this, they can be manipulated along a single stimulus dimension (e.g. luminance, contrast, orientation, spatial frequency) to probe the behavioral and physiological changes that the given dimension evokes. This makes artificial stimuli very useful when designing controlled experiments (Rust & Movshon 2005, Felsen & Dan 2005).

There is little doubt that experiments with artificial stimuli have revealed a great deal about both the computations and neural mechanisms that underlie visional processing. Visual stimuli in the natural environment, however, vary in complex ways and are almost never simple patterns (Felsen & Dan 2005). Although simple stimuli are useful for isolating particular stimulus dimensions, it is difficult to know which stimulus dimensions to test without

an understanding of visual processing in the natural environment. A complete understanding of visual processing must ultimately account for the types of visual stimuli encountered in nature. This is true because pressure from natural selection has strongly driven the design and function of the visual system to maximize performance in natural tasks. Therefore, the design and function of any sensory system is adapted to the particular statistical properties of the stimuli it encounters in nature.

Some researchers have found a compromise between the usefulness of simple stimuli and the complexity of natural stimuli. For example, detection and visual search experiments have used stimuli that match the amplitude spectra of natural images (Bradley, Abrams & Geisler 2014, Najemnik & Geisler 2008). Freeman et. al. (2013) showed that stimuli designed to match several statistical properties of natural textures elicited responses in visual cortical area V2, but not in V1, suggesting that V2 is selective to the structure found in natural scenes. Still, even these naturalistic stimuli lack the full complexity of real natural images. Because of this, these types of experiments may fail to capture important aspects of visual processing.

With the advantages of using natural stimuli also come challenges. First, natural image statistics are difficult to measure. One of the earliest studies of natural image statistics was conducted by Egon Brunswik in 1953. Brunswik painstakingly identified contours by hand on several photographs. These photographs were stills from the 1949 film *Kind Hearts and Coronets* taken from an issue of *Life* magazine. He then used these contours to show

that the Gestault rule for proximity grouping is predicted from natural images (contours that belong to the same object tend to be closer together than contours that belong to different objects). The fact that Brunswik used photographs from a magazine, and labeled the contours by hand illustrates just how difficult it once was to not only measure the scene statistics, but also to collect appropriate images. Thankfully, recent developments in computer technology and digital photography have made measuring natural image statistics much easier. There now exists several large databases online that have many thousands of calibrated natural images. Figure 1.1 shows an example image from Brunswik (1953) and an image from a modern natural image database (McCann 2015).

Even though natural scene statistics are more easily measured now than in the past, the relevant statistical properties of natural images are unknown for many perceptual tasks (Rust & Movshon 2005). Several studies have identified relevant stimulus dimensions for a number of natural tasks, such as defocus blur and disparity estimation (Burge & Geisler 2011, Burge & Geisler 2014), and foreground/background identification (Geisler, Najemnik & Ing 2009). Still, these approaches are limited in the types of stimulus features they can extract.

Recent technological advancements in eye tracking and virtual reality technology have made it possible to collect eye movement statistics while human subjects perform real world tasks. Hayhoe, Shrivastava, Mruczek & Pelz (2003) used head mounted eye trackers to measure eye movements and study

Figure 1.1: Examples of natural images used in scientific research. A. A sample image from Brunswik (1953). Images were black and white movie stills taken from *Life* magazine. In total seven images were used. B. An example image from a database of luminance and range images. Images were taken with a calibrated high resolution camera, and a laser range finder. Each image is a stereo pair, with corresponding depth information. In total there are 197 such images in the database.

motor planning in several tasks, including making a peanut butter and jelly sandwich. Sprague et. al. (2016) used a head mounted eye tracker along with a set of binocular cameras to measure the distribution of defocus blur encountered during a variety of tasks as subjects went about their day. Despite these advances, statistical properties of natural scenes that are important for

particular tasks are difficult to measure, because it is difficult to control these factors in a natural setting.

There is a large body of work measuring the statistical structure of natural scenes (for reviews see Simoncelli & Olshausen (2001) and Geisler (2008)). This work is directed at characterizing the general properties of natural stimuli (such as color (Dixon 1978), spatial structure (Olshausen & Field 1997), or depth (Yang & Purves 2003), and not the specific properties relevant to specific tasks.

The work presented in the following chapters focuses on one particular visual task: detection of luminance patterns. Detection is a critical bottleneck for almost every visual task, and there has been much work in the history of vision science in understanding the factors that affect detection. Specifically, the detectability of a target pattern added to a background decreases with increasing background luminance (König & Brodhum 1889, Hecht 1924, Mueller 1951, Cornsweet & Teller 1965), background contrast (Legge & Foley 1980, Burgess, Wagner, Jennings & Barlow 1981, Lu & Dosher 1999, Najemnik & Geisler 2008) and with the structural similarity of the background to the target (Campbell & Kulikowski 1966, Stromeyer & Julesz 1972).

The statistical properties of the area surrounding the target are also known to have an effect on detectability through surround inhibition mechanisms (Polat & Sagi 1993, Field, Hayes & Hess 1993, Xing & Heeger 2000). Uncertainty about certain target properties, such as the location and type of target, can also affect detectability. More recent studies suggest other back-

ground factors (such as the presence of edges verses texture) on detection in more naturalistic backgrounds (Chandler, Gaubatz & Hemami 2009, Alam, Vilankar, Field & Chandler 2014). In the next chapter, I will describe many of these basic findings in more detail.

One of the major goals of science is to measure lawful relationships in nature. In perceptual science, the goal is to measure lawful relationships between the external stimulus (rays of light entering the eye) and behavioral performance and neural computations. For detecting spatial targets, many of these relationships have been measured using simple stimuli and backgrounds. The ultimate goal of vision science, however, is to understand visual processing in natural images. Despite this, there has not been a systematic study of detection in natural image backgrounds. One reason that this has not been studied is that natural images are difficult to use in a controlled experimental setting. In chapter 3, I will describe a general framework (called constrained scene sampling) for testing hypothesis using natural image stimuli. Here, I apply this framework to the task of detection, but it could potentially be used for many other kinds of visual tasks.

In Chapters 4 and 5, I will describe a series of detection experiments in human observers using this new framework and natural image stimuli. First, I measure the effect of luminance, contrast, and structural similarity of the background to the target in natural image backgrounds. I show that the basic findings measured in simple backgrounds hold in complex natural backgrounds. Furthermore, I show that the entire set of results is predicted with a simple

matched template observer derived directly from the statistical properties of natural image backgrounds.

In the natural environment, the values along a number of local stimulus dimensions vary by an order of magnitude within a scene, and the intensity of the target an observer is trying to detect varies from one occasion to the next. Because of this, the visual system needs to handle large variations in background properties and target intensity quickly and efficiently. In Chapter 5, I describe the results of an experiment designed to test the effect of real world background and target uncertainty on performance. I show that human observers are remarkably unaffected by such large changes in the background statistics. This result is predicted by an ideal observer that normalizes the template response by an estimate of the local luminance, contrast, and structural similarity of the background to the target. This result suggests that the reason there exists rapid gain control mechanisms in the early visual system is to deal with the high degree of variability in the natural environment.

The experiments in Chapters 4 and 5 measure the effect of many of the major factors known to affect detectability. However, it is likely that there are other factors that were not controlled for. One such factor is variation in the distribution of contrast, luminance, and similarity under the target itself. In Chapter 6, I will describe a series of detection experiments in noise backgrounds designed to test how human observers detect targets where the statistical properties of the background change underneath the target (as they often do in natural images). This study reveals a new type of visual masking

that I call sub-masking. I argue that the results of this study should improve models of detection in natural images.

As mentioned earlier, much of what we know about how the visual system processes visual information is based on studies with simple artificial stimuli. The visual system evolved in order to maximize our chances of survival and reproduction in nature, with natural images. My dissertation work seeks to bridge the gap in our understanding between the lab and the natural world in which we live, for the basic task of detection. Although the work presented here does not achieve the lofty goals of the MIT summer project in the 1960's, it does bring us one step closer towards an understanding of how we see in the natural world around us.

Chapter 2

Detection: a fundamental visual task

2.1 Background

Our ability to identify spatial patterns in a cluttered environment is essential to almost every natural task. Because of this, detection tasks have been studied extensively, and several factors have been shown to effect our ability to detect spatial patterns. I will describe these factors in detail later in this chapter, however, it is important to first define exactly what is meant by detection. Detecting a visual object means identifying a particular pattern of light as that specific object and ruling out all other patterns of light as not that object (Geisler 2003). Detection is a special case of the task of identification, where the observer must identify a spatial pattern as belonging to one of a number of different categories. In detection, there are only two categories: the object or target of interest (the signal), and everything else (the noise). My work is concerned with detecting a known spatial pattern in natural backgrounds; in other words, identifying a specific object among all of the other objects and features found in the background.

Detection is often thought of as a primitive or low level visual process. Indeed, species with vision systems first emerged 530 million years ago, and

were simple luminance detectors, able to detect the direction and intensity of a luminance signal (Land & Nilsson 2012). These early visual systems lacked a lens or even a pupil, and were nothing more than a collection of photosensitive cells. Similar visual systems exist today in a species called planaria (a type of flatworm). Planaria use their visual system to detect the general direction of light, and they use this information to guide their behavior (Inoue 2015). In this sense, these primitive visual systems are like the human visual system because the main goal of any visual system is to use visual information to inform and guide behavior. Although the visual information we use and the tasks we perform are complex, we use visual information each day to guide behavior.

Despite being thought of as a basic low-level process, detection is essential in many high-level tasks. For example, to spot our friend in a crowd we must detect a specific face as belonging to a specific person, and rule out all other faces in the crowd. This process happens so quickly and effortlessly that we hardly notice it happening at all. Detecting the object of interest and ruling out all others is, however, a very complicated task in nature because of all the variability and visual clutter found in the natural environment. If we were unable to perform the task of detection quickly, we would not survive for very long.

Natural vision systems are tuned - through experience and evolution - to the tasks an organism performs and the stimuli it uses to guide behavior. For simple organisms, like the planeria, it is clear what information their visual

system uses to guide behavior because the only thing their visual system can sense is luminance in a general direction. Furthermore, it is clear that the ability of the planeria to detect a particular luminance signal is of course limited both by noise in the environment (for example photon noise), and noise in the internal neural signal that carries the information.

Unlike the planeria, which are sensitive only to luminance direction, the human visual system is sensitive to many visual features, including luminance, contrast, and spatial frequency and orientation. But what information is the human visual system using to detect objects of interest? And what factors make it easier or more difficult to detect a particular object? Understanding the factors that limit our ability to detect a target in space is fundamental to understanding visual processing. These factors can be broadly classified into two categories: internal and external.

2.2 External factors

We live in a world where many of the stimuli we encounter are highly variable in their statistical properties (Mante, Frazor, Bonin, Geisler & Carandini 2005). Because of this, a major factor that affects our ability to detect an object is the variability in the external stimuli themselves. To better understand this point, consider a simple observer that is shown a pattern of light intensity I, and is tasked with identifying whether this pattern contains some target signal (S), or just a background alone (N). In a completely noiseless world, the image containing the signal would always be the same on every presenta-

tion. Thus, an observer could be correct 100 percent of the time by simply subtracting the known background (N) from the input image (I); if the result is zero at every location then stimulus is background alone, otherwise it contains the target signal. Of course, in the real world the background can different on every presentation..

Signal detection theory provides a statistical framework for observers detecting patterns in noise (Green & Swets 1966). In a world where the stimuli vary on each presentation, the range of possible input images is no longer deterministic, but rather probabilistic. In this case, an optimal observer would compare the probability of the signal S given the input image I ($p(S|I)$) to the probability of the background alone, N, given the input image I ($p(N|I)$). The observer should assign the input image to the category with the highest probability. Using Bayes' Theorem, it can be shown that the optimal decision rule is to compare the ratio of the likelihoods to a criterion based on the prior category probabilities:

$$Response = \begin{cases} present, & \text{if } \frac{p(I|S)}{p(I|N)} \geq \frac{p(N)}{p(S)} \\ absent, & \text{if } \frac{p(I|S)}{p(I|N)} < \frac{p(N)}{p(S)} \end{cases}$$

where $p(I|S)$ is the likelihood of the image given the signal is present, $p(I|N)$ is the likelihood of the image given the signal is absent, and $p(S)$ and $p(N)$ are the prior probabilities of signal present and signal absent, respectively. Most experiments set the rate of signal and no signal trials to be equal, which makes the decision criterion be equal to 1.

Figure 2.1 shows an illustration of this decision process. Each distribu-

tion represents the probability of likelihood ratios (on the x axis) given that the input image belongs to either the noise category (shown in purple) or the signal plus noise category (shown in green). It is clear from this figure that choosing the distribution with the higher probability is equivalent to comparing an observed decision variable (likelihood ratio) with a fixed criterion (shown in black).

The detectability of a particular signal, S, can be represented as the signal to noise ratio, d-prime:

$$d' = \frac{|\mu(I|N) - \mu(I|S)|}{\sqrt{\frac{\sigma^2(I|N)+\sigma^2(I|S)}{2}}}$$

A higher d' corresponds to a higher detectability (greater accuracy). The detectability of a signal can increase either by increasing the signal strength (and therefore the mean of the signal distribution, as shown in Figure 3.3A) or by decreasing the noise (the variance of the likelihood distributions, as shown in Figure 3.3B). All else being equal, increasing noise will always result in lower detectability. Therefore, any noise present due to the variability of the signals from the natural environment fundamentally limits our ability to detect specific spatial patterns. A threshold for detection is defined as the signal strength needed for d' to reach a particular value. Many of the experiments I describe in this Chapter (and the experiments in the following chapters) set the threshold value of d' to be 1.

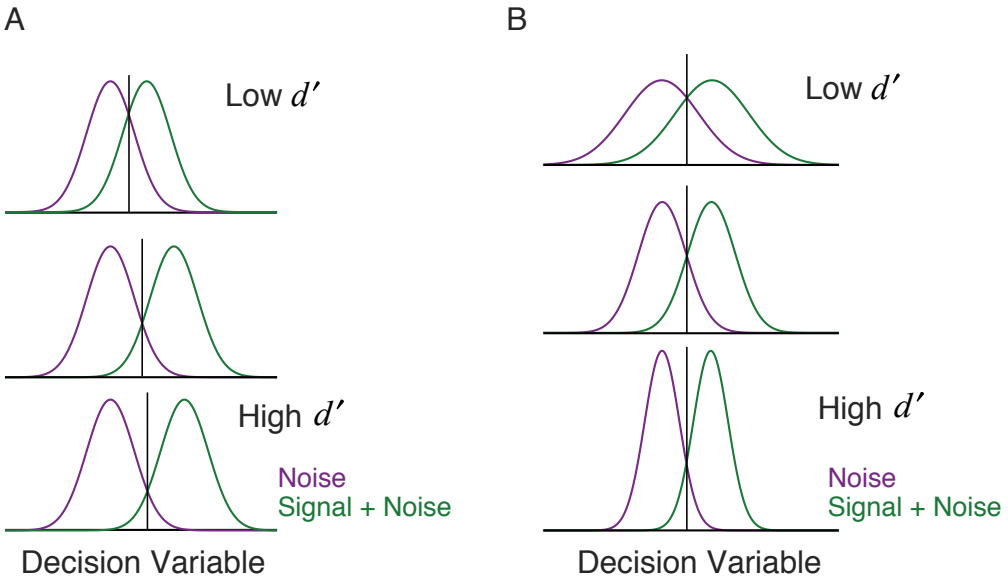

Figure 2.1: Effect of increasing signal and increasing noise on detectability. A. Increasing signal increases detectability. Each curve is the likelihood distribution of an input image coming from either the noise (purple) or signal (green) categories. The variance of the likelihoods represents the external variability in the input images. The optimal decision rule is to compare the input image to a fixed criterion (shown in black). The d-prime increases down the rows because the signal strength increases (and therefore the mean of the signal likelihood distributions increase). B. Decreasing noise increases detectability. Again, each curve is the likelihood distribution of signal given the image or noise given the image. The d-prime increases down the rows because the noise is reduced (and therefore the standard deviation of the likelihood distributions decrease).

Perhaps the most basic source of external variability found in the environment is due to the photon nature of light itself. The number of photons emitted by a light source and absorbed by the retina follows a Poisson proba-

bility distribution:

$$poiss(k|z) = \frac{z^k e^{-z}}{k!}$$

where k is the number of photons absorbed by the retina on a given presentation, and z is the mean number of photons absorbed by the retina given a particular light level.

As stated earlier, one of the most simple types of detection is detecting any change in luminance. Suppose an observer limited only by photon noise is tasked with detecting a change in luminance, Δz from a background luminance z_B. On each trial, the observer is shown randomly either z_B alone, or $z_B + \Delta z$. An ideal observer would compare the number of photons absorbed on a particular trial presentation (k) to the probability distributions $poiss(k|z_B)$ and $poiss(k|z_B+\Delta z)$ and choose the category corresponding to the distribution with the higher probability (greater likelihood ratio).

Each distribution in Figure 2.2 is the likelihood a particular number of photons, (k) given background alone (z_b), or the background plus an increment light ($z_b+\Delta z$). Just as before, the optimal decision rule is to choose the higher likelihood (respond signal present if the likelihood ratio is greater than 1), given the probability of background and background plus increment are equal). As the magnitude of the luminance increment increases, d-prime also increases.

An observer limited only by photon noise (in other words an observer that count the number of photons absorbed on every trial to make a judgement)

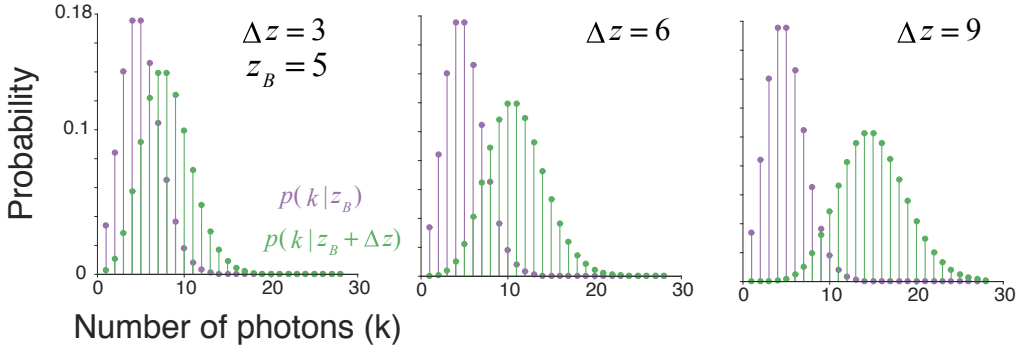

Figure 2.2: Detection limited by photon noise. Each set of points is the likelihood of a given category (background along or background plus increment) given the stimulus observed. The photon increment and detectability both increase across rows. An ideal observer chooses the category with the highest likelihood by comparing the observed number of photons to a criterion.

has increment thresholds that follow a square-root relationship:

$$\Delta z \propto \sqrt{z_B}$$

where z_B is the mean number of photons in the background only condition, and Δz is the increment threshold . This square-root relationship is sometimes referred to as the DeVries-Rose law from de Vries (1943) and Rose (1948) who both described the law.

Experiments with human observers have shown this square-root relationship to hold under low light conditions in near total darkness (Barlow 1957, Blakemore & Rushton 1965). This suggests that (under certain conditions) the human visual system is limited in its ability to detect changes in light by the noisy nature of light itself. Later, I will show that this square-root

relationship does not hold in conditions with higher light levels, a result that suggests limits in the internal processing by the human visual system.

In addition to the noisy nature of light itself, stimuli in the natural environment vary in their structure and statistical properties. The effect of this variability has been studied experimentally using simple noise patterns, where on each trial a different noise pattern is presented. A classic noise pattern used in detection experiments is Gaussian white noise, where the luminance at every pixel is sampled from a Gaussian distribution with mean μ and standard deviation σ. The mean luminance of the noise pattern increases with the mean of the Gaussian, and the contrast of the noise pattern increases with increasing standard deviation.

Consider an observer that is shown a particular image that either contains the noise pattern alone ($N(x, y)$) or the noise pattern plus some target signal with a particular amplitude ($N(x, y)+aT(x, y)$). If the noise is Gaussian white noise, it can be shown that an ideal observer computes a template match between the input image and the target template: $Resp = \sum T(x, y)I(x, y)$, and compares this response to a fixed criterion ($\gamma = \frac{aT \cdot T}{2}$). As the variance of the noise background increases, the threshold contrast squared for the target increases proportionally:

$$C_t^2 \propto \sigma^2$$

Written another way, contrast threshold power rises in linear fashion with

background contrast power.

$$C_t^2 = k_c C_b^2 + C_0^2$$

where C_t^2 is the contrast power at threshold, k_c is a constant, C_b^2 is the contrast power of the background noise, and C_0^2 is the baseline contrast power (threshold in a zero contrast background). For a true ideal observer the baseline contrast would be the power of the photon noise.

This relationship has been found to hold across several target and noise types, although the baseline contrast power parameter is much higher than predicted from photon noise (Burgess et al. 1981, Legge, Kersten & Burgess 1987, Lu & Dosher 1999). Lu and Doshier (1999) measured contrast thresholds for an oriented Gabor target in random Gaussian white noise. On each trial a noise frame was presented for 16.6 ms before and after a Gabor of a fixed spatial frequency and orientation was presented in for 8.3 ms. Subjects responded with the orientation of the Gabor. Contrast thresholds are initially flat across a low background contrast range, then increase with increasing background contrast, in agreement with the equation above (Figure 2.3B).

Legge et al. (1987) measured contrast increment thresholds under two conditions: a 2 cycle per degree sine wave embedded in dynamic white noise, and a small luminance disk embedded in static Gaussian white noise. In both cases threshold contrast power was found to rise in proportion to background contrast power. The results from one subject in the static noise experiment are shown in Figure 2.3B. This relationship also holds across the visual field

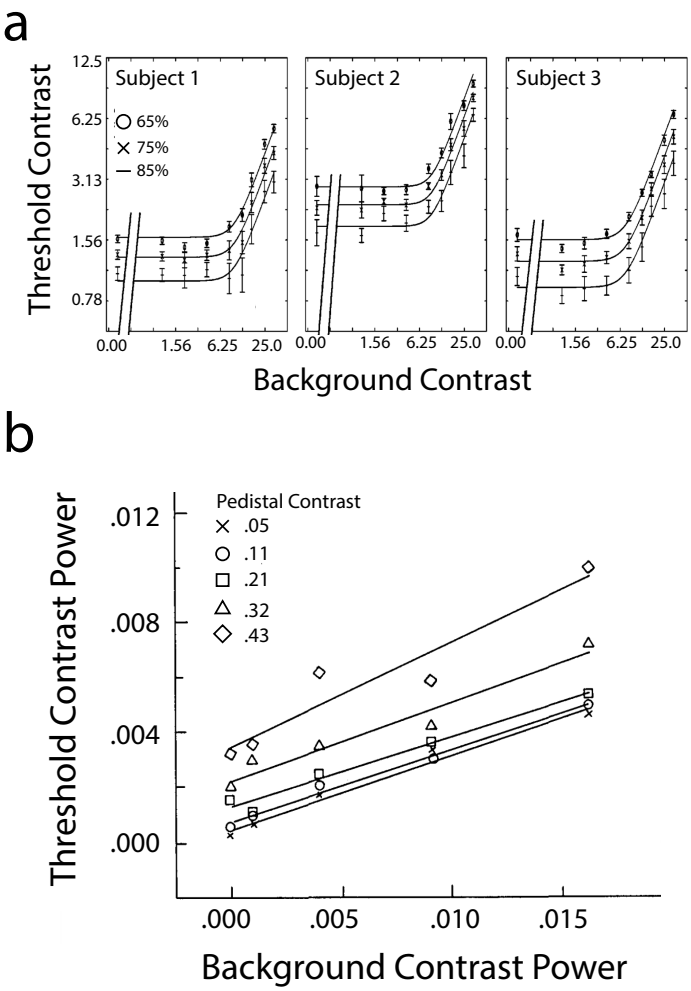

Figure 2.3: Effect of background contrast on detectability. A. Threshold contrast plotted against background contrast for three subjects in Lu and Dosher (1999). B. Threshold contrast power vs. background contrast power for one subject in Legge, Kersten and Burgess (1987). In both cases contrast power at threshold rises in proportion to background contrast power.

with different types of noise backgrounds (Najemnik & Geisler 2008, Bradley et al. 2014).

The findings I've described up to his point have all measured thresholds using a fixed intensity target at a fixed location in the visual field. Under natural conditions, however, the signal strength and location of the target typically vary on every occasion. Furthermore, in natural viewing, the target of interest typically appears against a different background scene on every occasion (this is also true for noise backgrounds, but the backgrounds are not as variable as under natural conditions).

Because of this, it is possible that the signal strength needed to detect a particular target may be higher under natural viewing conditions. For simple stimuli, it is known that the uncertainty created by the random amplitude and location of the target (target uncertainty), and the random variation in the properties of the background (background uncertainty) are additional factors that may reduce detection performance (Pelli 1985, Eckstein, Ahumada & Watson 1997, Michel & Geisler 2011).

2.3 Internal factors

Besides the external variability in the stimuli themselves, we are also limited in what we can detect by internal factors. The first and perhaps most obvious factor is that any vision system is limited in what it can detect by the sensory organ itself. In the human visual system, this sensory organ is the eye. Visual processing begins when light is refracted by the cornea and focused by the lens onto the retina. The modulation transfer function the lens system in the human eye drastically reduces power in spatial frequencies above 10 cycles

per degree (Patel 1966, Van Nes & Bouman 1967, van Meeteren 1974). Even under the best conditions (a small pupil size) spatial frequencies above 60 cpd cannot be imaged on our retina with our lens system. [1]

In the natural environment luminance varies by many orders of magnitude across all scenes, and by typically by one or two orders of magnitude within a given scene (Hood 1998, Mante et al. 2005). The human visual system evolved several mechanisms for dealing with such massive variations in overall luminance. First, we have a pupil that controls the amount of light entering the eye by adjusting the physical opening to the eye, much like an aperture in a camera system. The pupil dilates or contracts based on several factors (including vergence and accommodation (Campbell & Westheimer 1960, Alpern, Mason & Jardinico 1961), but it responses most strongly to luminance changes in the environment. Pupil size ranges from around 2 mm on a bright day, to 8 mm in darkness. This change accounts for about a factor of 16 change in the amount of light that enters the eye, which clearly is not enough to cover the huge dynamic range needed in natural viewing.

In addition to the pupil, the visual system also has two types of photosensors: rods (for scotopic or low luminance conditions) and cones (for photopic or high luminance conditions). Switching between the rod and cone system can adjust sensitivity by several orders of magnitude. This switching does not happen quickly, and even within each system there is a slow adap-

[1] Interestingly, the Nyquist frequency of the central fovea is around 60 cycles per degree (Hirsch & Curcio 1989), a likely consequence of the lens system and retina evolving together.

tation to the overall light level. Data from Hecht, Haig & Chase (1937) are shown in Figure 2.4. Subjects first adapted to a uniform light field, and then absolute thresholds for light detection were measured over time. Thresholds fell steeply at first, then slowed. After a period of several minutes, the visual system switched to using the scotopic system, and thresholds fell again. After a period of 30 to 40 minutes, thresholds leveled out and detectability did not continue to improve. Note that the sensitivity of an observer that was only limited by the external photon variability of light would switch instantly with changes in background luminance.

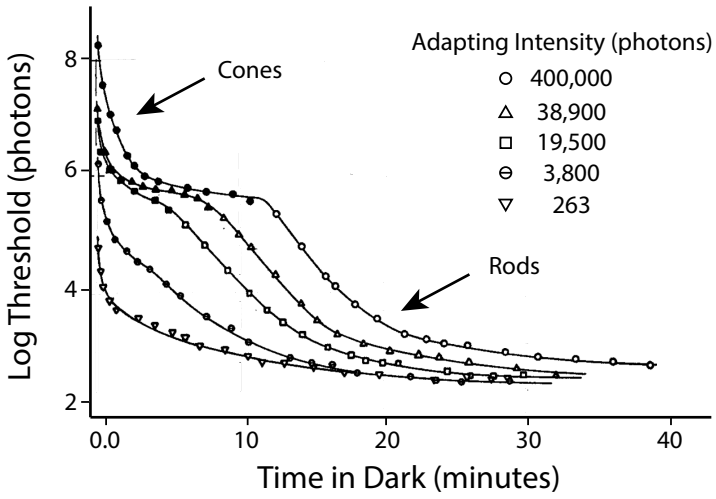

Figure 2.4: Time course of dark adaptation for different adapting intensities (figure adapted from Hect et. al. 1937). Subjects first adapted to a uniform light field, then absolute thresholds for light detection were measured over time. Thresholds initially fell steeply, then bowed as the system switched from cone to rod vision.

The specific properties and density of the photoreceptors across the

visual field also sets limits on detection performance. For example, threshold changes as a function of the wavelength (Hecht & Hsia 1945, Pokorny, Graham & Lanson 1968) due to the spectral sensitivity of the photoreceptors. The location of the target in the visual field determines how easily it is detected, because the density of photoreceptors and ganglion cells falls off rapidly away from the fovea (Curcio, Sloan, Kalina & Hendrickson 1990, Dacey 1993, Drasdo, Millican, Katholi & Curcio 2007, Watson 2014, Bradley et al. 2014).

Many detection experiments use a fixed stimulus, meaning trials where the signal is present are always exactly the same, and trials where the signal is absent are always exactly the same (no external variability except for photon noise). An ideal observer would (except for the weak effects of photon noise) produce the same response on each trial under such conditions. However, detection by human observers is much worse than predicted by photon noise alone. Some of this is due to the noise that is internal to the nervous system. One component of this internal noise is the variability of the underlying neural representation of the stimuli, due to the fact that individual neurons in the visual system do not respond deterministically to the same stimulus (Tolhurst, Movshon & Dean 1983). Other components of internal noise include decision noise (Beck, Ma, Pitkow, Latham & Pouget 2012), motor response noise, and the variability in the attentional state of the observer.

In the signal detection framework, adding a fixed internal noise is equivalent to adding a fixed variance to likelihood distributions. After adding internal noise, The standard deviations of the signal and signal + noise distributions

become:

$$\sigma_t = \sqrt{\sigma_e^2 + \sigma_i^2}$$

where σ_t is the total standard deviation, σ_e is the noise due to external factors, and σ_i is noise due to internal factors. As noted before, all else being equal, increasing the standard deviation of the likelihood distributions always decreases the detectability of the signal. Therefore, any noise internal to the observer will raise thresholds for detection.

Detection experiments using simple backgrounds have also revealed other additional processing limitations in the visual system. A classic finding in vision science is that the luminance needed to detect a target (ΔL) increases in proportion with the background luminance (L_b):

$$\Delta L \propto L_b$$

This relationship is known as Weber's Law, after Ernst Weber, a German physician who first described the law.

Weber's Law for luminance was first measured scientifically in the 16th century by a French scientist named Pierre Bourguer (Bouguer 1760, Robson 1993) using candles and a wooden rod. Studies in the 20th century that used more sophisticated apparatus have confirmed and extended these earlier findings. Data from Cornsweet and Teller (1965), are plotted in Figure 2.5A. Here a spot of light was displayed on a background disk, and increment thresholds were measured as a function of the disk luminance. After an initial flat period (in which internal noise and photon noise dominates), thresholds rise linearly

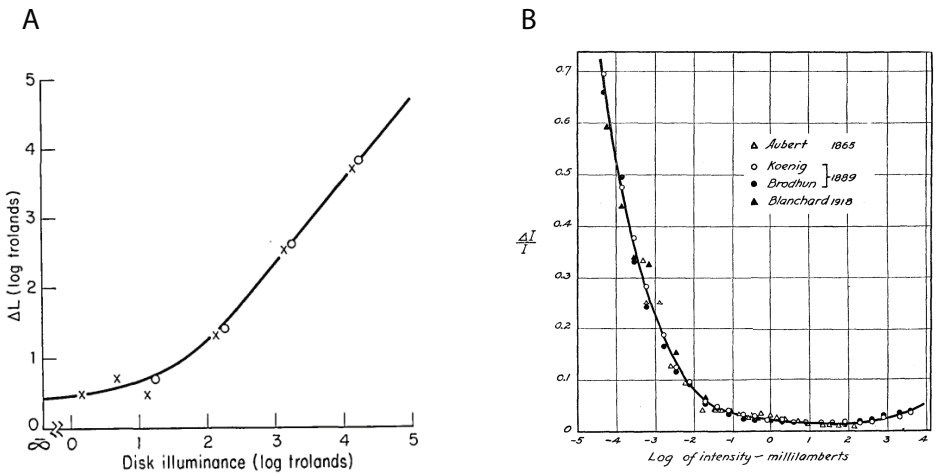

Figure 2.5: The effect of increasing background luminance on detection thresholds. Luminance thresholds follow a Weber's law relationship. A. Data from two subjects (x and o) in Cornsweet and Teller (1965). After an initial flat period, thresholds rise according to Weber's Law. Background illumination and thresholds are plotted as log trolands. B. Weber relationship demonstrated in a number of studies (from Hect 1924). The figure shows the Weber fraction (increment threshold over background illumination) plotted against the background illumination (in log millilamberts). Over a large portion of the range shown, the Weber fraction remains constant (Weber's Law).

with background luminance. This finding has been confirmed many times over (König & Brodhum 1889, Hecht 1924, Mueller 1951).

Weber's law relationships are also shown in Figure 2.5B. Here, the data from three studies (Aubert 1865, König & Brodhum 1889, Blanchard 1918) spanning almost 50 years are plotted (Hecht 1924). The figure plots the Weber fraction (increment threshold over background illumination) as a function of the background illumination. One can see that $\frac{\Delta L}{L_b}$ is constant over quite a

large range. It is important to note here that the Weber fraction for luminance ($\frac{\Delta L}{L_b}$) is a measure of contrast. Therefore, although the luminance needed to detect a target increases linearly with background luminance, the threshold contrast remains fixed.

A photon limited ideal observer predicts a square root relationship between the background luminance and the increment luminance threshold. While this law does hold in some conditions, the linear relationship of Weber's Law holds under most conditions and shows that internal factors other than photon noise strongly limit detection performance. Indeed, Weber's law for luminance is consistent with luminance gain control during the early stages of visual processing (for review see (Hood 1998)).

In the previous section, I showed that a matched template observer is optimal in Gaussian white noise backgrounds, and predicts a linear relationship between the background contrast power and threshold contrast power. However, several studies have also showed similar power law effects of background contrast for backgrounds that do not change from trial-to-trial. Legge and Foley (1987) measured contrast increment thresholds for sine wave targets added to backgrounds of sine waves with fixed contrast. The target sine wave was added to the a background sine wave with fixed properties that did not change from trial to trial (in other words there was no external variability in the stimulus). An ideal detector would be limited only by photon noise and have very low thresholds, and would have a constant threshold independent of background contrast. However, Legge and Foley observed a power

law relationship between background sine wave contrast and threshold contrast. This suboptimal performance can be explained by contrast gain control mechanisms in the early visual system. These mechanisms are similar to the luminance gain control mechanisms mentioned earlier, and are thought to exist in order to keep responses within the dynamic range of the neurons(Albrecht & Geisler 1991, Heeger 1991, Hood 1998, Carandini & Heeger 2012).

The sine wave grating is a classic and well studied target stimulus in vision science. Sensitivity ($1/c_t$) to sine wave grating was first measured by E.W.H Selwyn, a scientist at Kodak (Selwyn 1948, Robson 1993). Selwyn used photographs of sine waves of various spatial frequencies, and an optical system to vary their contrasts to measure sensitivity of human observers, and found that the peak sensitivity was around 4 or 5 cycles per degree. Selwyn attributed the falloff in sensitivity at higher spatial frequencies to inadequacies in the photographic process which became more obvious as spatial frequency increased. However, later measurements of human contrast sensitivity using sine wave gratings displayed on oscilloscopes also showed that sensitivity to sine wave stimuli peaks around 5 cycles per degree and falls off rapidly above that frequency (Figure 2.6A).

The contrast sensitivity function is thought to be the envelope of many underlying neural "channels", each sensitive to a separate narrow band of spatial frequencies and orientations. Figure 2.6B shows the results of an adaptation experiment designed to test whether such channels exist (Blakemore & Campbell 1969). First, the contrast sensitivity function was measured for

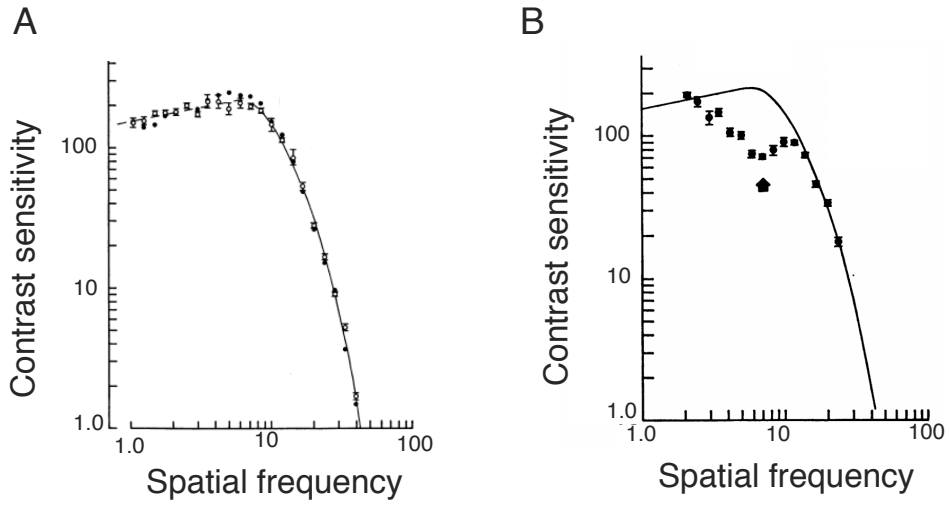

Figure 2.6: A. Contrast thresholds measured as a function of spatial frequency (From Blakemore and Campbell 1969). Contrast decreases along the y-axis, and spatial frequency (in cycles per degree) increases along the x-axis. Sensitivity is highest around some intermediate spatial frequency (4-6 cycles per degree). B. Contrast sensitivity function measured before (solid line) and after (points) adapting to a 10 cpd grating (Blakemore and Campbell 1969).

a human observer (shown as the black line). Next, the observer adapted to a grating of 10 cycles per degree. Finally, the contrast sensitively function was remeasured after adaptation (show as the black points). After adaptation, there is a notch or dent in the contrast sensitivity function, meaning that gratings at those locations needed more contrast to be detected. This dent is centered around the adapting frequency of 10 cycles per degree, but has also spread to other spatial frequencies that were not adapted, which is consistent with visual information being processed in separate but overlapping spatial-frequency channels.

Figure 2.7: Example of similarity masking in the form of camouflage. Right: an owl sits against a tree. Left: The owl is revealed when he opens his eye. The owl is difficult to detect because it evolved to match the structure of its surrounding. (Photographs by Graham McGeorge, http://www.grahammcgeorge.com/)

One consequence of processing visual information in overlapping spatial frequency and orientation channels is that the particular spatial structure of the background has a strong effect on detectability. Specifically, the extent to which the background structure matches the structural properties of the target determines the amount of masking. Many organisms take advantage of this effect in the form of camouflage. There is strong evolutionary pressure on species to match the structure of their surroundings in order to make themselves more difficult to detect by either predators or prey. Figure 2.7 shows an example of camouflage (and similarity masking) in nature.

There are several ways in which a target can match the structure of

the background, but here I will focus on three: orientation, spatial frequency, and phase. Campbell & Kulikowski (1966) measured the effect of changing the relative orientation between a target object and the background. Sine wave grating stimuli were generated on two oscilloscopes and superimposed optically. The orientation of the target sine wave remained fixed vertically. The orientation of the background sine wave was varied, and detection thresholds were measured by adjusting the contrast of the target.

When the orientation of the background was the same as the target, contrast thresholds rose proportionally with the contrast of the background in agreement to the contrast background findings I mentioned earlier. However, when the angle between the background and the target gratings was increased, thresholds fell exponentially. At a separation of 12 degrees, thresholds were reduced by a factor of 2, independent of the background contrast. This finding demonstrates a strong effect of background orientation similarity on detectability.

Stromeyer & Julesz (1972) studied the effect of spatial frequency similarity on detection thresholds. Vertical sine wave grating target stimuli were embedded in spatial frequency filtered vertical stripe noise. The stripe noise varied across time with a frequency of 60 Hz. The contrast of the sine wave grating was varied to measure thresholds in two conditions. First, the frequency of the target grating was held constant and the frequency of the noise was varied. Next, the frequency of the grating was varied and the frequency of the noise was held constant.

Thresholds decreased by half when the noise cutoff frequency was 0.5 octaves away from the target stimulus in the first condition. Similarly, in the second condition, thresholds again decreased by half when the target frequency was 0.5 octaves away from the noise pattern. Taken together, these results show that as the spatial frequency band of the noise approaches the spatial frequency of the target, the target becomes more difficult to detect. Wilson, McFarlane & Phillips (1983) found a similar masking effect of spatial frequency using fixed sine wave gratings as backgrounds.

The effect of background phase on detection thresholds was studied by Foley & Boynton (1994). In this study, targets were vertical Gabor patterns, and backgrounds were sine wave gratings of the same spatial frequency and orientation. Contrast thresholds were measured in the fovea for three background phases: -90, 0, and 90 degrees relative to the target. Phase was found to have a small effect on detection thresholds relative to the effect of spatial frequency and orientation.

Properties in the region surrounding the target can effect detection through surround suppression mechanisms. Specifically, the surrounding luminance (Cornsweet & Teller 1965, De Valois, Webster, De Valois & Lingelbach 1986) , and contrast (Xing & Heeger 2000) can effect the perceived brightness and contrast of a target. Furthermore, the orientation (Blakemore & Nachmias 1971), position (Field et al. 1993), and spatial frequency (Polat & Sagi 1993) of the surrounding stimulus all effect the detectability of the central target.

2.4 Detection in natural image backgrounds

It is clear that background luminance, contrast, and spatial structure have a strong influence on target detectability. Like most aspects of vision, however, our understanding of the factors that affect the visual system's ability to detect a spatial pattern comes from experiments with simple laboratory stimuli. Despite the long history of research on detection tasks, it is not known what factors affect detection in natural images. Lawful relationships (such as Weber's law for luminance) have been measured extensively on uniform or simple backgrounds. But it is unknown whether or not these laws hold in the natural environment, where the particular properties of the background vary in complex ways.

There have been only a few studies of detection in natural backgrounds. Bex (2010) found that human observers were sensitive to distortions in novel natural images even when the distortions did not affect the amplitude spectrum of the image. Winkler & Susstrunk (2004) measured human detectability thresholds of random noise patterns embedded in a database of 30 natural images. Here, the target was the noise pattern, and the mask was the natural image background. They found that these noise patterns were easiest to detect in featureless regions of the image, and most difficult to detect in regions with texture. Here it is important to note the distinction between detecting an object in the environment (and ruling out all others) and detecting any change in the image. In natural behavior, we are often concerned with detecting particular objects, not just any object or change. Therefore, it is important

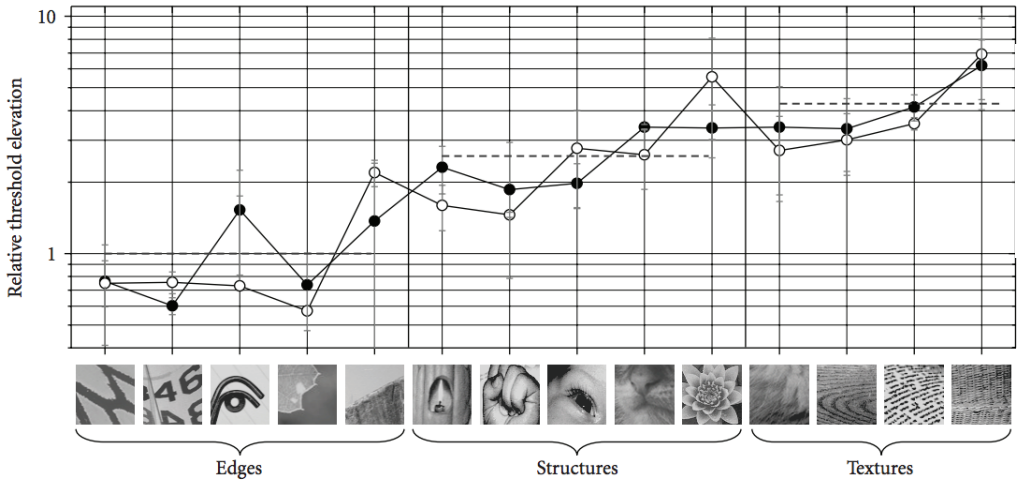

Figure 2.8: The effect of edges and structure on detection thresholds (Chandler et al. 2009). Thresholds relative to a baseline are shown as a function of the background type. In each case, the background contrast was fixed to either 0.32 (open circles) or 0.64 (filled circles). Thresholds are lowest in edge structures, and highest in textures.

that models of detection account for this.

Alam et al. (2014) measured contrast detection thresholds for a large database of natural images using a noise pattern as the target. The noise pattern was created by filtering noise with a vertical Gabor filter having a bandwidth of 1 octave. This work was an extension of previous work by Chandler et al. (2009). In both studies, amplitude thresholds were found to increase with background luminance and contrast. Thresholds also varied with different types of spatial patterns in the background image, independent of the background contrast and luminance (Fig. 2.8).

Both Alam et al. (2014) and Chandler et al. (2009) used a 3AFC

paradigm, where an identical background stimulus was presented three times in a single interval, and one of the backgrounds contained the target pattern. This paradigm is similar to detecting any change in the background, as opposed to detecting a particular target. Of course, when we try to detect a particular object under natural conditions, it is rare that we get to compare the identical background sample with and without the object. Thus, in this sense, the paradigm is not representative of natural detection tasks.

Bradley et al. (2014) used a single interval 2AFC paradigm to study the effect of increasing background contrast in natural images on detection thresholds across the visual field for a number of targets. Contrast power thresholds were found to increase linearly with background contrast power (Fig. 2.9). The slope of the increase changed as a function of eccentricity. The natural images used in that study, however, were transformed such that their luminance histograms were normally distributed. In natural images the luminance histograms are often skewed. Furthermore, the purpose of this study was not to characterize the laws of vision in natural images, but rather to build a biologically plausible model for detection in arbitrary scenes.

There is a large body of literature reporting measurements and models of detection in noise backgrounds in and in medical images (see Eckstein et al. (1997), Watson & Ahumada (2005), Burge & Geisler (2011)). Like natural images, medical images have complex statistical structure. It is likely, however, that this statistical structure is different in many ways for that of natural images of the typical human environment. Nonetheless, models of detection in

medical images may be relevant for understanding detection in natural images.

There has also been extensive work in the engineering field of image processing concerning the human visual system's ability to detect artifacts in a natural photographic images, along with a related engineering literature (Eskicioglu & Fisher 1995, Pappas, Safranek & Chen 2000, Wang, Bovik,

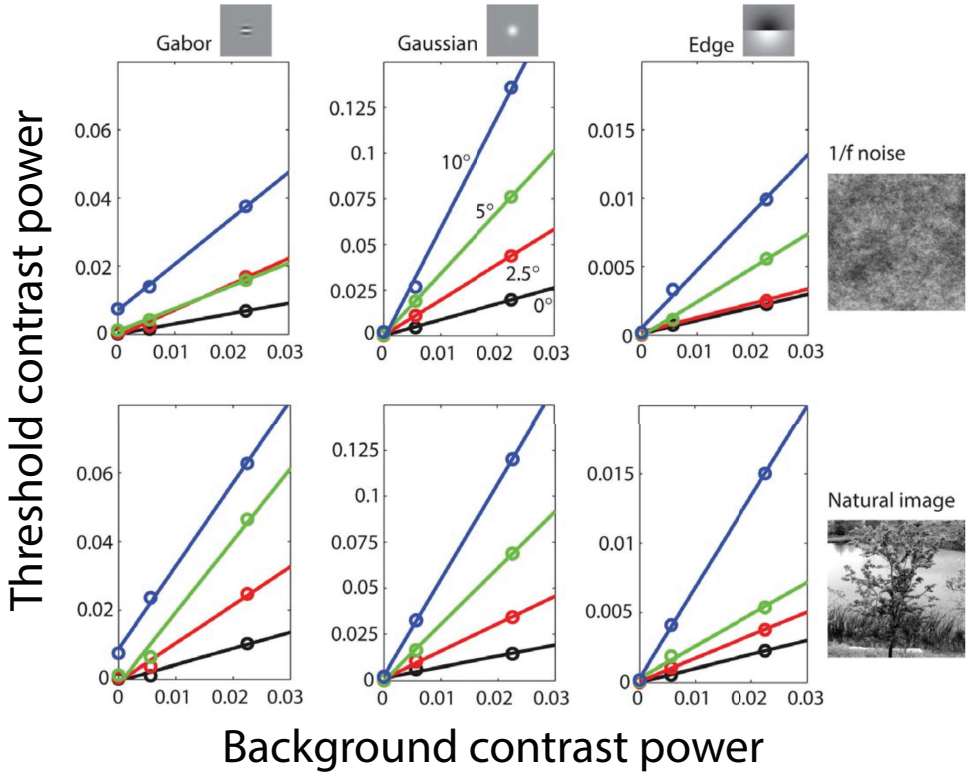

Figure 2.9: Contrast thresholds (shown on the y axis) were measured as a function of background contrast (shown on the x axis), and eccentricity (shown as different colored lines). The top row shows thresholds in 1/f noise, and the bottom row shows thresholds in natural images with Gaussianized luminance histograms. Each column shows data for a different target. In all cases, contrast power at threshold rises linearly with background contrast power.

Sheikh & Simoncelli 2004) concerned with predicting the subjective quality of images (images with detectable artifacts are likely to have lower subjective quality). These engineering models are designed for many practical applications. For example, when building an image compression algorithm any image artifacts left after compression that are below or near the detection limit are acceptable. These models, however, differ from models of natural detection in that they seek to capture sensitivity to any change in the image (e.g., compression artifacts) rather than the visual system‘s ability to detect a specific pattern.

2.5 Next steps

It is clear from the work described in this chapter that many factors affect our ability to detect spatial targets. The work presented in the rest of this dissertation will focus on three of these factors: the background luminance, contrast, and similarity to a specific target. The effect of these three factors is demonstrated in Figure 2.10. The first row shows mean luminance decreasing across columns. Backgrounds on the second row are noise patterns were each element is sampled from a Gaussian distribution. The standard deviation of this distribution decreases across columns, but the mean luminance stays fixed. Targets in the third row are added to a line noise pattern that consists of vertical and horizontal stripes. The noise orientation relative to the orientation of target grating decreases across columns. In each case, the target becomes easier to detect as the background property changes.

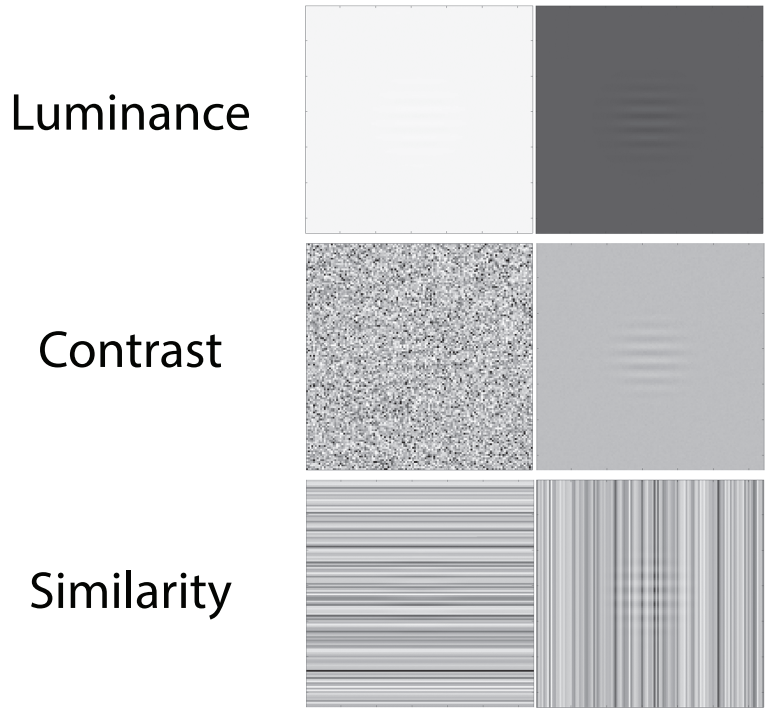

Figure 2.10: Example of three background properties on target detectability. A Gabor target of a fixed frequency, orientation, and amplitude is added to each background. As luminance, contrast, and similarity of the backgrounds decrease, the target becomes easier to detect.

There is a large body of work in vision science that has shown lawful relationships between detection thresholds and background properties (such as Weber's law for luminance). However, it is currently unknown how these factors affect detection in natural image backgrounds which have complex structure. One major reason that target detection has not yet been studied extensively in natural images is the difficulty of using natural image stimuli. For simple patterned backgrounds, a new sample of noise (such as Gaussian

white noise) can be generated on each trial. In the next chapter, I will present a framework by which natural image stimuli can be used in a similar way. Instead of presenting a new noise pattern on each trial, a new natural image was presented.

Chapter 3

Constrained scene sampling

3.1 The approach

Almost every study of visual pattern detection has used artificial backgrounds. A major advantage of this approach is that it allows for precise experimental control over the properties of the background. The ultimate goal of vision science, however, is to describe behavior and visual processing in the natural environment. Here I present a new method, called constrained scene sampling, that allows images of natural scenes to be used as backgrounds in a detection experiment, while maintaining the precise experimental control of artificial noise backgrounds. Although I apply this method to a detection task, this approach can theoretically be used in many different visual tasks, so long as a database of images exists that is appropriate for the task.

In constrained scene sampling, a large collection of natural images is first split into millions of smaller patches. Each patch is then sorted into one of 1000 narrow bins according to its luminance, contrast, and similarity to a given target. Patches from these bins are sampled to be used as background patterns in an experiment.

3.2 A natural image database

I used a database of 1204 images of nature scenes collected around the Austin, Texas area (Geisler & Perry 2011). The database can be found online at http://natural-scenes.cps.utexas.edu. These images contained a variety of natural scenes, including both dense and sparse foliage, small bodies of water, and various landscapes. Images in the database were 14 bits in pixel depth and linear in luminance, meaning each pixel step corresponds to a fixed step in luminance. Each image was 2844x4284 in size. Grayscale cropped sections of several images are shown in Figure 3.1.

For the present analysis I am interested in luminance stimuli (as opposed to chromatic stimuli). Therefore, the RGB images were converted to the luminance space by first converting to the XYZ color space and then taking the Y values. For a detailed description of the camera calibration procedure and color space conversion see Ing (2010) and Geisler & Perry (2011).

The distribution of luminance values in a particular natural image tends to be skewed towards high values. In order to increase the range of presentable images on the experimental display, the top 1 percent of all pixels in each image were set to the maximum value. The images were quantized to 8 bits, and the inverse gamma of the experimental display was then applied. These images were used in the experiment. For analysis, the gamma of the experimental display was applied. This was done in order to convert the images to the luminance values that would be displayed on the monitor in the experiments.

Figure 3.1: Example images from the natural image database. Images in the database contain a variety of natural scenes. The images shown here have been preprocessed (see text). Each image is a 1000x1000 pixel section that was cropped from the larger 2844x4284 images.

Each of the 1204 images was split into 101x101 pixel patches. This patch size corresponded to 0.8 degrees of visual angle in the experiment, and was the same size as the target stimuli in the experiment. The spacing between each patch was 10 pixels. In total there were 71,435,728 patches in the database.

3.3 Measuring natural image statistics

Several statistics were computed for each of the image patches. These statistics were mean luminance, RMS contrast, and similarity of the background to a given target. Here I consider two targets, which I will refer to as $T(x, y)$. One target is narrowband in frequency and has a single dominant orientation (a 4 cpd horizontal windowed grating in cosine phase), and one narrowband in frequency with two dominant orientations (a windowed plaid consisting of a vertical 4 cpd grating and horizontal 4 cpd grating). These targets are shown in Figure 3.3.

Both targets are mean zero, and were windowed with a raised cosine envelope, $env(x, y)$. The statistics were computed under the envelope of the target normalized to a volume of 1.0:

$$w_T(x, y) = \frac{env(x, y)}{\sum_{x,y}[env(x, y)]}$$

.

3.3.1 Mean luminance

The mean luminance for patch $B(x, y)$ is given by

$$L(x, y) = \bar{B}(x, y) = w_T(x, y) \cdot B(x, y)$$

Here, I measure mean luminance in units of the percentage maximum value. In the experiment, a mean luminance of 100 corresponds to 98 cd/m^2. The distribution of mean luminance values is shown in Figure 3.3A. The average luminance value in the database was 21.05.

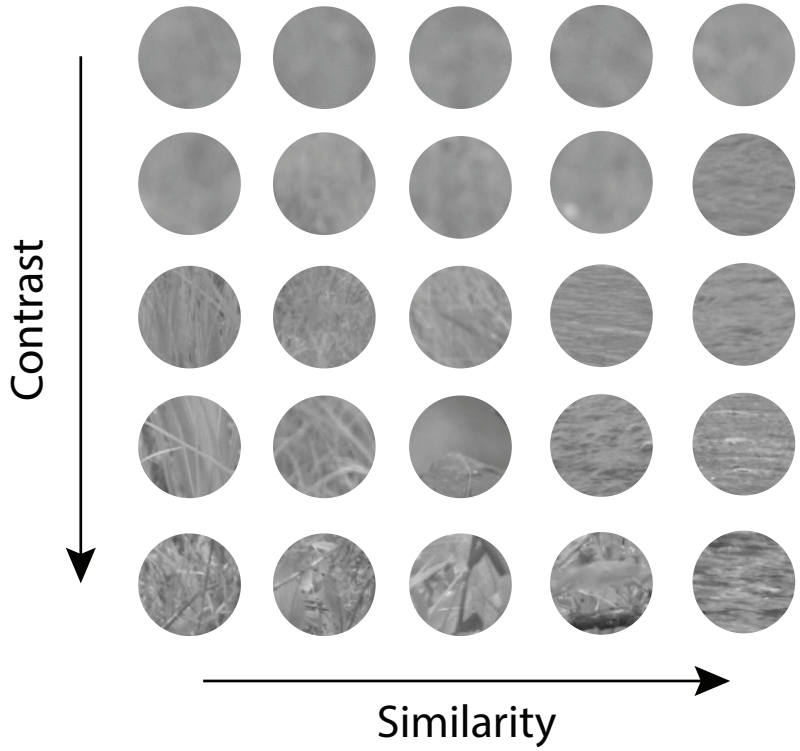

Figure 3.2: Example patches from the database. Similarity is shown for the horizontal Gabor target. Similarity increases across columns, and contrast increases down the rows. All patches shown have the same mean luminance.

3.3.2 RMS contrast

The RMS contrast for patch $B(x, y)$ is defined as:

$$C(x, y) = \|c(x, y)\sqrt{w_T(x, y)}\|$$

where $c(x, y) = \dfrac{B(x, y) - L(x, y)}{L(x, y)}$ is the contrast image at position (x, y).

The distribution of RMS contrasts is shown in Figure 3.3A. The mean value of RMS contrast was 0.508.

3.3.3 Similarity to target

Two measures of background similarity to a target were measured: phase-specific ($Q(x, y)$) and phase invariant ($S(x, y)$). Each measure is designed to capture the extent to which the background matches the spatial structure of the target. For the phase-specific similarity, $Q(x, y) = 1$ means the target exactly matches the spatial frequency, orientation, and phase of the target. For the phase-invariant similarity, $S(x, y) = 1$ means the target exactly matches the spatial frequency and orientation, but not necessarily the phase of the target. In both cases, the similarity measure is invariant to the contrast of the target and the background. In the present study, when I refer to similarity I am referring to the phase-invariant similarity $S(x, y)$ unless stated otherwise.

The phase-specific similarity is given by the cosine of the vector angle between $T(x, y)$ and the difference signal $\Delta B(x, y)$

$$Q(x, y) = \frac{T(x, y) \cdot \Delta B(x, y)}{\|T\| \|\Delta B 1_T\|}$$

where $\Delta B(x, y) = B(x, y) - \bar{B}(x, y)$ is the difference between the patch and its mean luminance, and 1_T is the indicator function

$$1_T = \begin{cases} 1, & \text{if } w_T > 0 \\ 0, & \text{if } w_T = 0 \end{cases}$$

Note that this definition of similarity is identical to the correlation between the target and the background.

The phase-invariant similarity is the cosine of the vector angle between the amplitude spectra of the target, $A_T(u, v)$ and the amplitude spectra of

the background difference signal $A_{\Delta B}(u,v)$. Before converting to the Fourier space, a flat topped raised cosine window was applied to remove the hard edge of the patch itself. This window is defined as

$$w(x,y) = \begin{cases} 1.0, & \text{if } \sqrt{(x^2+y^2)} < r_1 \\ 0.5 + 0.5cos\left(\frac{\pi(\sqrt{(x^2+y^2)} - r_1)}{(r_0 - r_1)}\right), & \text{if } r_0 < \sqrt{(x^2+y^2)} < r_1 \\ 0.0, & \text{if } \sqrt{(x^2+y^2)} > r_0 \end{cases}$$

Here r_1 was chosen to be 91 pixels and r_0 was 10 pixels.

Using this window the Fourier spectrum for the target is $T(x,y)w(x,y) \xrightarrow{F} F_T(u,v)$ and the amplitude spectra is $A_T(u,v) = |F_T(u,v)|$. Similarly, the Fourier spectrum for the patch is $\Delta B(x,y)1_T(x,y)w(x,y) \xrightarrow{F} F_{\Delta B}(u,v)$ and the amplitude spectra is $A_{\Delta B}(u,v) = |F_{\Delta B}(u,v)|$.

The phase-invariant similarity is given by

$$S(x,y) = \frac{A_T(u,v) \cdot A_{\Delta B}(u,v)}{\|A_T\|\|A_{\Delta B}\|}$$

The distributions of phase-invariant similarities for both the Gabor and plaid targets are shown in Figure 3.3A. The mean similarity for the Gabor target was 0.261, and the mean similarity for the plaid target was 0.349.

3.4 Binning natural image patches

Patches were sorted into one of 1000 (10 along each dimension) bins according to their mean luminance, RMS contrast, and phase-invariant similarity. The phase-specific similarity was used in building a model for detection

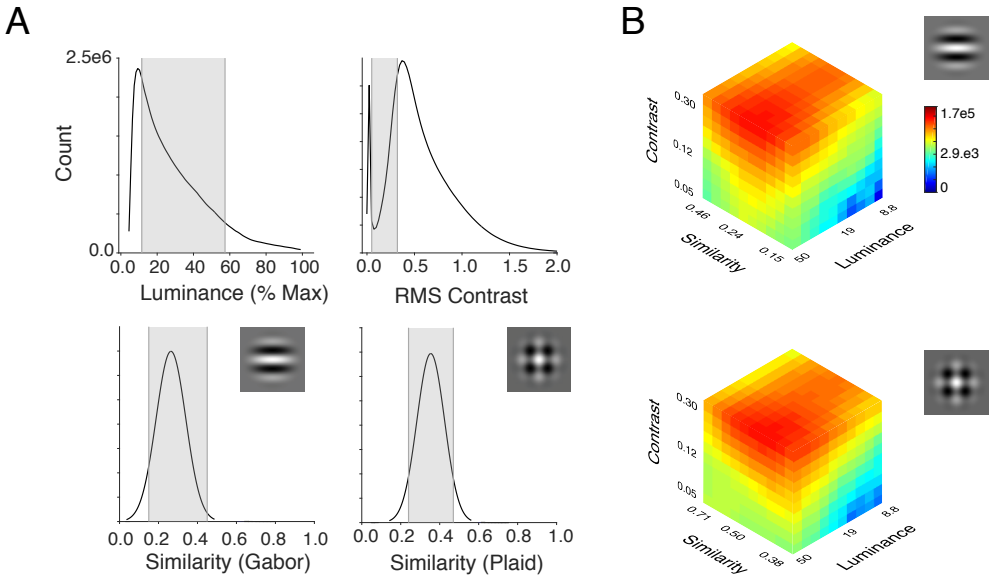

Figure 3.3: Distribution of image statistics. A. Individual distributions of each statistic. The shaded area is the area measured in the experiment. B. Joint distribution of all three statistics for both targets. Color represents the count in each bin. The color range is logarithmic to show the distribution of the statistics. Each bin in the histogram had at least 300 patches, and many bins contained several thousand patches.

(Chapter 5). Bin edge spacing was chosen so that the variation in the masking effect was approximately the same within each bin. If Weber's law holds then the bin boundaries in a geometric series satisfy this constraint. Therefore, the ith bin boundary, x_i is given by

$$x_i = x_{min} \left(\sqrt[n]{\frac{x_{max}}{x_{min}}} \right)^i$$

where , x_{min} is the minimum bin boundary, x_{max} is the maximum bin boundary, and n is the total number of bins (10 along each dimension).

The bounds x_{min} and x_{max} where constrained by the limitations of the display used in the experiment. Patches that were too high in contrast or luminance were not binned because a threshold for those bins could not be measured without clipping. Patches that were too low in luminance or contrast were not binned because it would not be possible to measure a threshold for those patches given the bit depth of the monitor. Furthermore, bin edge spacing was chosen to ensure that each bin contained at least 300 patches. For luminance, $[x_{min}, x_{max}]$ was $[8.0, 55.0]$. For contrast, $[x_{min}, x_{max}]$ was $[0.05, 0.32]$. For similarity, $[x_{min}, x_{max}]$ was $[0.15, 0.45]$ for the Gabor target, and $[0.24, 0.47]$ for the plaid.

3.5 Sampling in an experiment

Typical detection experiments use noise patterns generated in software to study the effect of different background properties on masking. For example, an experimenter would measure thresholds for Gaussian white noise of increasing contrast to test the effect of background contrast on detection thresholds. Generating the backgrounds in software gives the experimenter precise control over the stimulus.

In the same way, sampling patches from the 3D histogram bins along different dimensions allows one to test the effect of these dimensions on masking. In this case, instead of generating a new noise background in software, a natural background is sampled from one of the bins. Each of the 1000 bins in the 3D histograms contains at least 300 image patches with the approximately

equal (within a narrow range) luminance, contrast and similarity.

In the next chapter, I will describe a series of experiments using this method that were designed to test the effect of luminance, contrast, and similarity on detection thresholds in natural image backgrounds.

Chapter 4

Detection in natural scenes

4.1 Motivation

Perhaps the most fundamental visual task is to detect target objects in the natural backgrounds that surround us. In chapter 2, I described many of the factors that affect our ability to detect spatial targets. In this chapter, I will focus on understanding how the specific properties of the background influence detectability in natural image backgrounds. There is a large body of work in vision science that has shown that the detectability of a target pattern added to a background decreases with increases in background luminance (König & Brodhum 1889, Hecht 1924, Mueller 1951, Cornsweet & Teller 1965), background contrast (Legge & Foley 1980, Burgess et al. 1981, Lu & Dosher 1999) and with the similarity of the spatial properties of the background to that of the target (Campbell & Kulikowski 1966, Stromeyer & Julesz 1972, Wilson et al. 1983). Many of these findings show that thresholds change in a predictable way as the background properties change (for example, luminance thresholds increase linearly with background luminance). In addition to the background properties underneath the target, the properties of the region surrounding the target have been shown to have an effect on detection (Blakemore & Nachmias 1971, Polat & Sagi 1993, Xing & Heeger 2000).

Visual systems are the result of evolution by natural selection, and as a consequence their design is strongly constrained by the properties of natural visual stimuli and by the specific visual tasks performed to survive and reproduce. Thus, to understand the human visual system it is critical to characterize natural visual stimuli and performance in natural visual tasks. Despite this, much of our understanding of the background properties that affect detection has come from studies using either fixed backgrounds (such as uniform fields or gratings), or artificial noise patterns (such as white noise or 1/f noise).

The relatively few studies that have used natural images as stimuli either use just a few natural images (Caelli & Moraglia 1986, Rohaly, Ahumada & Watson 1997, Winkler & Susstrunk 2004, Alam et al. 2014), or a detection task that is not representative of detection in the natural environment (Caelli & Moraglia 1986, Alam et al. 2014). The latter studies allowed subjects to compare a sample of the background with the target to the same sample of background without the target in order to make a judgement. Of course, this type of direct comparison task is not common in the natural environment.

No study to date has directly measured in natural images the effect of background factors that are known to affect thresholds in simple backgrounds (namely luminance, contrast, and similarity). Because of this, it is unknown whether or not detection thresholds in natural image backgrounds follow the same types of relationships that have been measured in simple backgrounds. Furthermore, it is unknown how these factors - along with the underlying neural mechanisms for detection - are related to the statistical properties of

natural scenes.

One reason that few studies have used natural images as stimuli is because they are difficult to control in an experimental setting. In the previous chapter, I described a new experimental approach I developed called constrained scene sampling. In this approach, natural images are split into small local patches, and sorted into many bins according to their statistical properties. Image patches are sampled from these bins in order to be used as background stimuli in an experiment. Using this approach, natural images backgrounds can be used in a controlled experiment. On each trial of the experiment, instead of generating a new noise pattern (with a fixed luminance, contrast, and similarity) to be used as a background, a new natural image background (with a fixed luminance, contrast, and similarity within a narrow range) was sampled from a particular bin.

4.2 Experimental setup and methods

Detection thresholds were measured in the fovea using natural image patches as background stimuli. Thresholds were measured for two different targets types: one with a single dominant orientation (a windowed 4 cpd grating), and one with two dominant orientations (a windowed 4 cpd plaid). The targets are shown in Figure 3.3 on page 48.

4.2.1 Experimental procedure

Three experienced psychophysical observers participated in the experiments. Subjects were males whose ages ranged from 29-33 years old. All subjects had normal vision.

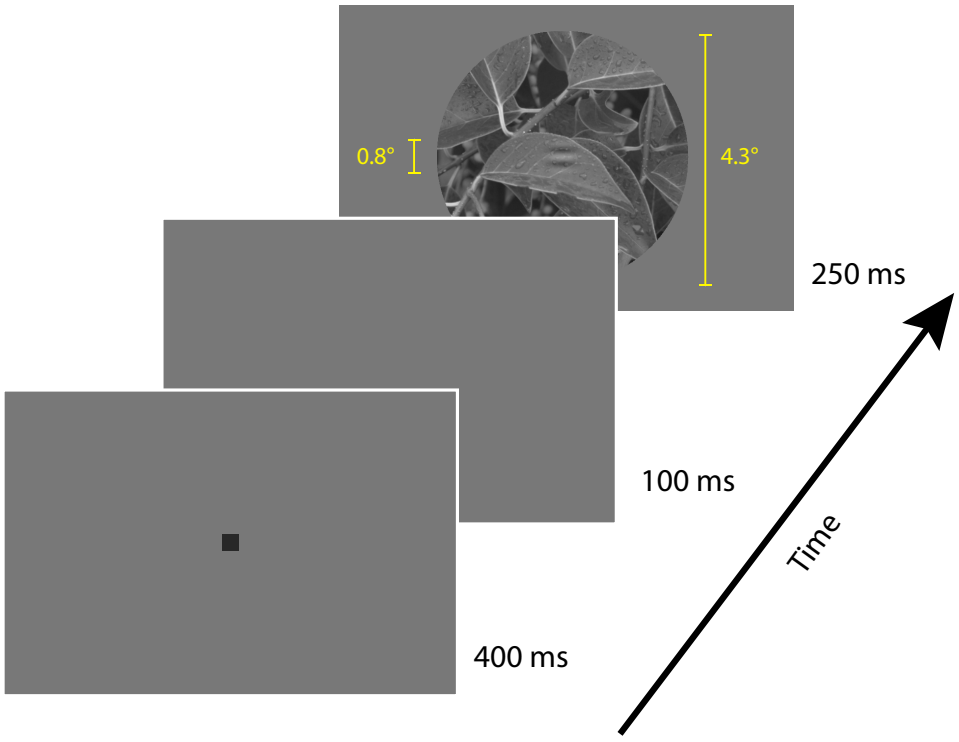

Figure 4.1: Experimental procedure in the detection task. On each trial, a low contrast fixation point was presented in the center of the screen for 400 ms. Next, a uniform field was presented for 100 ms. Finally, a natural image background (with or without the target) was presented for 250 ms. The task was to report whether or not the target was present. In experiment 1, the full 4.3 degree surround was presented (as shown). In experiment 2, the surround was tapered via a raised cosine window to the mean luminance (not shown).

Trials were blocked by the target type (Gabor or plaid), target amplitude, and the image database bin that the background images were sampled from. On each trial (Figure 4.1), a low contrast fixation point was presented in the fovea for 400 ms. Next, the fixation point was turned off for 100 ms. Finally, a background patch was randomly selected (without replacement) from the bin being tested. Randomly on half the trials the target was added to the background as follows:

$$I(x,y) = \begin{cases} B(x,y), & \text{if target absent} \\ B(x,y) + aT(x,y), & \text{if target present} \end{cases}$$

where $B(x,y)$ is the background patch, $T(x,y)$ is the target, and a is the target amplitude. The resulting image, $I(x,y)$ was then gamma compressed, based on the calibration of the display (SONY GDM-FW900), quantized to 256 10-bit gray levels, and displayed at a resolution of 120 pixels per degree for 250 ms. Stimulus presentation and response collection were programmed in MatLab, using PsychToolbox (Brainard 1997, Pelli 1997, Kleiner, Brainard, Pelli, Ingling, Murray, Broussard & others 2007).

At the end of each trial, the subject reported whether the target was present or absent. Subjects were given auditory feedback that indicated whether or not their choice was correct. This single interval yes/no task was chosen because it more closely matches detection tasks in the natural world, where observers do not make comparison judgments between the same background with and without the target.

In the first experiment, the background that was displayed also included

a context region that was 4.25 degrees wide surrounding the target region (See Figure 4.1). In the second experiment, the background, $B(x, y)$, was windowed to the mean luminance field with the target cosine window envelope $B_w(x, y) = B(x, y)env(x, y)$. After windowing, the target (if present) was added to the windowed background in the same way as shown previously. In both cases, the rest of the display contained a fixed luminance which corresponded to the mean luminance of the patches in the bin being tested.

Psychometric functions were measured separately for several bins in each experimental session. Each psychometric function was measured twice on each subject; the second after all the psychometric functions had been measured once. For each bin, five target amplitude levels were measured. At each amplitude level, 36 trials were collected (each with a different natural image background). To help the subject adopt the appropriate decision criterion, the first trial in a block always contained the target (this trial was not included in the data analysis). In total, 9,100 different natural image patches were used in the experiments.

4.2.2 Fitting the psychometric function

For each subject, all the psychometric data for each bin (350 trials) were fit with a generalized cumulative Gaussian function using a maximum-likelihood procedure. The probability of a hit is given by

$$P_h = \Phi\left(\frac{1}{2}\left(\frac{a}{a_t}\right)^{\beta} - \gamma\right)$$

and the probability of a false alarm is given by

$$P_{fa} = \Phi\left(-\frac{1}{2}\left(\frac{a}{a_t}\right)^{\beta} - \gamma\right)$$

where a is the amplitude of a target on a given trial, a_t is the amplitude at threshold, β is the slope parameter of the psychometric function, and γ is the bias parameter. Threshold was defined to be the target amplitude corresponding to 69% correct ($d' = 1$).

The threshold (a_t), slope (β), and bias (γ), were fit by maximizing the log-likelihood of all the responses to a given condition with the number of hits, misses, false alarms, and correct rejections for a particular amplitude. A single slope parameter (β) was fit for each subject, target, and experiment. In all experiments, the bias parameter (γ) was found to be small, and therefore was set to 0.

4.3 Experiment 1: Natural images with surrounding context

The goal of the first experiment was to characterize the effect of background luminance, background contrast, and background similarity to a target on the detectability of a known target (see Chapter 3 for a description of these factors) in natural images.

In Chapter 3, I described the creation of a database of image patches that were sorted into one of 1000 bins, each with a narrow range of luminance, contrast, and similarity. Fully characterizing this space would require the

collection of 2000 psychometric functions (one for each bin for each target), which is, of course, impractical. Previous studies using simple backgrounds have shown that thresholds vary smoothly as a function of the background dimension. Therefore, in order to characterize the effect of each dimension, I measured detection thresholds for a subset of the bins, under the assumption that thresholds would vary smoothly, and only a subset would be needed to characterize the space.

In Experiment 1, detection thresholds were measured in the fovea at five bins along each of the three dimensions in the 3D histogram of image patches. Along each dimension, the other two dimensions were held in fixed bin positions. In the luminance condition, the luminance values ranged from 8.85 to 50.18 percent of the maximum monitor luminance. The contrast was held fixed at 0.12 and the similarity was held fixed at 0.25 for the Gabor and 0.32 for the plaid. In the contrast conditions, the five RMS contrast values ranged from 0.06 to 0.29. The luminance was held fixed at 23.2, and the similarity was held fixed at 0.25. The similarity values measured ranged from 0.16 to 0.38 for the Gabor target, and 0.25 to 0.43 for the plaid target. For both the Gabor and the plaid, the luminance was fixed at 23.2, and the RMS contrast at 0.12.

The results for three different subjects are shown in Figure 4.2. The top row of panels shows the results for the Gabor target, and the bottom panels show the results for the plaid target. Each plot shows threshold amplitude (on the y-axis) as a function of the value along a background dimension (on the

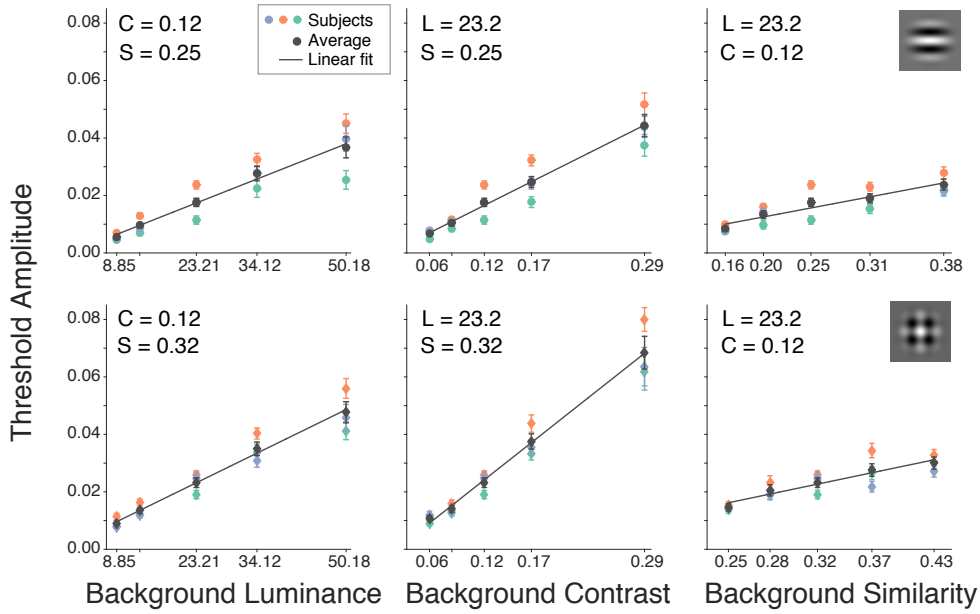

Figure 4.2: Results from experiment 1. The top row shows results for the Gabor target, and the bottom row shows results for the plaid target. The colored points show the data from three subjects, and the black points show the average. Error bars are bootstrapped standard errors. The black line shows the best fitting line to the average data.

x-axis). Values on the x-axis represent the middle value of the bin that was measured. Subject data are shown in the colored points, and the average is shown in black. The best fitting line to the average points is shown in black.

For both targets, threshold amplitude increases approximately linearly with local mean luminance in natural backgrounds (Figure 4.2, first column), in agreement with the classic finding of Weber's law reported for detection in uniform backgrounds (König & Brodhum 1889, Hecht 1924, Mueller 1951). The threshold amplitude increases approximately linearly with background

contrast (Figure 4.2, second column), in agreement with the classic finding for detection in white noise (Legge & Foley 1980, Burgess et al. 1981, Lu & Dosher 1999). This result is also in agreement with more recent findings for targets in $1/f$ noise backgrounds (Najemnik & Geisler 2008, Bradley et al. 2014), and in Gaussianized natural backgrounds (Bradley et al. 2014). Finally, amplitude threshold increases approximately linearly with the similarity of background to the target (Figure 4.2, third column). Previous studies have shown an effect of background orientation and spatial frequency on detection thresholds (Campbell & Kulikowski 1966, Stromeyer & Julesz 1972), however this linear relationship is a result not previously reported. There is individual variation between subjects (some are overall more sensitive than others), however the approximate linear trend holds for all subjects. The slope parameters for the psychometric functions are shown in Table 4.1.

Slope (β) parameters in Experiment 1		
	Gabor	Plaid
Subject 1	0.997	1.411
Subject 2	1.695	1.910
Subject 3	1.300	1.390

Table 4.1: Slope parameters in Experiment 1 for all three subjects

The primary conclusion from this experiment is that thresholds increase approximately linearly as a function of luminance, contrast, and similarity for detection in natural backgrounds. This largely confirms the findings of previous studies that used simple backgrounds (although the similarity result here is new). It might at first not seem surprising that detection thresholds

follow similar laws in natural images as in simple backgrounds. However, given the structure of natural images, the complex ways that luminance, contrast, and similarity vary, and the variation between images on each trial, it was very possible that linear relationships would not be obtained. In section 4.5 of this chapter, I show that this pattern of results follows directly from a principled signal-detection theory analysis of detection in natural images.

4.4 Experiment 2: Natural images with no surrounding context

The first experiment measured detection thresholds when the backgrounds were substantially larger (4.3 deg in diameter) than the target (0.8 deg in diameter). To examine the effect of the background region surrounding the target, I carried out a second experiment where the background was windowed to size of the target. A subset of the conditions measured in Experiment 1 was measured in Experiment 2. All backgrounds in Experiment 2 were sampled from a fixed luminance bin. The background contrast and similarity were varied in order to measure the effect of these dimensions on detection thresholds.

Figure 4.3 shows detection thresholds in Experiment 2 as a function of background contrast and similarity. For comparison, the best fitting lines to the average data from Experiment 1 are also plotted in a lighter shade. For both targets, threshold increases approximately linearly with the background contrast and similarity to the target. The slope (β) parameters for the

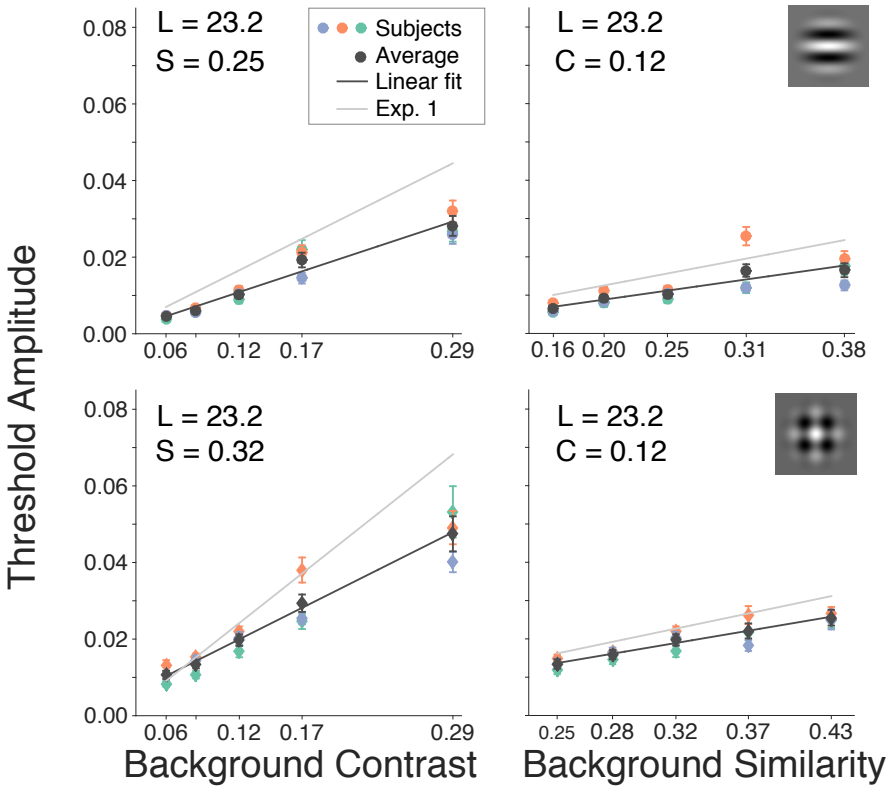

Figure 4.3: Results from Experiment 2. The top row shows results for the Gabor target, and the bottom row shows results for the plaid target. The colored points show the data from three subjects, and the black points show the average. Error bars are bootstrapped standard errors. The black line shows the best fitting line to the average data. The gray line shows the best linear fit from the average data in Experiment 1.

psychometric functions from each subject are shown in Table 4.2.

It is clear from Figure 4.3 that the overall magnitude of the thresholds in the windowed condition is lower than in the surround condition.This result appears to demonstrate that there is a substantial effect of the surrounding

context on detection thresholds. However, while the windowing procedure removed the surround region, it attenuated part of the background under the target (the half height of the cosine window used was 0.4 degrees, which was half of the target size). Because of this, some of the energy in the background that was under the target was removed by the window.

It is therefore possible that the measured thresholds are lower because masking energy was removed under the target. In the next section, I will show that this is the most likely explanation, and hence that the surrounding context region has little or no effect.

Slope parameters in Experiment 2		
	Gabor	Plaid
Subject 1	0.849	1.395
Subject 2	1.302	1.801
Subject 3	1.138	1.506

Table 4.2: Slope parameters in Experiment 2 for all three subjects.

4.5 Signal detection analysis of detection

4.5.1 Matched template observer

Experiments 1 and 2 showed that thresholds vary approximately linearly along all three dimensions (in agreement with previous findings using simple stimuli). But why should thresholds vary in this way? To gain some insight into this, I evaluated a simple signal detection model known as a matched-template observer (Figures 4.4 and 4.5). On each trial the matched-template observer computes the dot product of the target waveform $T(x, y)$ with the

input image $I(x, y)$:

$$R = \sum_{x,y} T(x, y)I(x, y)$$

where the target waveform is scaled so that its energy is 1.0 (the energy of the target waveform is the dot product of the waveform with itself). In this case, the dot product is equivalent to computing the response of a receptive field exactly matching the luminance profile of the target. If the template response (R) exceeds a decision criterion value (γ), then the observer reports that the target is present, otherwise that the target is absent.

The matched template observer is the optimal (ideal) observer when the target is known (as it was in experiments 1 and 2) and the background is Gaussian white noise (Peterson, Birdsall & Fox 1954, Green & Swets 1966, Burgess et al. 1981). A matched template observer is often not too far from optimal if the background is Poisson white noise or is correlated Gaussian noise (Geisler 2003). Although natural image backgrounds have a more complex

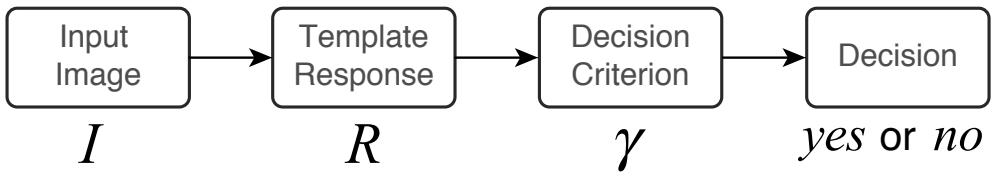

Figure 4.4: The matched template observer model. On each trial, an template response R is computed on the input image I. This response is compared to a decision criterion γ to determine whether or not the target was present.

statistical structure than Gaussian noise, the matched-template observer is a simple, principled signal detection model. It is therefore a good starting point for investigating detection in natural images.

4.5.2 Matched template observer in natural images

The purple histograms in Figure 4.5 show the distributions of template responses for the Gabor target, for all the non-windowed background patches in three of the 1000 bins in the image database (see chapter 3). The mean of the template response distribution is 0, because the template waveform sums to 0. The standard deviation of the response distribution is due to the structure and variability of the natural images themselves. If a target of amplitude a is added to the background, then the distribution remains identical in shape (with the same standard deviation), but is shifted to the right by a (the green histograms). The distributions are approximately Gaussian as indicated by the curves, which are Gaussian distributions having the same means and standard deviations as the measured distributions.

If the goal of the observer is to be as accurate as possible in the detection task - where the target is present on half the trials - then the decision criterion should be placed halfway between the two distributions ($\gamma = a/2$). In Figure 4.5, the target amplitudes have been set so that the accuracy of the matched template detector is 85% correct for the three example bins. As the standard deviation of the response distribution to background alone increases, the mean of the response distribution to background plus target must also

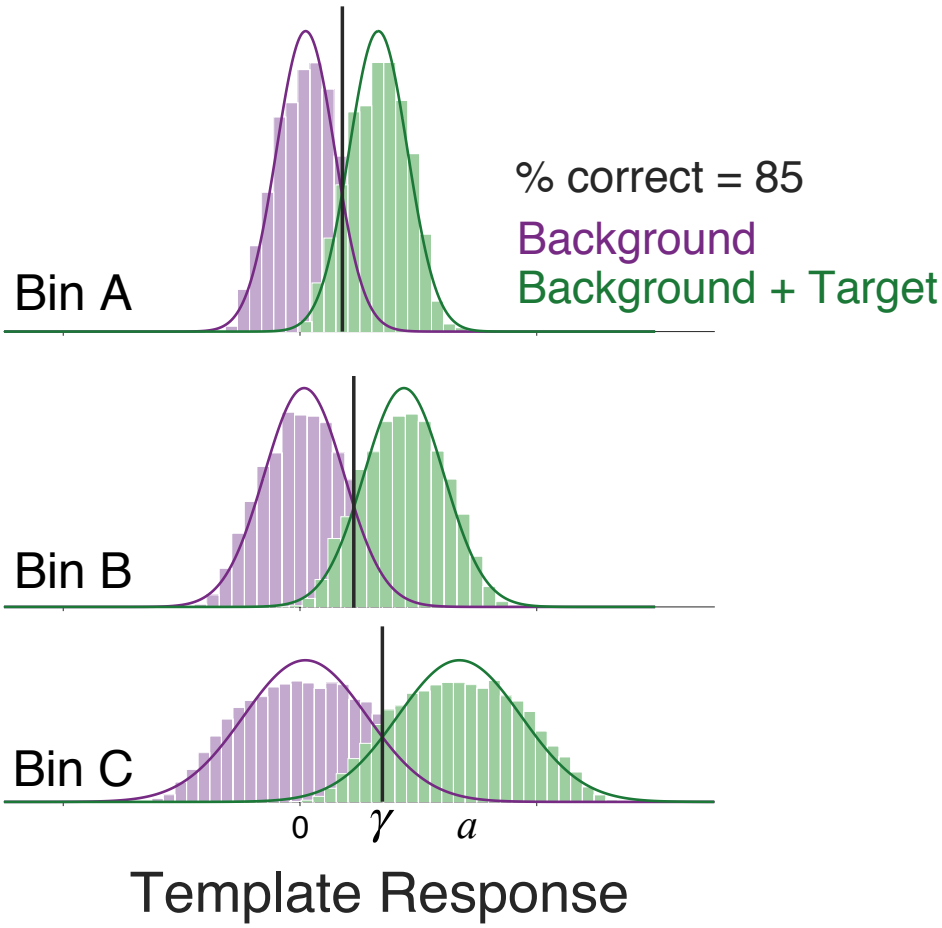

Figure 4.5: Template response distributions (for the Gabor target) for three of the 1000 bins in the experiment. Response to the background alone is shown in purple, and response to the background plus the target is shown in green. The mean of the background alone condition is 0, and the mean of the background plus target condition is the amplitude of the target. The amplitude here was chosen to give 85% correct in each condition.

increase to maintain 85% correct performance, and therefore the amplitude of the target must increase. But how is this related to detectability in natural image backgrounds?

In Chapter 2, I mentioned that the standard way to measure performance in a detection task is by using the signal to noise ratio, d'. For a particular bin with luminance, contrast, and similarity centered at L, C, and S, respectively, this ratio is

$$d' = \frac{|\mu_b - \mu_t|}{\sigma(L, C, S)}$$

where μ_b is the mean of the background response distribution (0), μ_t is the mean of the background plus target response distribution (a), and $\sigma(L, C, S)$ is the standard deviation of template responses to image patches in that bin. Plugging in the values for μ_b and μ_t, d' becomes

$$d' = \frac{a}{\sigma(L, C, S)}$$

Therefore, the target amplitude needed for a particular detectability level is

$$a = d'\sigma(L, C, S)$$

Thresholds in experiments 1 and 2 where defined as the amplitude corresponding to a $d' = 1$. In this case, the matched-template observer's threshold for detection of images in a particular bin is simply the standard deviation of the template responses for that bin

$$a_t = \sigma(L, C, S)$$

From this equation, it is clear that for the template observer the amplitude threshold for a particular bin is the standard deviation of the template response for the images in that bin. The standard deviation of the template response distribution is shown as a function of similarity, contrast, and luminance in Figure 4.6. Each plot shows a different luminance, each shows a different contrast level, and each symbol type shows a different target (circles for Gabor, and diamonds for plaid). The template response distributions were also computed for the windowed backgrounds (see Appendix A).

The curves through the data in Figure 4.6 show the fit of the following formula, which is separable in luminance, contrast, and similarity

$$\sigma(L, C, S) = k_0(L + k_L)\sqrt{C^2 + k_C}(S + k_S)$$

where k_0, k_L, k_C, and k_S are free parameters. These four parameters were fit to all 2000 data points (one for each of the 1000 bins per target) and were found to be $k_0 = 0.0275$, $k_L = -0.0768$, $k_C = -0.0154$, $k_S = -0.0712$. This descriptive model accounted for 99.6% of the variance in the data, implying that the thresholds of the matched template observer are linear in all three dimensions.

Here I am only showing the template response standard deviations for the non-windowed backgrounds. Appendix A shows the result for the windowed backgrounds. The linear relationship holds in that case, although the magnitude of the standard deviations is overall lower. This is likely due to the windowing removing structure from the background, and therefore lowering

the variability of the template responses. This result shows that thresholds in the second experiment should indeed be lower due to the windowing of the background under the target, and not due to the removal of the surrounding image region.

The fact that the standard deviations of the template responses increase as backgrounds increase along these dimensions is quite surprising, and there was no obvious reason to think that should be the case beforehand. This increase in the variability of the template response is due to the variability in the structure of the natural images themselves. This result demonstrates that the fact that human thresholds are approximately linear along all three dimensions is predicted from first principles. In other words, an observer that is limited only by the variability found in natural image structure has detection thresholds that increase linearly as a function of luminance, contrast, and similarity.

4.5.3 Human detection performance and the structure of natural images

The prediction of linear threshold functions for each dimensions is an important result but do the matched-template statistics actually predict the variation in slopes and intercepts across the different background dimensions and targets in Experiments 1 and 2 (shown in Figures 4.6 and A.1)? For detection in noise, humans do not reach the absolute levels of performance of the matched-template (ideal) observer, and thus the relative performance

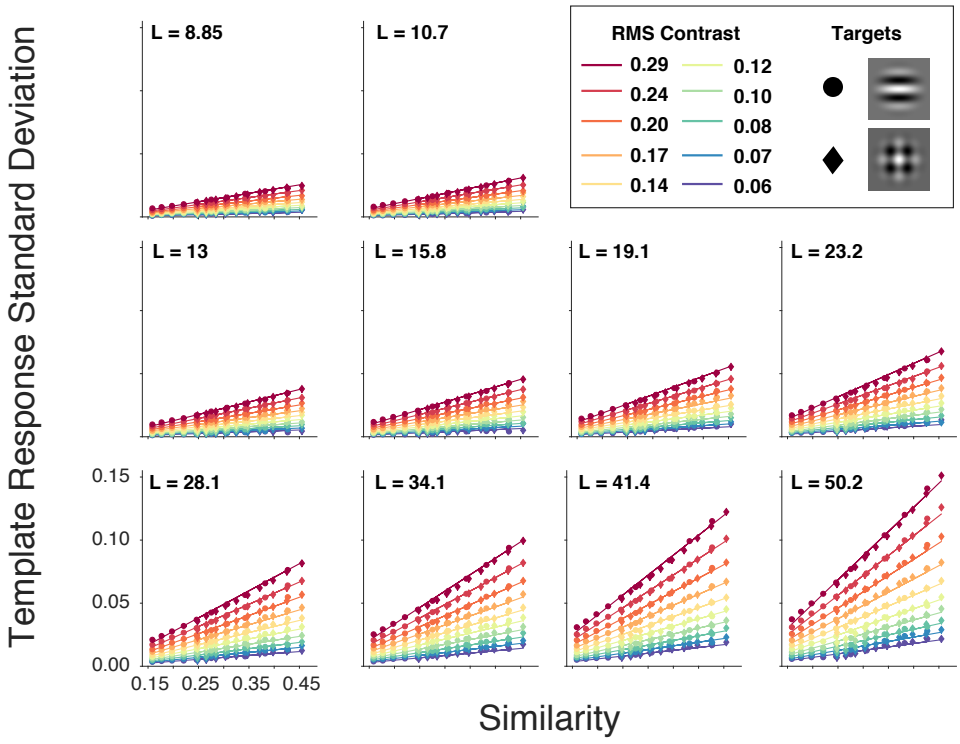

Figure 4.6: Standard deviation of template responses for each bin in the experiment (without windowing) for the Gabor target (circles) and plaid target (diamonds). Bins vary as a function of similarity on the x axis. Each color represents a different contrast level. Each plot represents a different luminance level. The solid lines are fits with a 4 parameter model (see text). Appendix A shows plots of template response standard deviations for windowed backgrounds.

of human and ideal observers is usually compared by introducing an overall efficiency parameter, η, that effectively scaled up the variance of the matched-template responses, or equivalently, scales up all the matched-template ob-

server's thresholds by a constant (Burgess et al. 1981, Geisler 2011):

$$a_t = \sigma(L, C, S)/\sqrt{\eta}$$

A single efficiency parameter was applied to the matched template observer's thresholds across all conditions, target types, and experiments (with and without windowing). For the windowed condition, the standard deviation of the template responses to the windowed patches was used. The efficiency parameter that minimized the mean squared error between the template observer and human observers was found to be 0.124.

Figure 4.7 shows the average human thresholds for both targets (Gabor in red circles, and plaid in blue diamonds), and for both experiments (the windowed surround data are shown in the lighter colors). The predictions of the matched template model for a single fixed value of the efficiency parameter are shown as the open gray symbols. As can be seen, the values of the slopes and intercepts across luminance, contrast, and similarity for both targets, and in both experiments are predicted quite well from the statistics of the template responses to natural backgrounds (with only a single free scaling parameter).

The correlation between the model threshold amplitude without the efficiency parameter and the human thresholds is 0.98 (as seen in Figure 4.8). This high correlation shows that the detection mechanisms in the human visual system are tightly matched to the statistical properties of natural scenes.

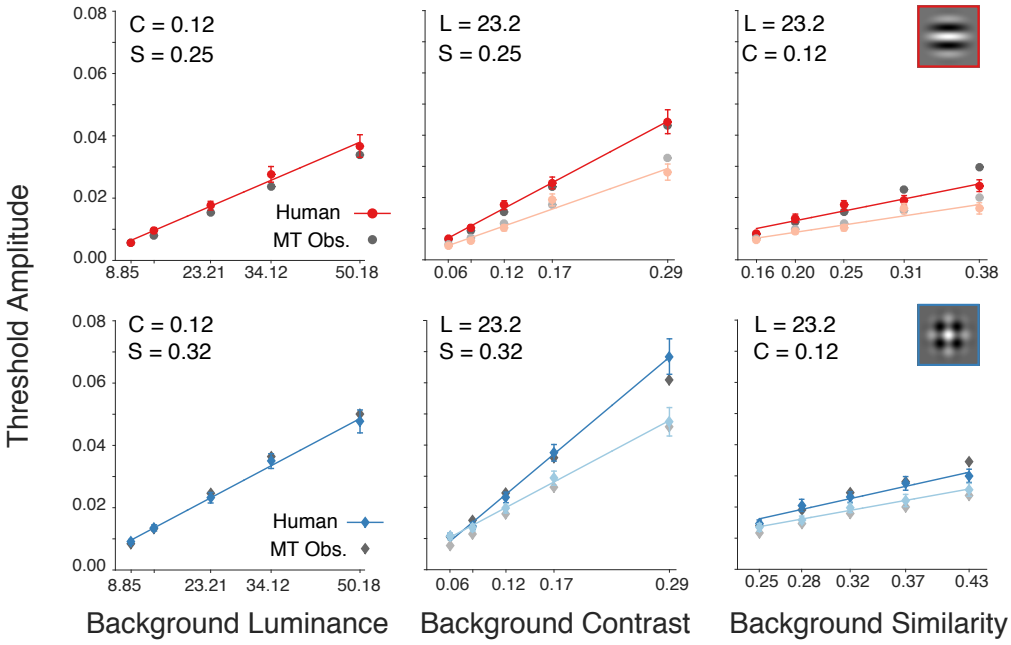

Figure 4.7: Matched template and human observer data for detection in natural scenes. The average of three human subjects is shown by the colored points (Experiment 1 is shown as the darker color, and Experiment 2 is shown in the lighter shade). The matched template observer (with a single efficiency parameter, 0.124) is shown by gray symbols. The lines are the best linear fit to the human data. The performance of the matched template observer tracks the performance of human observers.

4.6 Summary

The experiments presented in this chapter revealed lawful effects of the three dimensions on detection performance. Amplitude thresholds increased linearly with background luminance, contrast, and similarity with and without surrounding image context. These aspects of human detection performance are predicted quantitatively from first principles by a signal detection analysis of

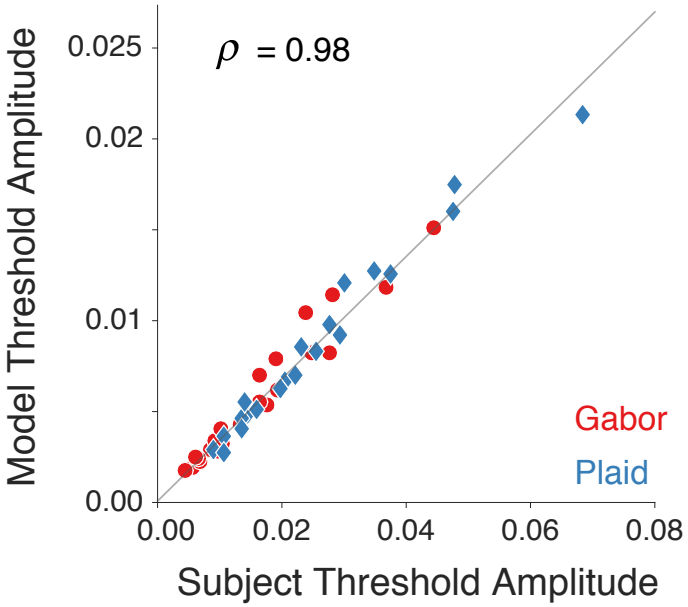

Figure 4.8: Plot of the correlation between the template observer thresholds (which are the template response distribution standard deviations) vs. subject thresholds. Each point represents the threshold as measured in a single bin. Both experiments are plotted here. The correlation was 0.98

the natural image stimuli. This analysis also reveals that there is little direct effect of removing the surrounding context on detection thresholds. The work presented here demonstrates that the ability of human observers to detect targets is highly similar to that of an observer that is limited only by the variability in natural images, and that human detection mechanisms are strongly linked to the statistics of natural scenes.

Chapter 5

Uncertainty in natural scenes

5.1 Motivation

In the previous chapter, I measured detection thresholds on human observers using over 9,000 different natural image stimuli as backgrounds. Experimental trials were blocked by the target amplitude, as well as the luminance, contrast, and similarity of the backgrounds. Despite the variety in the structure and content of the images presented, the luminance, contrast, and similarity of the background images varied by only a small amount trial-to-trial. The target amplitude (if present) did not vary trial-to-trial. Because of this, the human observers (as well as the matched template observer) could adopt a single decision criterion in order to maximize performance in this task.

In the natural environment, however, the luminance and contrast typically vary by an order of magnitude in a given scene (Hood 1998, Mante et al. 2005). The similarity between the background and a particular target varies by an order of magnitude on average in natural scenes (Appendix B). Furthermore, even when the shape of the target of interest is known, the strength (or amplitude) of the target often varies on every occasion. Under these conditions, matched-template observers would not be able to adopt a

simple decision strategy for detection.

Previous studies have shown that the uncertainty created by the random amplitude and location of the target, and the random variation in background properties can reduce detection performance (Tanner 1961, Eckstein 2011, Burge & Geisler 2011). However, in the natural environment, it is critical that we are able to detect targets of interest quickly. Therefore, it seems likely that there was strong evolutionary pressure to maximize detection performance in conditions of high uncertainty (like those found in nature). But how do human observers perform in more natural, uncertain conditions?

The goal of the work presented in this chapter was to measure and model detection performance under uncertain conditions. First, I measured the effect of uncertainty on human observers in a detection task where the background properties and the target amplitude varied randomly on each trial. Next, to test whether or not the surrounding image context helped observers under uncertain conditions, I measured the effect of background and target amplitude uncertainty with the surround windowed away (as in Chapter 4, Experiment 2). I tested the extent to which the matched template observer shown in Chapter 4 could account for the results in uncertain conditions. Finally, I developed an extension to the matched template model that that could account for the results in all three experiments.

5.2 Experimental methods

The same three observers that participated in Chapter 4, Experiments 1 and 2, participated in this experiment. The experimental timing and procedure for display was the same as in the previously mentioned experiments. In this experiment, trials were blocked by the target type (Gabor or plaid), and the performance level from the previous experiments. Four performance levels were tested: 65%, 75%, 85% and 95% correct.

On each trial, a bin that was measured in the first set of experiments was randomly selected, and a random patch was chosen from that bin without replacement. Randomly on half the trials, the target was presented. The am-

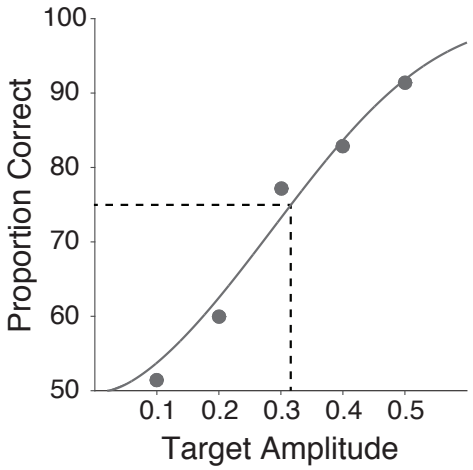

Figure 5.1: Target amplitude selection in the uncertainty task. The data shown are from a single subject and a single bin condition. In Experiment 3, trials were blocked by the performance level in Experiments 1 and 2. If the performance was 75% correct, then targets would be set to an amplitude of 0.31 for this subject and this bin.

plitude of the target was set to the amplitude that corresponded to a particular performance level (for a particular subject) on the psychometric curve for the data in that bin in the blocked experiments (Experiments 1 and 2 from Chapter 4). This is illustrated in Figure 5.1. The data shown are the responses of subject 1 in Chapter 4, experiment 1 for a particular bin condition. Suppose this bin was randomly chosen in the present experiment, and the performance condition being measuring was 75% correct. An image patch would be selected from the randomly chosen bin, and the target amplitude would be set to the value corresponding to 75% correct on this psychometric fit (around 0.31 in this case).

Performance in the uncertain condition was measured for each of the four accuracy levels, with the background bin randomly selected on each trial. For each level, 200 trials were collected separately in blocks of 50 trials. Under these circumstances, both the target amplitude and the background bin vary on each trial (unlike in Chapter 4, experiments 1 and 2, where both amplitude and background bin were blocked). If there was no effect of uncertainty, then performance should be unchanged from performance in the blocked conditions. Performance was measured separately for image patches with and without surrounding context.

Images were sampled only along a single luminance bin dimension (the median value), and therefore only the contrast and similarity randomly varied trial-to-trial. If the luminance bin was not fixed, then the surrounding uniform region of the display would either provide an unnatural cue to the luminance

(if it varied with the natural background), or it would produce brightness-contrast artifacts (if it was kept it at a fixed luminance).

5.3 Experiment 3: Uncertainy

The third experiment measured the effect of randomly varying the background and target amplitude on every trial on performance in the detection task. The uncertainty in this experiment is similar to the uncertainty we experience under natural conditions, where the background varies on every occasion.

First, performance was measured in images that contained the full surround (as in Chapter 4, Experiment 1). The left plot in Figure 5.2 shows the accuracy (averaged across the three subjects) in the random conditions plotted as a function of the accuracy in the blocked conditions of Chapter 4, Experiment 1. There is little (if any) effect of background and target uncertainty, even though subjects reported that the background appeared dramatically different on each trial.

As will be shown below, to perform well in this experiment a matched template detector would need to adjust its decision criterion on every trial. Perhaps the performance of the human observers is unaffected by uncertainty, because they are somehow able to use the surrounding background image context to adjust their decision criteria. To test this hypothesis, performance was measured in images where the surrounding context was windowed to the mean luminance (as in Chapter 4, Experiment 2). The right plot in Figure

5.2 shows the results from this experiment. The data points are plotted as a function of performance in Chapter 4, Experiment 2. Again, performance is largely unaffected by background and target amplitude uncertainty.

The fact that the performance of human observers remained unchanged under conditions of high background and amplitude uncertainty is a surprising result, because uncertainty is typically thought to harm performance in many tasks. Furthermore, the results from windowed images suggest that the human

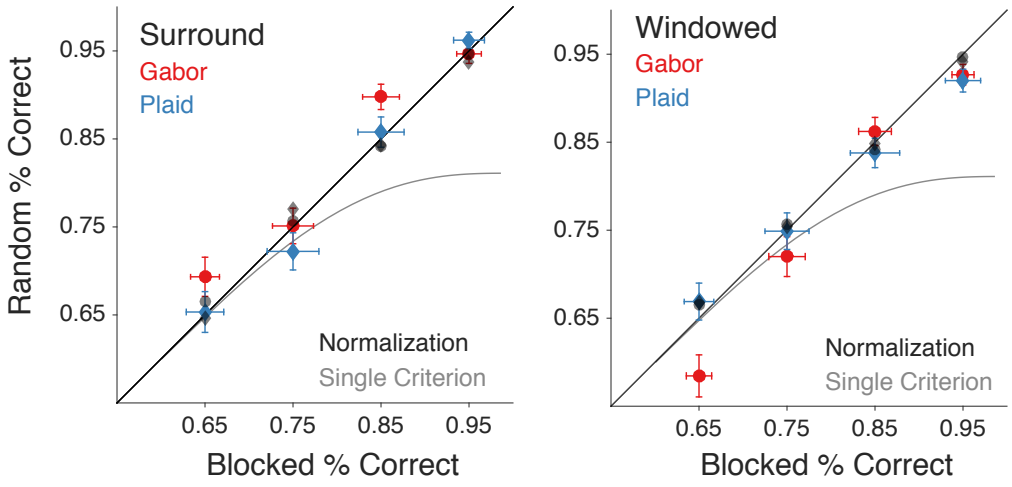

Figure 5.2: Results from Experiment 3. Shown is the performance in the uncertain condition as a function of performance in the blocked condition. The points show the average performance between the three subjects. The figure on the left shows data where the image was presented with a 4.3 degree surround (as in Experiment 1), and the figure on the right shows data where the surround was windowed (as in Experiment 2). The gray line shows the performance of the standard matched template observer, and the black lines shows the performance of the normalized matched template observer (see text for details).

visual system is somehow able to correctly adjust the decision criterion on each trial by using information only at the location of the target itself. In the next section, I will first describe how a matched template observer would perform under these uncertain conditions, and then describe an enhanced matched template observer.

5.4 Signal detection analysis of detection in uncertain conditions

5.4.1 Matched template observer

Consider the standard template matching observer for the target detection task I described in Chapter 4. On each trial, this observer computes the dot product of the background and a template of the target waveform. If the dot product exceeds a particular criterion the observer reports that the target is present. An ideal observer would set the criterion to the value that maximizes overall accuracy.

Figure 5.3 (left) shows the distributions of template responses for the Gabor target in three specific bins from the database. Here, the target was set to the amplitude that produced 85% correct. If the background bin and the amplitude of the target are fixed, as in Experiments 1 and 2, then the matched template observer would set the criterion at the cross point of the two distributions in that bin ($\gamma = a/2$). In Figure 5.3 (left), the optimal criterion for each bin is shown as the black lines. For bin A the standard deviation is relatively low, and hence the target amplitude and the criterion

that produces 85% correct are relatively small. For bins B and C the standard deviations are higher, and the hence target amplitudes and the criterion that produces 85% correct are larger.

This simple matched template observer applied to the natural image backgrounds predicted the entire set of results shown in Chapter 2. How would this model perform when the background bin randomly varies on every trial, as in the experiments presented in this chapter? The simple template observer can only set a single decision criterion. In this case, the optimal single

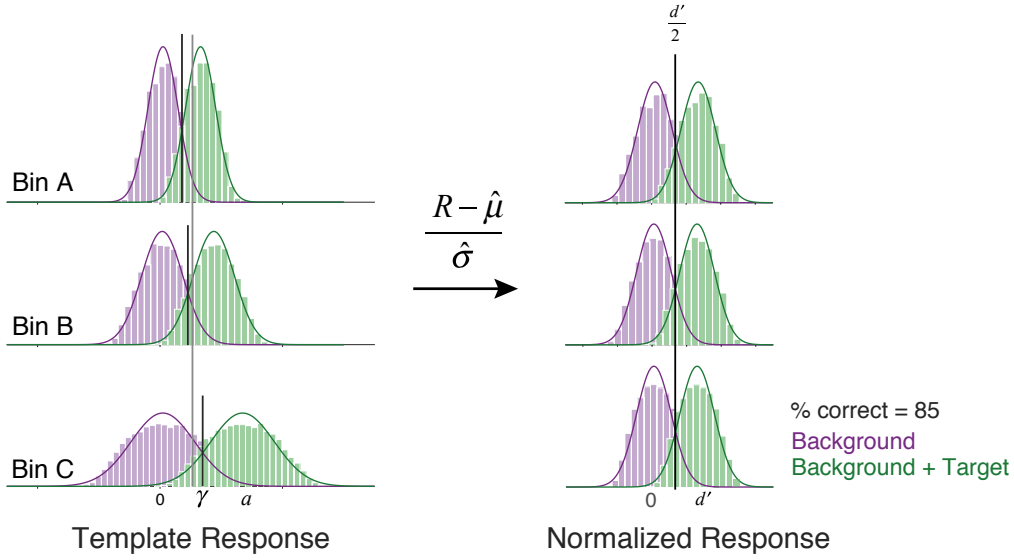

Figure 5.3: Normalized template response distributions. The distributions on the left show template responses for three bins in the image database. The gray line shows the best single criterion for detection across all bins. The distributions on the right are the same distributions, but normalized by their standard deviations. Here, only a single criterion is needed to perform optimally across all bins.

criterion corresponds to some intermediate compromise value. The value of single criterion is shown as the gray line in Figure 5.3. Although this criterion is optimal if only one criterion could be chosen, it is clear from the figure that is it substantially far from the optimal criterion in each bin. Thus, the simple matched template observer must perform worse in the uncertainty conditions.

The gray curves in Figure 4.1 show the difference in performance of the simple matched template observer in the blocked (the previous chapter) and uncertain (this chapter) conditions when the criterion is set at the best possible value for each accuracy level. It is clear that the human observers perform substantially better than what would be predicted from this simple observer. Thus, the results of this experiment suggest that human observers are not behaving like a standard matched template observer, but are more sophisticated. In effect, the human observers are able to dynamically adjust their decision criterion on each trial so that they are unaffected by the trial-to-trial changes in background bin and target amplitude.

5.4.2 Normalized matched template observer

How are the human visual and decision-making systems able to make this dynamic adjustment? One possibility is that they are able to estimate luminance, contrast and target similarity of the background (at the target location) on each trial, and then use those estimates to shift the criterion appropriately. Note that this is equivalent to normalizing the template response by the subtracting the mean and dividing by standard deviation implied by

the estimated luminance, contrast and similarity, and then use a fixed decision criterion. This model is shown in Figure 5.4.

The effect of properly normalizing the template responses is illustrated in Figure 5.3 (right). These distribution are the same as the distributions shown in the left of the figure, but they have been normalized by their standard deviation (which is given by the mean luminance, mean contrast, and mean similarity of their respective bins). Once normalized, the standard deviations all become 1.0 and separation between the distributions becomes the detectability (d') corresponding to 85% correct. After this normalization, optimal performance in Experiment 3 can be obtained with a single fixed criterion ($d'/2$) that depends on the accuracy level. Recall that the randomized experiment in this chapter was blocked by accuracy level, so observers could adopt a single criterion after normalizing the template response.

In the previous chapter, I showed that the standard deviation of the

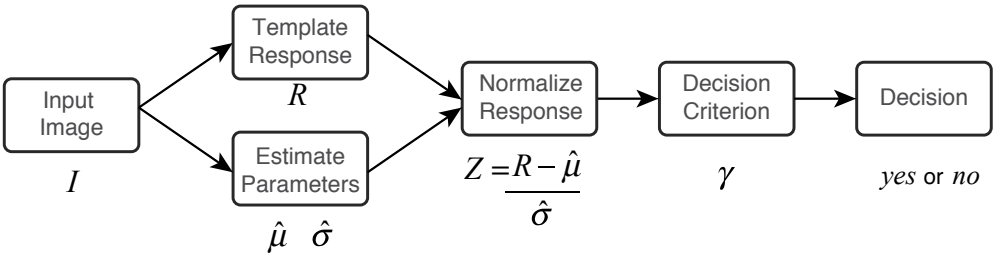

Figure 5.4: The normalized matched template observer model. The observer computes a template response, R, to an input image, then normalized this response by an estimate of the mean and standard deviation corresponding to the input image.

template response in a bin is a separable product of the luminance, contrast, and similarity for that bin (see Chapter 4, Section 4.5). The targets used in this study integrate to zero, and therefore the target absent distributions have a mean of zero. Thus, the response of the normalized matched template observer (Z) to a particular patch ($I(x,y)$) is given by

$$Z = \frac{\sum_{xy} T(x,y)I(x,y)}{k_0(\hat{L}+k_L)\sqrt{\hat{C}^2+k_C}(\hat{S}+k_S)}$$

where $\hat{L}$, $\hat{C}$, and $\hat{S}$ are the estimated luminance, contrast, and similarity for the image $I(x,y)$.

These three properties of the background image might the estimated from the background region surrounding the target region. This is plausible because the individual statistical properties of natural images are spatially correlated (nearby locations have similar contrasts for example). However, the experiment showed that human performance was unaffected by background and target amplitude uncertainty even when the surrounding region was removed by windowing. It is therefore also possible that these background properties could be estimated in the target region. This, however, might be more difficult because the estimates would be corrupted by the properties of the target on target present trials.

5.4.3 Estimation of local background properties

To evaluate the hypothesis that a matched template observer could remain unaffected by uncertainty by normalizing its response, I determined

how well the luminance, contrast, and similarity could be estimated by a simple linear model. Because there is only a weak correlation between luminance, contrast, and similarity in the natural environment, these three properties were estimated separability.

The performance of the normalized matched template observer depends on how accurately the properties of the background in the target region can be estimated. This is a potentially tricky problem because on each trial the observer did not know whether the target was present or absent. If the target was present it could bias the estimate of the background properties, thereby leading to a reduction in performance.

Two linear models for estimating the background properties were considered. The first model took into account the surrounding background context region and was only appropriate for Experiment 1 in Chapter 4. The second model only considered the background in the target region and therefore could be applied to either Experiment 1 or Experiment 2 in Chapter 4. In both cases, a separate linear weighting was learned for each background property. I trained the model by randomly sampling a large number of backgrounds from the entire space, and for half the samples, I added a target whose contrast was randomly sampled from a uniform probability distribution over a large range (0.01 to 0.35).

In the first model, I measured - for each training image - the value of each stimulus property at the target location and at 8 surrounding locations. These surrounding locations were each 0.8 degrees in size, and were

spaced without overlap around the center target region (Figure 5.5 top). I also measured the spatial similarity at the target region, whether or not the target was present in the training image. This spatial similarity measurement is described in Chapter 3. These measurements gave a vector of 10 numbers for each training image and background property (luminance, contrast, and similarity). Next, I applied linear regression to learn the 10 weights that best

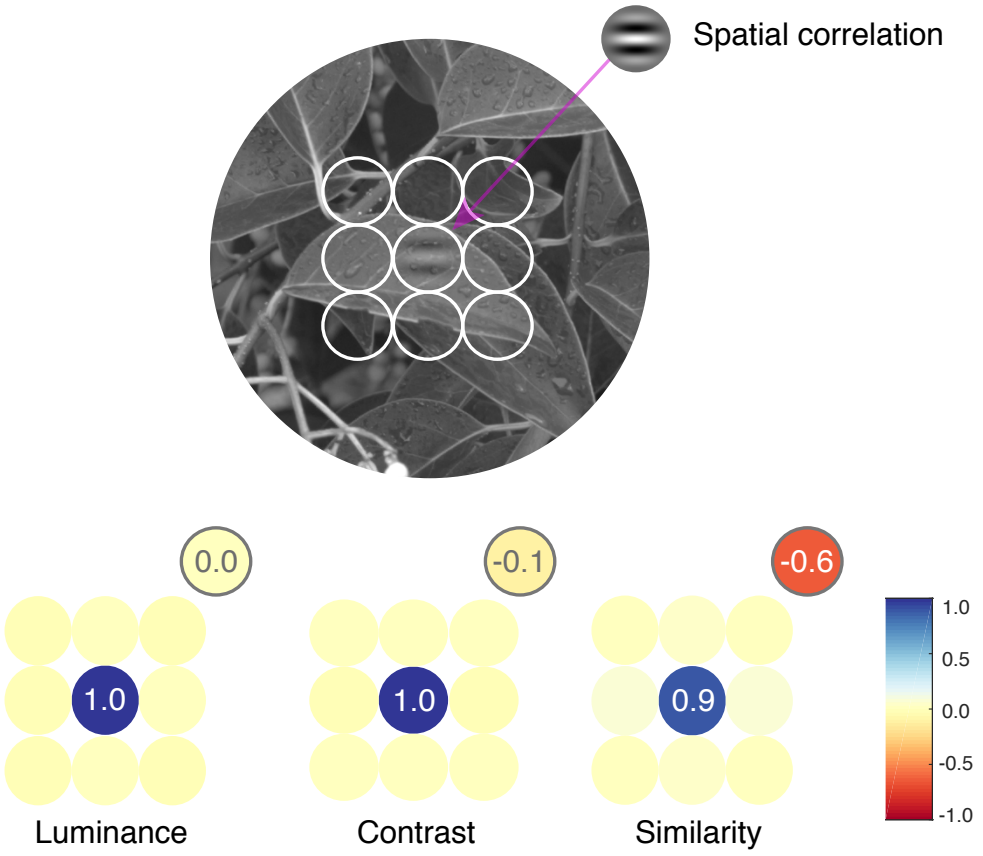

Figure 5.5: Linear weights for local background statistics estimation for the Gabor target.

predict the ground truth background property value at the target location.

Figure 5.5 shows the learned weights for the Gabor target for each of the three background properties. For each dimension, the most weight was put on the center (target) location and the second most on the template response. In fact, for luminance and contrast, the weights in the surrounding region were either zero or very close to zero. For similarity, the weights in the surrounding region were very close to zero, except for the region immediately next to the center in the x direction, where the weights were approximately 0.1. For contrast and similarity, the weight on the target correlation at the location of the target was negative (-0.1 for contrast, and -0.6 for similarity). This makes intuitive sense, because if the target present, then the target correlation would be high, and therefore the estimates of contrast and similarity should be down weighted. The weights for the plaid target were similar to those for the Gabor

In the second model, I measured - for the same set of training images as before - the value of each stimulus property only at the target location. I also measured the spatial similarity at the target location. Thus, there are only two weights to learn for each background property. This model returned weights that were similar to those of the first model. Later I will show that this second model has similar performance to the model with the surround, which suggests that the surrounding area is not needed to estimate the properties of the center area (even when targets are present on half the trials).

To assess how these estimates might be used for a matched template observer that normalized its response, I simulated the experimental trials in

experiment 3. On each trial, a template response was computed at the target location. Next, the linear weights learned on the training stimuli were used to estimate separately the luminance, $\hat{L}$, contrast, $\hat{C}$, and similarity $\hat{S}$ at the target region. These values were then plugged into the seperable equation to produce an estimate of the standard deviation of template responses, $\hat{sigma}$. Finally, the template response was normalized by this estimate, and compared to a fixed criterion (in this case $\gamma = d'/2$). If the response was greater than that criterion, the observer responded target present, if the response was less the observer responded target absent.

The black lines in Figure 5.2 shows the results from this simulation. It is clear that the normalized matched template observer's performance is not affected by uncertainty under the conditions measured. Also, the normalized matched template observer predicts the performance of human observers quite well. This suggests that the visual system is able to adjust its decision criterion (or normalize some internal response) on each trial.

The simple matched template observer was sufficient to predict the entire set of results presented in Chapter 4. However, it failed to predict the results presented here. Note that the normalized matched template observer would also have the same performance as the simple matched template observer in the previous experiments. Therefore, the normalized matched template observer predicts all the results in Chapters 4 and 5, and is a more general model of human performance in detection in natural images.

5.5 Summary

The results presented in this chapter showed that human observers are remarkably unaffected by large variations in background properties and target amplitude. This is a surprising result, given that in most cases uncertainty harms performance. However, in the natural environment, there is always uncertainty about the properties of the background at a potential target location. Because of this, it seems likely that the human visual system would have evolved to deal with this uncertainty.

I showed that a matched template observer that normalized its response by the separable product of the estimated local luminance, contrast, and similarity is able to perform as well as humans in this task. There is a large body of work that dealing with both luminance and contrast gain control mechanisms in the early visual system. This type of gain control is similar to the normalization of the matched template model. Therefore, it is possible that these rapid gain control mechanisms evolved to deal with the uncertainty found in the natural environment.

Chapter 6

Submasking

6.1 Motivation

The human visual system has a remarkable ability to detect and identify objects in the complex backgrounds found in the natural environment. In Chapter 4, I used constrained scene sampling (a method that was introduced in Chapter 3) to examine how several factors affect detection in natural backgrounds. I found that detection threshold amplitude for a particular target increased approximately linearly with local mean luminance, local RMS contrast, and with the similarity of the background to the target. Within a typical natural image, these three factors caused the threshold for detection to vary by more than a factor of 10. In Chapter 5, I found that detection thresholds were not affected by trial to trial variability in the target strength and background properties.

A simple detection model that was limited only by the variability found in natural scenes was strongly correlated with human performance. Furthermore, the entire set of results in Chapters 4 and 5 was predicted by a matched template model that normalized its response based on the statistical properties of the background under the target. The matched template observer relied on

the assumption that the statistical properties of the background underneath the target were uniform. However, in the natural environment this is almost always not the case, due to the complex spatial structure of natural images (Figure 6.1).

Does this complex structure affect detection thresholds? Subjectively, the answer appears to be yes. The subjects in Experiments 1 and 2 (Chapter 4) viewed thousands of trials, and each trial had a different background of the same luminance, contrast, and similarity within a narrow range. Despite controlling for these factors, the subjects reported that on a subset of trials the target was trivial to detect. These easier trials seemed to occur when the properties of the background under the target were inhomogeneous (such as in Figure 6.1, center). For example, if there was a lower contrast subregion, then the part of the target that fell into that subregion was much easier to detect, and often this visible part of the target was sufficient for identification. One potential reason for this was that the signal coming from the high contrast region was more strongly masked, and perhaps the visual system gave less weight to that location when determining whether or not the signal was present. Here, I refer to this phenomena as submasking.

Previous detection studies that used natural images as background stimuli have also noted this phenomenon. Chandler et al. (2009) and Alam et al. (2014) found that detection thresholds rose when there was more texture (as opposed to simple structure) underneath the target region. However, neither study quantified exactly what was meant by texture, and neither study

Figure 6.1: Examples of submasking. Left image: Submasking in the natural environment due to occlusion. Middle image: Example of submasking in Chapter 4, Experiments 1 and 2. Left image: Modulated white noise image used in the present experiment and analysis.

developed a theory for why detection thresholds should be higher in such backgrounds. Despite the long history of detection experiments in vision science, the effect of modulating background statistics underneath the target region (as shown in Figure 6.1 right) has not been studied directly.

The goal of the work presented in this chapter was to measure the effect of submasking on human observers in a simple case, and to develop a theory of how to optimally take advantage of variations in masking strength within the target region. Specifically, I measured the effect of modulating the contrast of the background underneath the target. First, I created a simple stimulus using Gaussian white noise that could be used to measure this effect in human observers. Next, I derived an observer that optimally exploits the variation in the reliability of the information underneath the target. Finally, I measured detection thresholds on human observers, and compared the results to the

predictions of the ideal observer. I found a significant effect of modulating the contrast underneath the target (in agreement with the subjective reports from my previous experiments). This effect was predicted by an ideal observer that normalizes its response by an estimate of the contrast at each location under the target.

6.2 Modulated white noise

In Chapters 4 and 5, I used constrained scene sampling to measure the effect of luminance, contrast, and similarity on detection thresholds in natural images. These factors are known to affect thresholds in simple stimuli, and these statistics are easily computed in natural stimuli. The effect of background changes underneath the target, however, has not been previously studied with simple background stimuli. Furthermore, images in the natural environment have structure that varies in complex ways which makes it difficult to parameterize this structure.

Because of this, I used a new simple stimulus designed specifically to test the effect of background statistical modulation underneath the target in a detection task. Gaussian white noise patterns are a common background used in detection experiments. The luminance at each pixel location is sampled from a Gaussian distribution of a particular mean and standard deviation. The contrast of the background is determined by the standard deviation of the Gaussian from which the image pixels were sampled. Previous studies use backgrounds where every pixel was sampled from a single Gaussian distribu-

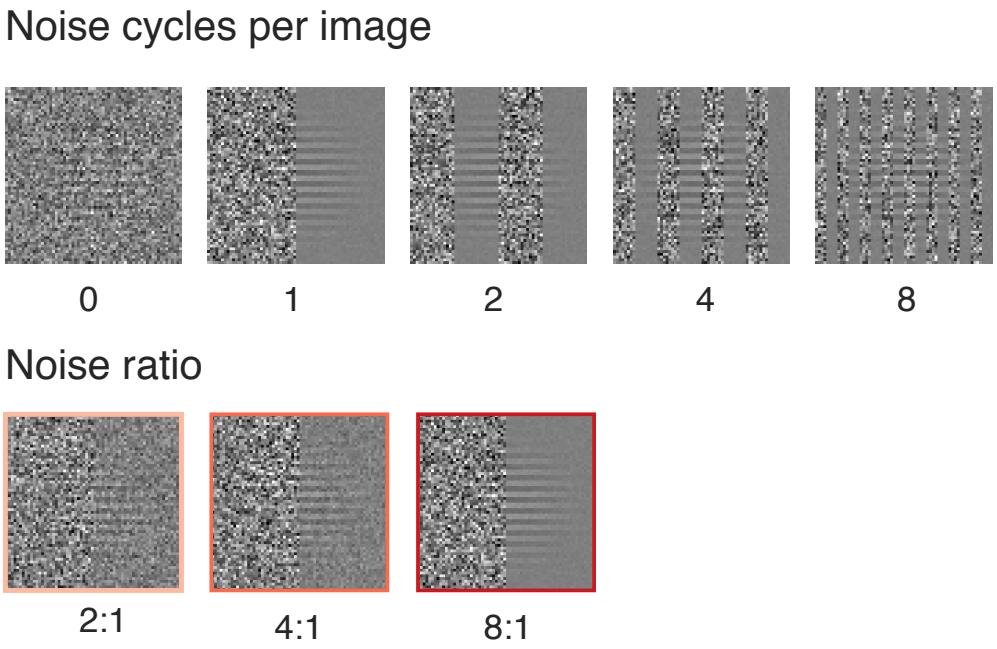

Figure 6.2: Example patches of modulated white noise used in the analysis and experiment. Each background has the same overall contrast energy. Every background contains the target (an 8 cpd square wave windowed with a cosine window). In the experiment the backgrounds were 2.13 degrees of visual angle. The colored outlines will be used to key subsequent data figures.

tion. In other words, the contrast energy of the background was uniformly distributed underneath the target.

In the models and experiments presented in this chapter, I use a modification of the Gaussian white noise stimulus, called amplitude-modulated white noise. As in previous experiments, each pixel was sampled from a Gaussian distribution of a particular standard deviation. However, in some regions of the background noise pattern, pixels were sampled from a Gaussian with a

high standard deviation, and some were sampled from a Gaussian with a low standard deviation. The high and low regions alternated in vertical stripes from 0 noise cycles per image (which is uniform Gaussian white noise) to 8 noise cycles per image (shown in the top row of Figure 6.2).

The ratio between the high and low standard deviations was also varied between 2 to 1 and 8 to 1 (shown in the bottom row of Figure 6.2. In all cases, however, the average variance (and therefore the total contrast energy) in the background was the same (here the overall contrast was fixed at 0.246). As a baseline, I also measured the thresholds for the standard case where the ratio between low and high noise was 1 to 1 (first panel in the upper row of Figure 6.2). Because the overall contrast was the same in each condition, a standard matched template observer (which integrates over the entire image) would give on average the same response to all of the stimuli in Figure 6.2, and hence the would have the same threshold. Subjectively, this is not the case, as the target is much easier to detect in the high noise ratio condition. In the following sections, I develop an ideal observer for detection in modulated white noise patterns, and compare the performance of this observer to human detection thresholds.

6.3 Reliability weighted matched template observer

How does an ideal observer use the changing background information in a detection task? Consider an observer tasked with detecting a target $T(x, y)$ of amplitude a in Gaussian white noise. The ideal observer in this case is

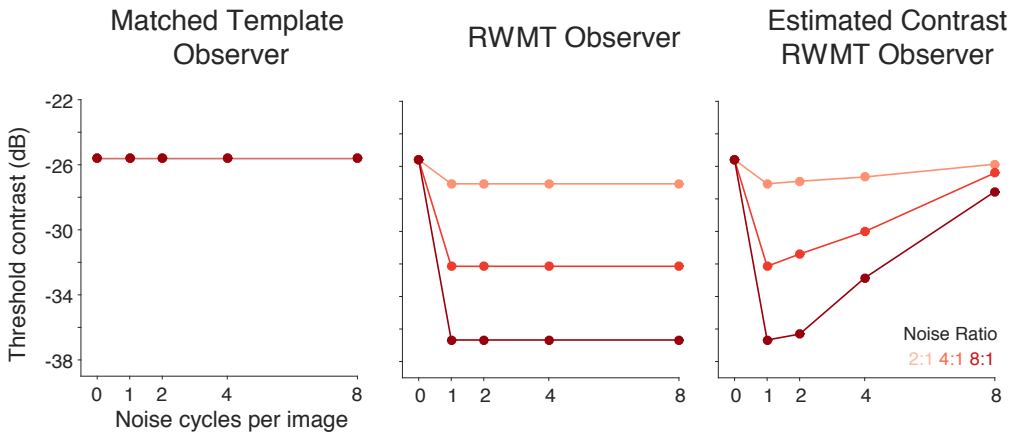

Figure 6.3: Model observer predictions for modulated white noise. Each plot shows predictions from different models observers. Contrast threshold (in dB) is plotted on the y axis, and the number of noise cycles per image is plotted on the x axis. The different shades of red represent different noise ratios in the modulated white noise backgrounds. See text for a description of each model.

the matched template observer. This observer first computes a dot product between the input image, $I(x, y)$, and the template of the target, $T(x, y)$,

$$Resp = \sum_{x,y} I(x, y)T(x, y)$$

This response is then compared to a criterion (equal to $\frac{aT \cdot T}{2}$), to determine whether or not the target was present. The target used in this analysis and experiment was an 8 cycle per degree, horizontally oriented square wave that was windowed with a cosine window. The diameter of the cosine window was 1.7 degrees at half height.

Because this observer integrates the dot product over the entire image, the predicted detection thresholds will always be the same in images where the overall contrast energy is equal. The right most plot in Figure 6.3 shows the

predicted thresholds for a standard matched template observer in modulated white noise. On the x-axis is the noise cycles per image, and the different shades of red represent the different noise ratios. On the y axis is the predicted threshold contrast (shown here is in decibels (dB) where $c_{dB} = 20\log_{10}(c)$). Not surprisingly the standard matched template observer (which is optimal in Gaussian white noise) predicts no improvement as the ratio between the high and low noise regions increases.

An ideal detector in modulated white noise should take into account the reliability of the template response at every pixel location. Target regions that fall into areas with high contrast are less reliable than regions that fall into areas of low contrast. In the detection task in modulated white noise, the observer receives an input image, $I(x, y)$ on each trial:

$$I(x,y) = \begin{cases} N(x,y), & \text{if target absent} \\ aT(x,y) + N(x,y), & \text{if target present} \end{cases}$$

where $T(x, y)$ is the target signal, and $N(x, y)$ is a background of modulated Gaussian noise, with mean μ, and a variance of $\sigma^2(x, y)$. This variance corresponds to the average local contrast energy (the square of the RMS contrast) of the background at location (x, y).

The ideal observer should choose the category (*signal + noise* or *noise alone*) with the highest probability given the input image (i.e. if $p(S|I) > p(N|I)$ choose signal present). Using Bayes' rule, this decision rule reduces to

$$if \ \frac{p(I|S)p(S)}{p(I|N)p(N)} > 1 \text{ choose present } else \text{ choose absent}$$

where $p(I|S)$ and $p(I|N)$ are the likelihoods of the image given the signal + noise and the noise alone respectively, and $p(S)$ and $p(N)$ are the prior probabilities of the signal + noise and the noise alone, respectively. Taking the natural logarithm of the equation and rearranging gives:

$$if\ X > \gamma \text{ choose present } else \text{ choose absent}$$

where the decision variable X is the ratio of the log likelihoods, $X = \ln \frac{p(I|S)}{p(I|N)}$, and γ is the ratio of the log priors, $\gamma = \ln \frac{p(S)}{p(N)}$. In the case where the prior probabilities of the signal and the noise trials are equal, γ becomes 0.

For the case of Gaussian white noise, the likelihood distribution of the input image given noise alone is a Gaussian with mean μ and standard deviation $\sigma(x, y)$:

$$p(I|N) = \frac{1}{\sqrt{2\sigma^2(x,y)\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2(x,y)}}$$

. The likelihood distribution of the input image given the signal plus noise is a Gaussian with mean $aT(x, y) + \mu$ and standard deviation $\sigma(x, y)$:

$$p(I|S) = \frac{1}{\sqrt{2\sigma^2(x,y)\pi}} e^{-\frac{(x-(aT(x,y)+\mu))^2}{2\sigma^2(x,y)}}$$

Again, it is important to note here that the standard deviation varies at each point, (x, y), in modulated white noise. In standard Gaussian white noise $\sigma(x, y)$ is a single number.

Plugging the equation of a Gaussian for the likelihoods into the previous equation for X shows that the decision variable is

$$X = \sum_{x,y} -\frac{[I(x,y) - aT(x,y) - \mu]^2}{2\sigma^2(x,y)} + \frac{[I(x,y) - \mu]^2}{2\sigma^2(x,y)}$$

Simplifying this equation further gives

$$X = a \sum_{x,y} \frac{T(x,y)\Delta I(x,y)}{\sigma^2(x,y)} - \frac{a^2}{2} \sum_{x,y} \frac{T^2(x,y)}{\sigma^2(x,y)}$$

where $\Delta I(x,y) = I(x,y) - \mu$. If the prior probabilities of the signal and the noise trials are equal, then the optimal decision rule becomes

$$if \ \frac{1}{\sum_{x,y} \frac{T^2(x,y)}{\sigma^2(x,y)}} \sum_{x,y} \frac{T(x,y)}{\sigma^2(x,y)} \Delta I(x,y) > \frac{a}{2} \ else \text{ choose absent}$$

Therefore, an ideal observer in modulated white noise scales the template values by 1 over the variance (the reliability) of the background at each location, and then applies this scaled template to the input image (minus the mean luminance of the image). This response is then divided by the energy of the reliability weighted template and the result compared to criterion that is half the amplitude of the target. In the case where the contrast is fixed underneath the target, this observer reduces to the standard matched template observer, because the $\sigma^2(x,y)$ terms cancel out.

Detection thresholds for a reliability weighted matched template (RWMT) observer that knows exactly the standard deviation of the underling noise distribution at every pixel is shown in the middle plot in Figure 6.3. Unlike the standard matched template observer, the RWMT observer's thresholds improve as the noise modulation increases. The thresholds remain constant above 0 cycles per image, because this observer has perfect knowledge of the noise standard deviation at each pixel location.

Of course, it is not realistic to expect any visual system to have perfect information regarding the underlying noise distributions. Rather, a real visual system would have to estimate the contrast at each location. Thus, I consider a near-ideal observer where the noise variance is estimated at each image location. Specifically, we assume that the variance is computed under a raised-cosine window with radius ρ:

$$w(x,y) = \begin{cases} 0.5 + 0.5\cos(\pi\sqrt{x^2+y^2}/2), & \sqrt{x^2+y^2} < \rho \\ 0, & \sqrt{x^2+y^2} \geq \rho \end{cases}$$

For this observer, the optimal decision rule is the same as before, but instead estimates of the variance at each points are used in the decision rule:

$$if \ \frac{1}{\sum\limits_{x,y} \frac{T^2(x,y)}{\hat{\sigma}^2(x,y)}} \sum_{x,y} \frac{T(x,y)}{\hat{\sigma}^2(x,y)} \Delta I(x,y) > \frac{a}{2} \ else \text{ choose absent}$$

Experimental thresholds for this observer with a fixed window width ρ are shown in the leftmost plot in Figure 6.3. Like the previous model, this model shows an improvement with increasing noise ratios. However, thresholds increase as the number of noise cycles in the image increase. This is because as the regions of low and high noise get closer together, the estimate of the underlying variance becomes inaccurate. Changing the window width for the variance estimation would change this pattern of thresholds. In the following sections, I show the results from detection experiments in modulated white noise in human observers, and compare these results to this near-ideal observer.

6.4 Experimental procedure

Three experienced psychophysical observers participated in the present experiment. All subjects had normal or corrected to normal vision. Stimuli were presented on a SONY GDM-FW900 monitor at a resolution of 30 pixels per degree. Each stimulus was first generated in software, and then gamma compressed (based on the same calibration used in the experiments shown in Chapters 4 and 5). Finally, the stimulus was quantized to 256 10-bit gray levels and displayed. Stimulus presentation and response collection were programmed in MatLab, using PsychToolbox (Brainard 1997, Pelli 1997, Kleiner et al. 2007).

The trial sequence for the present experiment is shown in Figure 6.4. On each trial in the experiment, a low contrast bounding box was presented in the center of the screen for 500 ms. The rest of the screen was set to the mean luminance of the modulated white noise pattern (48 cd/m^2). Next, the stimulus appeared inside the bounding box for 250 ms. Randomly, on half the trials the target was added to the noise background. Subjects responded with a button press to indicate whether or not the target was present. Auditory feedback was given after every trial.

Trials were blocked by target contrast (amplitude), noise modulation ratio, and the number of noise cycles per image. In order to help subjects set their decision criterion, the first trial of each block always contained the target stimulus. This trial was not included in the analysis of the data. To reduce adaptation effects, the low and high noise regions were reversed between each

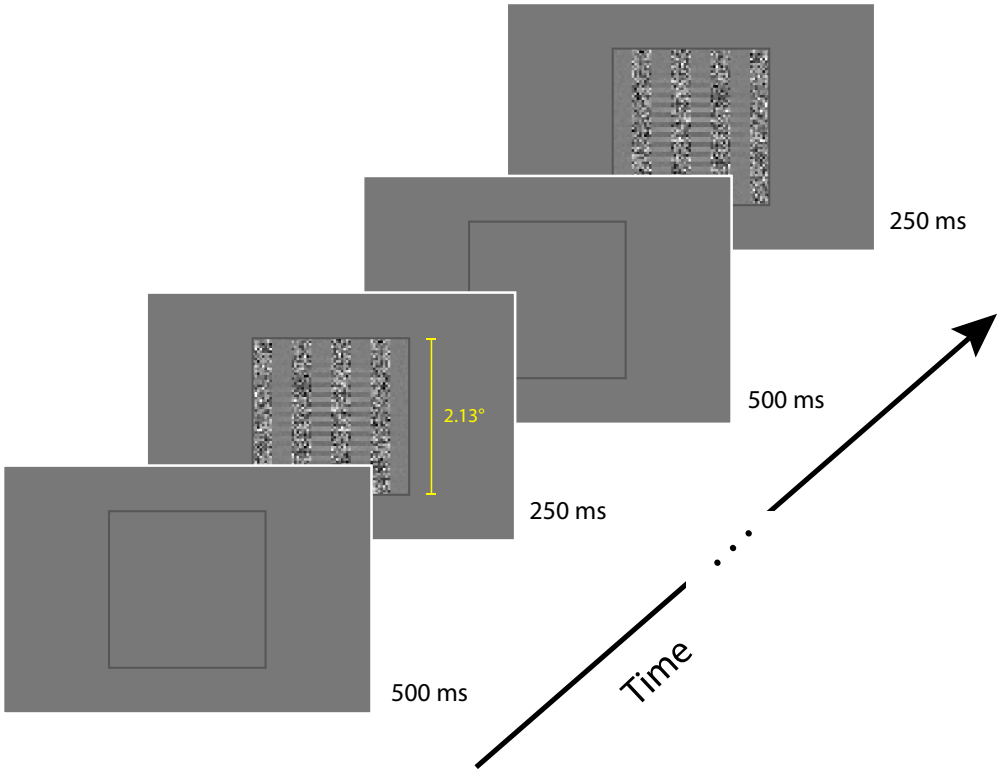

Figure 6.4: Trial sequence for submasking experiment (Experiment 4).

trial (as shown in Figure 6.4).

For each experimental condition, five target contrast levels were measured. Each level contained 72 trials measured in blocks of 36 trials. Subject responses were fit with the same psychometric function used in Experiments 1 and 2 from Chapter 4. For each subject only one slope parameter (β) was fit, and the bias parameter (γ) was set to zero. The slope parameters for each subject are shown in Table 6.1.

6.5 Results

The goal of this experiment was to measure how well human observers exploit the variation of reliability in simple stimuli, and the spatial scales over which the visual system estimates reliability in these stimuli. The rationale for measuring spatial scale is that the visual system may have evolved to estimate reliability at specific spatial scales.

To measure the effect of changes in contrast under the target, I measured detection thresholds in modulated white noise where the ratio of low to high noise in each condition was 2 to 1, 4 to 1, and 8 to 1. For each noise ratio, the noise cycles per image was varied: 1, 2, 4, and 8. These manipulations made it possible to estimate the spatial scale over which reliability was estimated in the human observers. The average noise power of the background

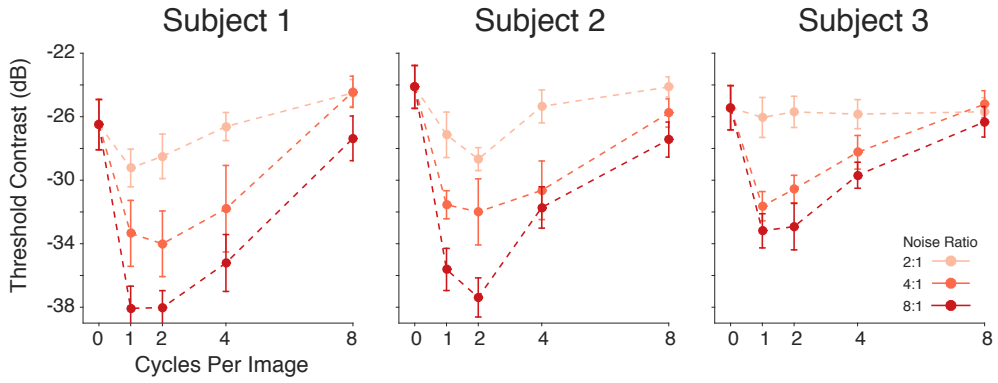

Figure 6.5: Modulated white noise experimental results. Each plot shows the threshold contrast (in db) for each subject. The dotted lines connect each point for each noise ratio (shown as different shades of red). Error bars represent bootstrapped standard errors.

was fixed at 0.07 across all conditions (thus a simple matched template detector would have identical thresholds for all conditions). The mean luminance was also held fixed across all conditions at 48 cd/m^2. To get a baseline for performance, I also measured thresholds in a white noise pattern where each pixel was sampled from a single Gaussian distribution.

The experimental results for the three subjects are shown in Figure 6.5. Each panel plots threshold contrast (in decibels) as a function of the noise cycles per image for one of the subjects. The different shades of red correspond to the noise modulation ratios. The point at 0 cycles per image is the threshold for classic white noise background where the contrast is fixed underneath the target.

While there are some individual differences, the effect of modulating the contrast of the background under the target was substantial for each subject. For all subjects and all noise ratios threshold is a u-shaped function of noise cycles per image that reaches a minimum in the range of 1-2 noise cycles per image. As modulation ratio increases, the minimum threshold declines. The maximum effect is approximately 8.5 dB (a drop in threshold by a factor of 7.1) for a modulation ratio of 8. This makes intuitive sense because it is

Slope (β) parameters in the modulated white noise experiment	
Subject	β
Subject 1	1.3021
Subject 2	1.7987
Subject 3	1.9919

Table 6.1: Slope parameters in the modulated white noise experiment

subjectively easiest to detect the target when there are regions of low contrast in the image.

Qualitatively, the pattern of thresholds look most like the reliability weighted observer that estimated the variance at each point (described in the previous section and shown in Figure 6.3). But how well are these results

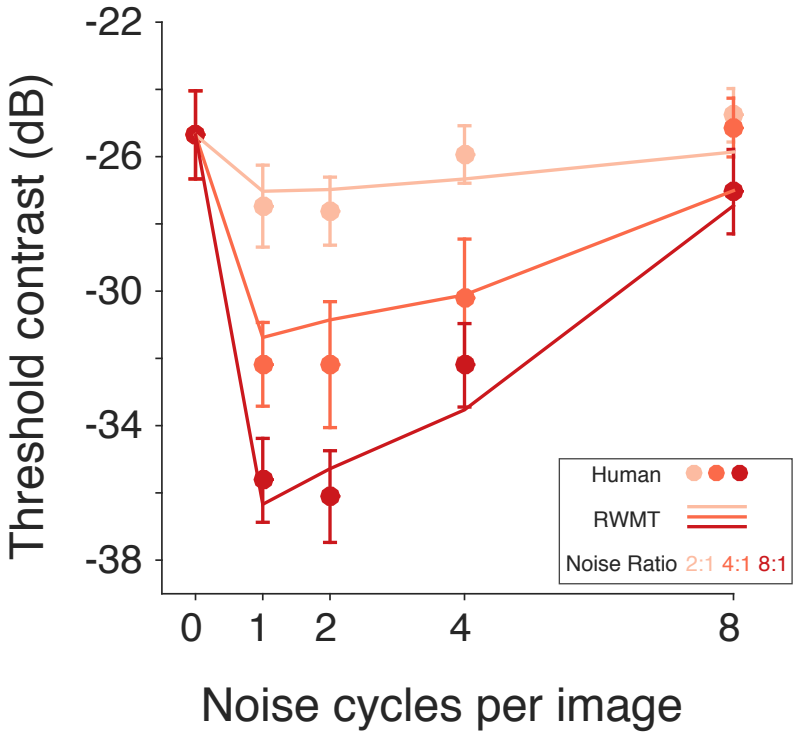

Figure 6.6: Modulated white noise experimental results (averaged across three subjects) and the reliability weighted matched template observer. The RWMT observer accounts for the pattern in thresholds, which suggests that the visual system is using the reliability of the target signal at each location to detect the target. The observer shown estimated the contrast using a window of 0.43 degrees in width.

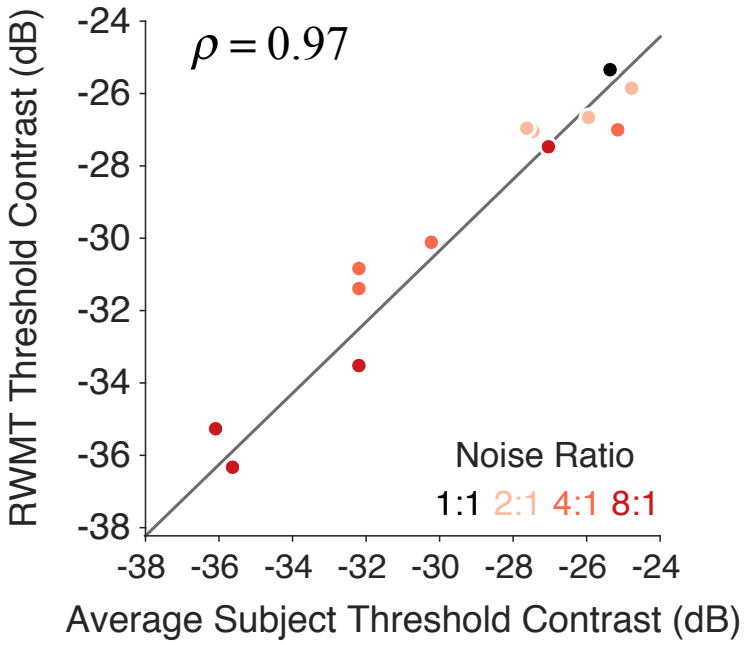

Figure 6.7: Plot of correlation between the thresholds of the RWMT observer vs. human subjects. Each point represents the average contrast threshold (in dB) averaged across the three subjects. The correlation was 0.97.

predicted by such an observer? Recall that the thresholds of the reliability weighted observer depend on the spatial scale over which it estimates the local contrast at each point. Furthermore, as mentioned in Chapter 4, an ideal observer usually performs much better overall than human observers and a standard way to account for this is to apply an efficiency parameter (η) to the model thresholds. Here, the efficiency parameter simply translates all the predicted thresholds vertically on the dB scale.

I fit the reliability weighted matched template observer to the human

data by minimizing the mean squared error between the model estimates and the human data by adjusting the size of the estimation window. The efficiency was set so that the model and human thresholds at 0 noise cycles per image matched (here $\eta = 0.1797$). Figure 6.6 shows the model fit (solid lines) and the average thresholds of the three subjects. The RWMT observer captured both the improvement in thresholds with increasing noise ratio, as well as the decrease in performance with increasing number of noise cycles in the image. The window width parameter that best fit the data was 0.43 degrees of visual angle. The RWMT observer thresholds are shown as a function of the average subject thresholds in Figure 6.7. The correlation between the model and the subjects threshold is 0.97.

6.6 Summary

The work in Chapters 4 and 5 (along with almost every previous study of pattern masking) assumed that the properties of the background underneath the target were uniformly distributed. Of course, in the natural environment, that is often not the case due to the complex structure of natural images. In this chapter, I studied the effect of background changes in contrast underneath the target on detection. I found that varying the contrast underneath the target signal had a substantial effect on detectability in simple white noise backgrounds. I derived an ideal observer for these stimuli that computed the response of a template of the target weighted at each point by an estimate of the inverse noise contrast energy (or reliability) at that location. The pattern

of human detection performance was predicted by the ideal observer, which suggests that the human visual system is able to use the varying properties of the background under the target in an near optimal way.

Chapter 7

Discussion

7.1 Summary

I used constrained scene sampling to examine the factors that affect detection of known targets in natural backgrounds. Background patches from a database of calibrated natural images were sorted into bins according to their mean luminance, RMS contrast, and (phase-invariant) similarity to the target. To test the effect of these three dimensions on detection thresholds, I measured psychometric functions in a single-interval forced choice experiment in a subset of the bins. Amplitude thresholds increased approximately linearly along all three dimensions, both when the background region extended well beyond the target region, and when the background was windowed to the size of the target region. A simple matched template detector (with a single efficiency parameter) predicted the entire set of results, including the case where the surrounding region was removed by windowing with a raised cosine envelope.

The experiments in Chapter 4 had a different image background on each trial, but were blocked by the bin the images were sample from, and the target amplitude. In Chapter 5, I examined the effects of background and

amplitude uncertainty by randomly sampling a background from a randomly selected contrast and similarity bin on each trial. The amplitude of the target (if present) also depended on the randomly selected bin. I found that under these conditions subject performance was essentially the same as in the blocked experiments, and therefore there was no effect of uncertainty. These results could not be explained by a simple matched template observer, but could be explained by a normalized matched template observer. This observer estimated the background luminance, contrast, and similarity in the target region via linear weighted summation of the local measurements of each property in the target region (or both the target and surrounding region). This normalized template observer quantitatively accounted for the results from not only the experiment in Chapter 5, but also the experiments in Chapter 4. Thus, human thresholds in natural backgrounds were accurately predicted from first principles.

Although the matched template model described the entire set of results in Chapters 4 and 5, there was (at least subjectively) at least one other important factor for detection that was not captured by the model. Subjects noted that detection was trivial in cases where only a portion of target fell on a region of background with low contrast, luminance, or similarity (an effect I referred to as submasking). In order to test this effect, I designed a stimulus where bands of low and high contrast white noise alternated. I developed an ideal observer (reliability weighted matched template observer) for this case, and showed that this observer predicted the pattern of human detection thresh-

olds. This suggests that submasking is an important factor in detection in the natural environment, and that human observers are able to use the reliability of signal information at each point in the image in an approximately optimal way.

7.2 Shape of the template response distributions

The histograms of the template responses in each bin were approximately Gaussian for all bins and both targets. The average excess kurtosis (a Gaussian has an excess kurtosis of 0), was -0.01 for the Gabor target, and 0.07 for the Plaid target. So on average, the template response distributions were very close to Gaussian in their kurtosis. This is perhaps surprising, because a classic finding is that the response distributions of oriented Gabor filters (like the Gabor target used in my experiments) to natural images are highly non-Gaussian, with sharp peaks at zero and heavy tails (Field 1987, Daugman 1989). One hypothesis is that this is due to the higher-order structure (contours, edges, lines, etc.) in natural images. However, all of the patches in any bin in the image database contained such structure. Thus, this result suggests instead that the heavy-tailed distributions arise from the mixture of standard deviations from the different bins (a mixture-of-Gaussians model that does not depend on the local phase structure of natural images). Also, when the distributions of template responses were normalized by the patch luminance, contrast, and similarity, then they became Gaussian with the same variance ($\sigma^2 = 1$) for all bins. To the extent that cortical neuron

responses are consistent with such normalization, they do not provide a sparse code by producing a heavy-tailed distribution of responses to natural images (Olshausen & Field 1997).

7.3 Detection in the real world

In the experiments in Chapter 4 (Detection in natural scenes), stimulus uncertainty was minimized by blocking both the target amplitude and the bin from which the background was sampled. In Chapter 5 (Uncertainty in natural scenes), uncertainty was increased, but was still constrained by blocking trials to a fixed level of performance, and therefore detectability (d'). Under these circumstances, target amplitude and background properties co-varied in a way that allowed the normalized matched template observer to adopt a single optimal fixed criterion for the block ($\gamma = \frac{d'}{2}$).

However, in the real world there is generally no reason to expect the amplitude and the background properties to covary. Nonetheless, the normalized matched template observer would still support a simple optimal decision strategy. Under conditions where amplitude is completely unconstrained, a rational strategy would be to choose a decision criterion that gives a desired false alarm rate. This is the same strategy that is used in standard statistical tests. For the normalized matched template observer this corresponds to placing the criterion at a fixed value. This in turn maximizes the hit rate for a given false alarm rate. For example, a criterion of 1.65 gives a false alarm rate of 5%, independent of target amplitude and background properties. There is

no such fixed criterion for the simple matched template observer.

In this dissertation, I only considered detection under background and target amplitude uncertainty; however, the normalized matched template observer would also be appropriate for other forms of target uncertainty, such as location or orientation uncertainty. In these cases, the normalized matched template would be applied over the region of uncertainty and the decision criterion would be applied to the maximum of the normalized template responses. Unlike amplitude uncertainty, these forms of target uncertainty usually cause a substantial unavoidable decrease in accuracy (Green & Swets 1966, Pelli 1985, Burge & Geisler 2011, Geisler 2011).

7.4 Stimulus uncertainty and gain control

The classic laws of detection - increases in threshold with background luminance, contrast and similarity to the target - were primarily discovered and then explored using simple backgrounds that did not randomly vary from trial to trial (Mueller 1951, Campbell & Kulikowski 1966, Legge & Foley 1980). Furthermore, the effects observed with these fixed backgrounds are similar to those I found in Chapter 4. On the surface, this fact seems puzzling. A matched template detector for backgrounds that do not vary from on each trial, will always perform perfectly, independent of background luminance, contrast, or similarity. This is because the template response is always the same when the target is present, and always the same when the target is absent. So, why should there be a close relationship between the thresholds

obtained with random (natural) backgrounds (that have complex structure) and those obtained with fixed backgrounds (like a uniform luminance field or a grating)?

The answer most likely lies in the fact that the visual system evolved to operate under conditions of high stimulus uncertainty. In the natural environment, both the background and the amplitude of the target (if it is present) are generally different on every occasion. What the scene statistics measurements and modeling in my dissertation show is that the detrimental effects of this uncertainty can be optimally reduced by dividing the template response by the product of background luminance, contrast, and similarity (see Chapter 5, section 5.4.2 for details of this model).

This sort of normalization (or gain-control) for luminance and contrast is observed in the early visual system. Because the visual system is almost always performing detection under uncertainty, it is reasonable to expect evolution to have placed the adjustments for this uncertainty into the early, automatic levels of the visual system. However, the side effect of this is that under laboratory conditions where the background does not change from trial-to-trial (and therefore there is no uncertainty), these gain control mechanisms lead to highly sub-optimal performance, because the gain control reduces the signal level relative to subsequent neural-processing and decision noise.

Undoubtedly, if our ancestors had lived in a simple environment with just a few specific backgrounds, then our visual system would have evolved a very different solution, (such as estimating which of the few possible back-

grounds is present and then subtracting it from the input). Here, I argue that the rapid and local neural gain control mechanisms, and the psychophysical laws of detection and masking, are most likely the result of having evolved a near optimal solution to detection in natural backgrounds under conditions of high uncertainty.

A standard explanation for early gain-control mechanisms is to keep the responses of the neurons encoding the stimulus within the neurons dynamic range. This must be true for the slow changes in gain that occur with changes in ambient light level that typically varies slowly over many orders of magnitude (cameras also adjust their overall gain based on ambient light level). However, there are also gain control mechanisms that adjust rapidly to the local luminance, contrast (Albrecht & Geisler 1991, Heeger 1991, Hood 1998, Carandini & Heeger 2012), and perhaps similarity.

Indeed, to be useful, these gain control mechanisms must adjust nearly instantly (within a few tens of ms) because our eyes are in constant motion and local image statistics are largely uncorrelated across fixations (Mante et al. 2005, Frazor & Geisler 2006). It is these rapid gain control mechanisms that would seem to be optimal for detection when fixating around a given natural scene (the typical variation of luminance and contrast within a scene is somewhat more than one order of magnitude).

7.5 Constrained scene sampling

The constrained scene sampling approach described in Chapter 3 might prove useful for uncovering important principles of other natural tasks. The crucial requirements for constrained scene sampling are to have a large collection of relevant natural signals and to have hypotheses (or prior evidence) about what stimulus dimensions are likely to strongly influence task performance. A useful benefit of randomly sampling from the histogram bins without replacement is that, for each bin, the subjects make responses to a large number of different stimuli that are controlled simultaneously along the dimensions of interest. This makes it possible to analyze the stimuli and responses within a bin to discover other potential factors contributing to human and model observer performance.

7.6 Submasking in natural images

Many previous models for detection (including the matched template observers in Chapters 4 and 5) assume that the statistics of the background underneath the target are constant. In Chapter 6, I measured detection performance in background where the contrast changed underneath the target (an effect I called submasking). The pattern of human observer thresholds was strongly correlated with an ideal observer that took into account the reliability of the information at each pixel location. This suggests that the visual system is able to take advantage of the local background properties underneath the target in a near optimal way.

The natural images in the constrained scene sampling experiments often had inhomogeneous luminance, contrast, and similarity underneath the target region. Despite this, when averaged across trials, a matched template observer that integrated its response across the entire target region captured much of the pattern in human data. It is possible, however, that a reliability matched template observer (as described in Chapter 6) would better predict the trial-by-trial responses of human observers to the natural image backgrounds. Such an observer would normalize at each location by an estimate of the luminance, contrast, and similarity.

In Chapter 6, I only showed submasking for contrast, and fit the human data with one contrast estimation window. However, it is possible (and perhaps likely) that optimal estimation window for estimating the luminance, contrast, and similarity would be of different size. This would not only provide a potentially better trial-by-trial model of human behavior, but would also provide insight into the neural mechanisms underlying detection.

7.7 Image quality assessment

There has been extensive work in the engineering field of image processing concerning the human visual system's ability to detect any change in an image. This knowledge is of course useful for many practical applications. For example, when building an image compression algorithm any image artifacts left after compression that are below or near the detection limit are acceptable. These approaches are similar to other detection experiments in the sense that

the image is the mask and the artifact added to the image is the target.

An algorithm that produces a measure of image quality automatically is of great use in the field of image processing. Interestingly, in the image processing literature one of the most successful image quality metrics involves a combination of luminance, contrast, and similarity (Wang et al. 2004). The results of experiments in Chapter 4 suggest that the success of this metric has direct basis in the statistical properties of natural images.

Appendices

Appendix A

Standard deviation of template response distributions

A.1 Template responses for windowed image patches

In Chapter 4, I showed a plot of the standard deviation of template response (for non-windowed patches) as a function of luminance, contrast, and similarity (Figure 4.6 on page 70) . A similar figure is shown below (Figure A.1). For the windowed patches the standard deviation of template response was lower than for non-windowed patches. The same linear model that was fit to the full surround case was fit to these data. Here, the four parameters were found to be $k_0 = 0.0177$, $k_L = -0.1514$, $k_C = -0.0139$, $k_S = -0.0397$.

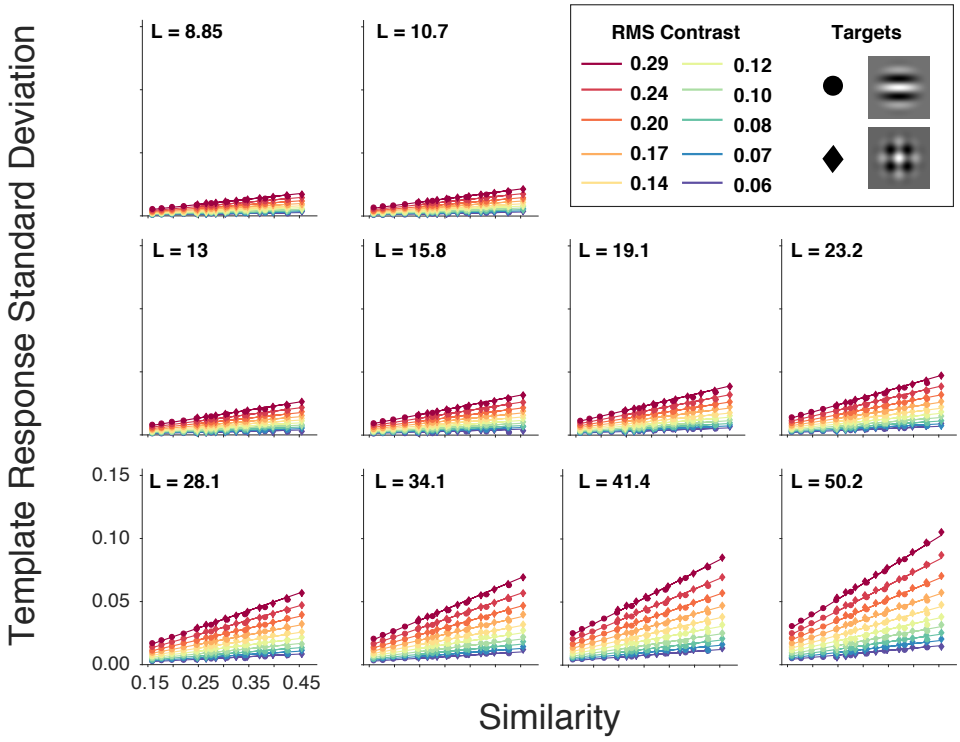

Figure A.1: Effect of background luminance, contrast, and similarity on the standard deviation of the template response for windowed patches.

Appendix B

Additional natural image statistics

B.1 Correlation between the center and surround values

The extent to which the center value for luminance, contrast, or similarity can be estimated from the surrounding values stongly depends on the correlation between the center and surrounding values. Each subplot in Figure B.1 shows the correlation of the value at 9 surrounding locations and the center value. It is clear from these plots that the center values for luminance have the highest correlation with the center value (the average correlation is 0.8). The surrounding values for contrast have a correlation with the center contrast of 0.47 on average. For both the Gabor and plaid targets, the correlation is weak between the center and surrounding similarities (on average the correlation is 0.3 for both targets).

B.2 Range of similarity values in natural images

Previous studies have shown that luminance and contrast in a typical natural image vary by around an order of magnitude (Mante et al. 2005, Frazor & Geisler 2006). However, no study has investigated the magnitude that similarity for specific targets varies in natural images. Figure B.2 shows two

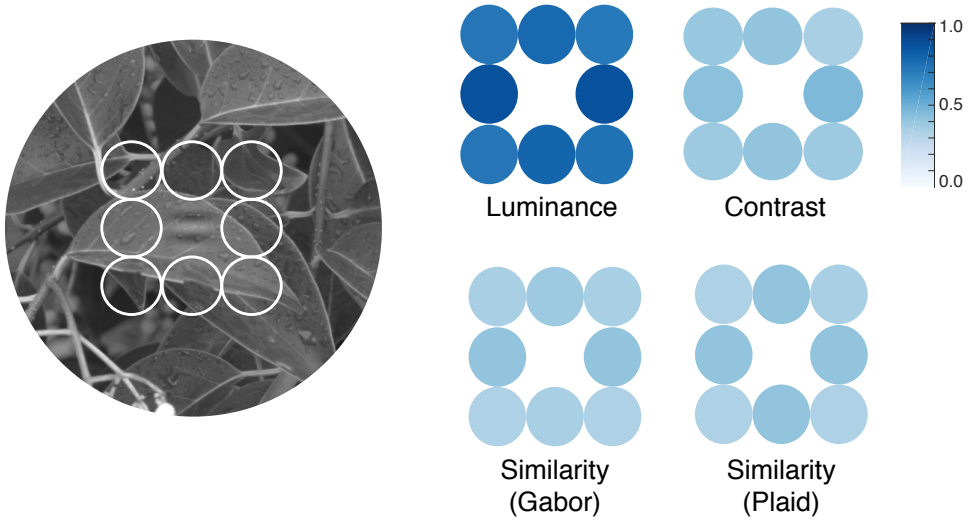

Figure B.1: Correlation between the values of the center and surrounding area in the natural image database.

histograms (one for each target) of the ratio between the maximum and minimum similarity for each of the 1,204 images in the database. The average ratio between the minimum and maximum similarity for the Gabor target is around 8, and the average ratio for the plaid is around 6.

B.3 Shape of the template response distributions

In Chapter 7, I mentioned that on average the shape of the template response distributions was Gaussian. To quantify the extent to which each distribution was Gaussian, I computed (for each bin in the image database) the excess kurtosis of the distribution of template responses. Excess kurtosis

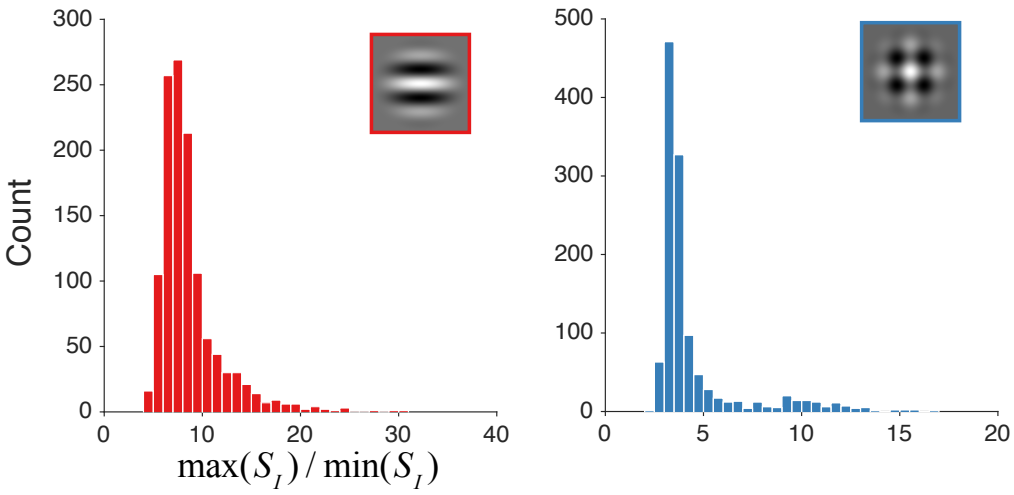

Figure B.2: Distribution of similarity magnitude change in natural images.

is defined as follows

$$Kurt[X] = \frac{E[(X-\mu)^4]}{(E[(X-\mu)^4])^2} - Kurt[N]$$

where $Kurt[N]$ is the kurtosis of a normal distribution. By definition, the excess kurtosis is 0 for a normal distribution.

Figure B.3 shows the distribution of excess kurtosis for each of the 2,000 bins in the image data base (1,000 for each target). The mean excess kurtosis is -0.01 for the Gabor, and 0.07 for the plaid. The median excess kurtosis was -0.17 for the Gabor, and -0.13 for the plaid. From this, it is clear that on average, the template response distributions had no more kurtosis than a Gaussian distribution.

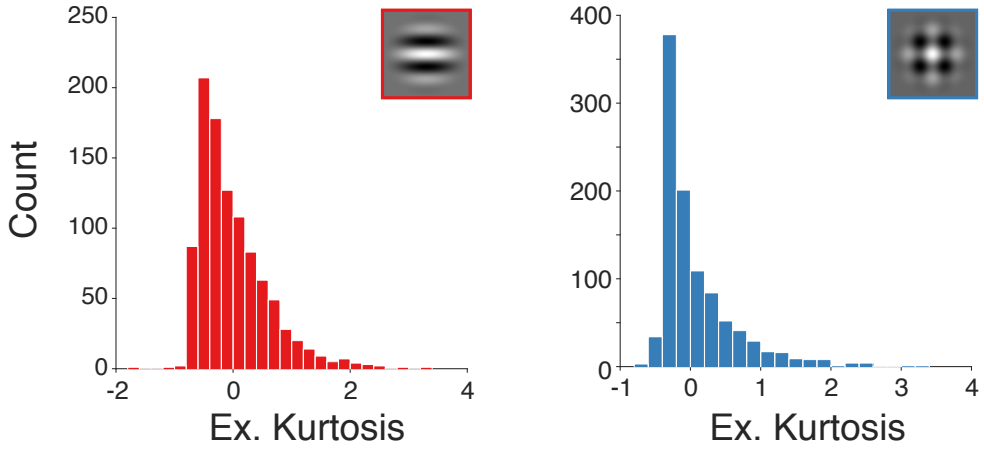

Figure B.3: Average excess kurtosis for the template response distibutions.

Bibliography

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

Alam, M. M., Vilankar, K. P., Field, D. J. & Chandler, D. M. (2014), 'Local masking in natural images: A database and analysis', *Journal of Vision* **14**(8), 22–22.

Albrecht, D. G. & Geisler, W. S. (1991), 'Motion selectivity and the contrast-response function of simple cells in the visual cortex', *Visual neuroscience* **7**(06), 531–546.

Alpern, M., Mason, G. L. & Jardinico, R. E. (1961), 'Vergence and Accommodation', *American Journal of Ophthalmology* **52**(5), 762–767.

Aubert, H. (1865), 'Physiologie Der Netzhaut', *Breslau* .

Barlow, H. B. (1957), 'Increment thresholds at low intensities considered as signal/noise discriminations', *The Journal of Physiology* **136**(3), 469–488.

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E. & Pouget, A. (2012), 'Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability', *Neuron* **74**(1), 30–39.

Bex, P. J. (2010), '(In) Sensitivity to spatial distortion in natural scenes', *Journal of Vision* **10**(2), 1–15.

Blakemore, C. B. & Rushton, W. A. (1965), 'Dark adaptation and increment threshold in a rod monochromat.', *The Journal of Physiology* **181**(3), 612–628.

Blakemore, C. & Nachmias, J. (1971), 'The orientation specificity of two visual after-effects', *The Journal of Physiology* **213**(1), 157.

Blakemore, C. T. & Campbell, F. W. (1969), 'On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images', *The Journal of Physiology* **203**(1), 237–260.1.

Blanchard, J. (1918), 'The Brightness Sensibility of the Retina', *Phys. Rev.* **11**, 81–99.

Bouguer, P. (1760), 'Optical treatise on the gradation of light'.

Bradley, C., Abrams, J. & Geisler, W. S. (2014), 'Retina-V1 model of detectability across the visual field', *Journal of Vision* **14**(12), 22–22.

Brainard, D. H. (1997), 'The psychophysics toolbox', *Spatial vision* **10**, 433–436.

Brunswik, E. & Kamiya, J. (1953), 'Ecological cue-validity of proximity and of other Gestalt factors.', *The American journal of psychology* **66**(1), 20–32.

Burge, J. & Geisler, W. S. (2011), 'Optimal defocus estimation in individual natural images', *PNAS* **108**(40), 16849–16854.

Burge, J. & Geisler, W. S. (2014), 'Optimal disparity estimation in natural stereo images', *Journal of Vision* **14**(2), 1.

Burgess, A. E., Wagner, R. F., Jennings, R. J. & Barlow, H. B. (1981), 'Efficiency of human visual signal discrimination', *Science* **214**(4516), 93–94.

Caelli, T. & Moraglia, G. (1986), 'On the detection of signals embedded in natural scenes', *Perception & psychophysics* **39**(2), 87–95.

Campbell, F. W. & Kulikowski, J. J. (1966), 'Orientational selectivity of the human visual system', *The Journal of Physiology* **187**(2), 437–445.

Campbell, F. W. & Westheimer, G. (1960), 'Dynamics of accommodation responses of the human eye', *The Journal of Physiology* **151**(2), 285–295.

Carandini, M. & Heeger, D. J. (2012), 'Normalization as a canonical neural computation', *Nature Reviews Neuroscience* **13**(1), 51–62.

Chandler, B. & Mingolla, E. (2016), 'Mitigation of Effects of Occlusion on Object Recognition with Deep Neural Networks through Low-Level Image Completion', *Computational Intelligence and Neuroscience* **2016**, 15.

Chandler, D. M., Gaubatz, M. D. & Hemami, S. S. (2009), 'A Patch-Based Structural Masking Model with an Application to Compression', *EURASIP Journal on Image and Video Processing* **2009**(1), 649316.

Cornsweet, T. N. & Teller, D. Y. (1965), 'Relation of Increment Thresholds to Brightness and Luminance∗', *Journal of the Optical Society of America* **55**(10), 1303–1308.

Curcio, C. A., Sloan, K. R., Kalina, R. E. & Hendrickson, A. E. (1990), 'Human photoreceptor topography.', *The Journal of comparative neurology* **292**(4), 497–523.

Dacey, D. M. (1993), 'The mosaic of midget ganglion cells in the human retina', *The journal of neuroscience* **13**(12), 5334–5355.

Daugman, J. G. (1989), 'Entropy reduction and decorrelation in visual coding by oriented neural receptive fields', *IEEE Transactions on Biomedical Engineering* **36**(1), 107–114.

De Valois, R. L., Webster, M. A., De Valois, K. K. & Lingelbach, B. (1986), 'Temporal properties of brightness and color induction.', *Vision Research* **26**(6), 887–897.

de Vries, H. L. (1943), 'The quantum character of light and its bearing upon threshold of vision, the differential sensitivity and visual acuity of the eye', *Physica* **10**(7), 553–564.

Dixon, E. R. (1978), 'Spectral distribution of Australian daylight', *Journal of the Optical Society of America* **68**(4), 437–450.

Drasdo, N., Millican, C. L., Katholi, C. R. & Curcio, C. A. (2007), 'The length of Henle fibers in the human retina and a model of ganglion receptive field density in the visual field.', *Vision Research* **47**(22), 2901–2911.

Eckstein, M. P. (2011), 'Visual search: A retrospective', *Journal of Vision* **11**(5), 14.

Eckstein, M. P., Ahumada, A. J. & Watson, A. B. (1997), 'Visual signal detection in structured backgrounds. II. Effects of contrast gain control, background variations, and white noise', *JOSA A* pp. 1–14.

Eskicioglu, A. M. & Fisher, P. S. (1995), 'Image quality measures and their performance', *IEEE Transactions on communications* **43**(12), 2959–2965.

Felsen, G. & Dan, Y. (2005), 'A natural approach to studying vision', *Nature Neuroscience* **8**(12), 1643–1646.

Field, D. J. (1987), 'Relations between the statistics of natural images and the response properties of cortical cells', *Journal of the Optical Society of America. A, Optics and image science* **4**(12), 2379–2394.

Field, D. J., Hayes, A. & Hess, R. F. (1993), 'Contour integration by the human visual system: evidence for a local "association field".', *Vision Research* **33**(2), 173–193.

Foley, J. M. & Boynton, G. M. (1994), New model of human luminance pattern vision mechanisms: analysis of the effects of pattern orientation, spatial phase, and temporal frequency, *in* T. B. Lawton, ed., 'Computational Vision Based on Neurobiology', SPIE, pp. 32–42.

Frazor, R. A. & Geisler, W. S. (2006), 'Local luminance and contrast in natural images', *Vision Research* **46**(10), 1585–1598.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P. & Movshon, J. A. (2013), 'A functional and perceptual signature of the second visual area in primates', *Nature Neuroscience* **16**(7), 974–981.

Geisler, W. S. (2003), 'Ideal observer analysis', *The visual neurosciences* **10**(7), 12–12.

Geisler, W. S. (2008), 'Visual Perception and the Statistical Properties of Natural Scenes', *Annu Rev Psychol* **59**(1), 167–192.

Geisler, W. S. (2011), 'Contributions of ideal observer theory to vision research', *Vision Research* **51**(7), 771–781.

Geisler, W. S., Najemnik, J. & Ing, A. D. (2009), 'Optimal stimulus encoders for natural tasks', *Journal of Vision* **9**(13), 17.1–1716.

Geisler, W. S. & Perry, J. S. (2011), 'Statistics for optimal point prediction in natural images', *Journal of Vision* **11**(12), 14–14.

Green, D. M. & Swets, J. A. (1966), 'Signal detection theory and psychophysics.', *John Wiley* .

Hartline, H. K. (1938), 'The Response of Single Optic Nerve Fibers of the Vertebrate Eye to Illumination of the Retina', *American Journal of Physiology* **121**(2), 400–415.

Hayhoe, M. M., Shrivastava, A., Mruczek, R. & Pelz, J. B. (2003), 'Visual memory and motor planning in a natural task', *Journal of Vision* **3**(1), 6.

Hecht, S. (1924), 'The visual discrimination of intensity and the Weber-Fechner law.', *Journal of General Physiology* **7**(2), 235–267.

Hecht, S., Haig, C. & Chase, A. M. (1937), 'The influence of light adaptation on subsequent dark adaptation of the eye', *Journal of General Physiology* **20**(6), 831–850.

Hecht, S. & Hsia, Y. (1945), 'Dark Adaptation Following Light Adaptation to Red and White Lights', *Journal of the Optical Society of America* **35**(4), 261–267.

Heeger, D. J. (1991), 'Nonlinear model of neural responses in cat visual cortex', *Computational models of visual processing* pp. 119–133.

Hirsch, J. & Curcio, C. A. (1989), 'The Spatial-Resolution Capacity of Human Foveal Retina', *Vision Research* **29**(9), 1095–&.

Hood, D. C. (1998), 'Lower-level visual processing and models of light adaptation', *Annual Review of Psychology* **49**(1), 503–535.

Howard, I. P. & Rogers, B. J. (2008), *Seeing in Depth: Volume 1: Basic Mechanics/ Volume 2: Depth Perception 2-Volume Set*, Oxford University Press.

Hubel, D. H. & Wiesel, T. N. (1962), 'Receptive fields, binocular interaction and functional architecture in the cat's visual cortex', *The Journal of Physiology* .

Hubel, D. H. & Wiesel, T. N. (1968), 'Receptive Fields and Functional Architecture of Monkey Striate Cortex', *The Journal of Physiology* **195**(1), 215–&.

Ing, A. D. (2010), 'Region grouping in natural foliage scenes: Image statistics and human performance', *Journal of Vision* **10**(4), 1–19.

Inoue, T. (2015), 'Planarian shows decision-making behavior in response to multiple stimuli by integrative brain function', pp. 1–15.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., Broussard, C. & others (2007), 'What's new in Psychtoolbox-3'.

König, A. & Brodhum, E. (1889), 'Experimentelle Untersuchungen über die psycho-physische Fundamentalformel in Beug auf den Gesichtssinn. ', *Zweite Mittlg. S. B. Preuss. Akad. Wiss.* p. 641.

Land, M. F. & Nilsson, D.-E. (2012), *Animal Eyes*, Oxford Animal Biology Series, Oxford University Press, Oxford.

Legge, G. E. & Foley, J. M. (1980), 'Contrast masking in human vision', *Journal of the Optical Society of America* **70**(12), 1458–1471.

Legge, G. E., Kersten, D. & Burgess, A. E. (1987), 'Contrast discrimination in noise.', *Journal of the Optical Society of America. A, Optics and image science* **4**(2), 391–404.

Lettvin, J. Y., Maturana, H. R. & McCulloch, W. S. (1959), What the frog's eye tells the frog's brain, *in* 'Proceedings of the ... '.

Levi, D. M. (2008), 'Crowding - An essential bottleneck for object recognition: A mini-review', *Vision Research* **48**(5), 635–654.

Lu, Z. L. & Dosher, B. A. (1999), 'Characterizing human perceptual inefficiencies with equivalent internal noise.', *Journal of the Optical Society of America a-Optics Image Science and Vision* **16**(3), 764–778.

Mante, V., Frazor, R. A., Bonin, V., Geisler, W. S. & Carandini, M. (2005), 'Independence of luminance and contrast in natural scenes and in the early visual system', *Nature Neuroscience* **8**(12), 1690–1697.

McCann, B. (2015), Naturalistic Depth Perception , PhD thesis, The University of Texas at Austin.

Michel, M. & Geisler, W. S. (2011), 'Intrinsic position uncertainty explains detection and localization performance in peripheral vision', *Journal of Vision* **11**(1), 18–18.

Mueller, C. G. (1951), 'Frequency of seeing functions for intensity discrimination at varous levels of adapting intensity', *The Journal of General Physiology* **34**(4), 463–474.

Najemnik, J. & Geisler, W. S. (2008), 'Eye movement statistics in humans are consistent with an optimal search strategy', *Journal of Vision* **8**(3), 4.

Olshausen, B. A. & Field, D. J. (1997), 'Sparse coding with an overcomplete basis set: A strategy employed by V1?', *Vision Research* **37**(23), 3311–3325.

Papert, S. (1966), 'The Summer Vision Project', Technical Report Memo AIM-100, Artificial Intelligence Lab, Massachussets Institiure of Technology.

Pappas, T. N., Safranek, R. J. & Chen, J. (2000), 'Perceptual criteria for image quality evaluation', *Handbook of image and video processing* pp. 669–684.

Patel, A. S. (1966), 'Spatial Resolution by the Human Visual System. The Effect of Mean Retinal Illuminance. ', *Journal of the Optical Society of America* **56**, 689–694.

Pelli, D. G. (1985), 'Uncertainty explains many aspects of visual contrast detection and discrimination.', *Journal of the Optical Society of America. A, Optics and image science* **2**(9), 1508–1532.

Pelli, D. G. (1997), 'The VideoToolbox software for visual psychophysics: Transforming numbers into movies', *Spatial vision* **10**(4), 437–442.

Peterson, W., Birdsall, T. & Fox, W. (1954), 'The theory of signal detectability', *Transactions of the IRE professional group on information theory* **4**(4), 171–212.

Pokorny, J., Graham, C. H. & Lanson, R. N. (1968), 'Effect of wavelength on foveal grating acuity', *JOSA* .

Polat, U. & Sagi, D. (1993), 'Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments.', *Vision Research* **33**(7), 993–999.

Robson, J. G. (1993), Contrast Sensitivity: One Hundred Years of Clinical Measurement, *in* R. Shapley & D. Mat-Kin Lam, eds, 'Contrast Sensitivity', MIT Press, Cambridge.

Rohaly, A. M., Ahumada, Jr., A. J. & Watson, A. B. (1997), 'Object detection in natural backgrounds predicted by discrimination performance and models', *Vision Research* **37**(23), 3225–3235.

Rose, A. (1948), 'The Sensitivity Performance of the Human Eye on an Absolute Scale∗', *Journal of the Optical Society of America* **38**(2), 196–208.

Rust, N. C. & Movshon, J. A. (2005), 'In praise of artifice', *Nature Neuroscience* **8**(12), 1647–1650.

Selwyn, E. W. H. (1948), 'The photographic and visual resolving power of lenses', *Photographic Journal* pp. 6–12.

Simoncelli, E. P. & Olshausen, B. A. (2001), 'Natural image statistics and neural representation.', *Annual review of neuroscience* **24**, 1193–1216.

Sprague, W. W., Cooper, E. A., Reissier, S., Yellapragada, B. & Banks, M. S. (2016), 'The natural statistics of blur', *Journal of Vision* **16**(10), 23.

Stromeyer, C. F. & Julesz, B. (1972), 'Spatial-Frequency Masking in Vision: Critical Bands and Spread of Masking*', *JOSA* **62**(10), 1221–1232.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J. & Fergus, R. (2013), 'Intriguing properties of neural networks', *CoRR* **abs/1312.6199**.

Tanner, W. P. (1961), 'Physiological Implications of Psychophysical Data', *Annals of the New York Academy of Sciences* **89**(5), 752–765.

Tolhurst, D. J., Movshon, J. A. & Dean, A. F. (1983), 'The statistical reliability of signals in single neurons in cat and monkey visual cortex.', *Vision Research* **23**(8), 775–785.

van Meeteren, A. (1974), 'Calculations on the Optical Modulation Transfer Function of the Human Eye for White Light', *Optica Acta: International Journal of Optics* **21**(5), 395–412.

Van Nes, F. L. & Bouman, M. A. (1967), 'Spatial modulation transfer in the human eye', *JOSA* .

Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. (2004), 'Image quality assessment: from error visibility to structural similarity', *IEEE TRANSACTIONS ON IMAGE PROCESSING* **13**(4), 600–612.

Watson, A. B. (2014), 'A formula for human retinal ganglion cell receptive field density as a function of visual field location', *Journal of Vision* **14**(7), 15–15.

Watson, A. B. & Ahumada, Jr., A. J. (2005), 'A standard model for foveal detection of spatial contrast', *Journal of Vision* **5**(9), 717–740.

Wilson, H. R., McFarlane, D. K. & Phillips, G. C. (1983), 'Spatial frequency tuning of orientation selective units estimated by oblique masking', *Vision Research* **23**(9), 873–882.

Winkler, S. & Susstrunk, S. (2004), 'Visibility of noise in natural images', **5292**, 121–129.

Xing, J. & Heeger, D. J. (2000), 'Center-surround interactions in foveal and peripheral vision.', *Vision Research* **40**(22), 3065–3072.

Yang, Z. & Purves, D. (2003), 'A statistical explanation of visual space', *Nature Neuroscience* **6**(6), 632–640.

Vita

Stephen Sebastian entered Purdue University in the Fall of 2005 to study Computer Science. There, he joined the Human and Computer Vision Lab of Dr. Zigmunt Pizlo. After graduating from Purdue, he entered graduate school at UT Austin, and joined the lab of Dr. Wilson Geisler. He currently lives in Austin with his fiancee Sally, and their dog Sophie.

Permanent address: sebastian@utexas.edu

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.