The Dissertation Committee for Chris Kent Bradley Certifies that this is the approved version of the following dissertation:

Retina-V1 Model of Detectability across the Visual Field

Committee:

Wilson S. Geisler, Supervisor

Lawrence Cormack

Eyal Seidemann

Mary M. Hayhoe

Alan C. Bovik

Retina-V1 Model of Detectability across the Visual Field

by

Chris Kent Bradley, B.A.

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

August, 2014

Retina-V1 Model of Detectability across the Visual Field

Chris Kent Bradley, Ph.D

The University of Texas at Austin, 2014

Supervisor: Wilson S. Geisler

A practical model is proposed for predicting the detectability of targets at arbitrary locations in the visual field, in arbitrary gray-scale backgrounds, and under photopic viewing conditions. The major factors incorporated into the model include: (i) the optical point spread function of the eye, (ii) local luminance gain control (Weber's law), (iii) the sampling array of retinal ganglion cells, (iv) orientation and spatial-frequency dependent contrast masking, (iv) broadband contrast masking, (vi) and efficient response pooling. The model is tested against previously reported threshold measurements on uniform backgrounds (the ModelFest data set and data from Foley et al. 2007), and against new measurements reported here for several ModelFest targets presented on uniform, 1/f noise, and natural backgrounds, at retinal eccentricities ranging from 0 to 10 deg. Although the model has few free parameters, it is able to account quite well for all the threshold measurements.

Table of Contents

Chapter 1 Introduction 1

Chapter 2 Modeling the Human Visual System 6

Optics of the human eye 6

Retinal processing 11

Processing in the visual cortex 23

Ideal observer analysis 37

Chapter 3 The Retina-V1 Model...42

Scope of the RV1 model 42

Basic outline of the RV1 model..43

Retinal image 46

Ganglion cell sampling and the magnification principle 47

Light adaptation 50

Ganglion cell responses 51

Effective masking power 52

Pooling 55

Detectability and contrast threshold..55

Implementation 57

Parameter estimation..58

Chapter 4 Testing the Retina-V1 Model..59

Experiment: Measuring detectability in 1/f noise and natural backgrounds.59

Stimuli...60

Procedure 62

Fitting psychometric functions and thresholds 63

Results...64

Model predictions 68

Uniform backgrounds 71

Non-uniform backgrounds...75

Chapter 5 Discussion 77

Components in the model and components not in the model 77

Comparison to previous models..80

Detectability maps 82

Extensions of the RV1 model 84

Applications of the RV1 model 89

Chapter 6 Concluding remarks 96

Appendix...98

Generation of ganglion cell mosaic 98

Ganglion cell receptive fields 99

Cortical filters 99

Target envelope...99

References..101

Chapter 1: Introduction

The science of vision is more than one and a half centuries old. During this time, many discoveries about how the human brain processes information from the light entering our eyes have been made. On the one hand, many key discoveries in vision science are like key discoveries in physics or chemistry. Rigorous experimentation, often in relatively controlled environments, has given us empirical relations that stand the test of time. Models of neural processing and human behavior are tested and either accepted, rejected, or some combination of both (something's right, but something's also wrong). The level of measurement precision and model accuracy may differ (there is a serious difference in trying to model the behavior of an electron vs. that of a human, especially since completely describing the behavior of one electron completely describes the behavior of all electrons), but the ability of vision science to produce discoveries that form a foundation for further discoveries, and that does not have to be completely torn apart later, is commendable.

However, there is also something very different about the trajectories of fields like physics and chemistry compared to that of vision science, or for that matter any other area of brain science or psychology. Thus far, it is hard to claim that any area of brain science or psychology has produced a "grand theory" of some large category of phenomena where the theory has been proven to be highly accurate. There is no theory in brain science or psychology that has a similar stature to Newton's theory of motion or Maxwell's theory of electromagnetism. It is even hard to say we have come close to formulating such a grand theory. This is true despite having many times more resources and researchers today than in the 17th century when Newton published his theory. It is also true despite having a comparable amount of time to produce such a discovery (approximately one and a half centuries if one goes back to the very early research in modern vision science – this is comparable to the time from the first discoveries in modern science in the mid-16th century to the publication of Newton's Principia in 1687). Some may suggest that such grand theories are not possible in fields like brain science and psychology, or even biology. Indeed, other than the theory of evolution, and perhaps because of it, most of biology looks like "stamp collecting", as the Nobel Prize winning physicist and chemist Ernest Rutherford once said (his exact quote is "All science is either physics or stamp collecting"). On the other hand, there are many areas in brain science and psychology where theories like those seen in physics exist, but on a smaller scale. For example, there is currently no general theory of learning. This is true whether we are talking about a mechanistic model of learning that describes the underlying neural mechanisms of learning, or a behavioral theory that predicts the probability of any response to any stimulus, assuming one knows the history of stimulus presentation (and in principle, assuming one knows the history of responses to those stimuli since the responses themselves can be associated with stimuli that seem to occur as a consequence). There is also no general theory in any of the major subdivisions of learning research, such as in associative learning, instrumental learning or in perceptual learning, among other areas. Nevertheless, one sees many successful models in these fields that resemble models in physics (and later, chemistry) from the 17th through 19th centuries. These models consist of relatively simple mathematical relations with components that map onto measurable (or if not measurable, then intuitive) physical, biological or psychological phenomena. The main difference is that they predict smaller subsets of phenomena and have not yet been unified into a larger, more general model like Newton's theory of motion. Importantly, the scope of these models has often increased over time, suggesting the problem is not necessarily that unifying theories cannot in principle be formulated, but simply that no one has yet succeeded.

Just as in the science of learning, the science of vision has a plethora of models that work well on smaller subsets of phenomena. Vision science is also full of reliable experimental results that could someday form the foundation for developing more general models. In the following sections, I will describe many such discoveries that deal with the optics of the human eye, processing in the retina, and processing in area V1 of the visual cortex. The examples are chosen to highlight what seem to be the key discoveries needed to build a more general model in one small but important area of vision science: the study of target detection. The study of target detection is the study of the probability with which an observer correctly determines that something it is looking for (the target) was present or absent in a background scene. The field is concerned with the neural mechanisms underlying this behavior, as well as the ability to predict human target detection performance accurately. Some people may ask why scientists should bother studying target detection in the first place. It may not be obvious what utility there is in studying something that "low level". The reason is that the ability to detect targets is fundamental to essentially every task the human visual system does. While target detection is rarely sufficient to accomplish a more complex task, it is almost always necessary. The current state of vision science is such that we still do not have a relatively complete understanding of the neural mechanisms underlying target detection, nor do we have a practical behavioral model that accurately predicts the probability of detecting a target in an arbitrary background. The model presented here, called the Retina-V1 model, is an attempt to include many of the mechanisms known to determine human detection performance into a relatively general and practical model of target detection.

There are many different models of target detection that have been published (e.g., Wilson & Bergen 1979; Watson & Solomon 1997; Watt & Morgan 1985; Morrone & Burr 1988; Foley 1994; Foley et al. 2007; Watson & Ahumada 2005; Arnow & Geisler 1996; Goris, et al. 2013). As previously stated, they tend to account for a very limited range of phenomena. One major limitation of many current models is that they predict detection performance only for certain artificial stimuli, like sine waves or Gabors (a Gabor is a sine wave with a Gaussian envelope). They also tend to predict detection performance for such stimuli only when they are presented in uniform backgrounds. These models are generally not designed to handle detection in non-uniform backgrounds where features in the background may adversely affect detection performance. Some models do attempt to predict detection performance in certain types of non-uniform backgrounds, such as white noise backgrounds. However, these models are not designed to predict detection performance in natural backgrounds – the backgrounds most relevant for the human visual system. Thus, one way in which the Retina-V1 (RV1) model is more general than most previous models is that it is designed to make predictions of detection performance in natural scenes. As the reader will see later, it does a relatively good job of this for the small number of stimuli tested (it takes a lot of experimental trials to get reliable data in our experiment, so resource constraints have limited us to testing the model in natural scenes for just 3 types of stimuli). Another way in which the RV1 model is more general than most previous models is that it takes into account the effect of foveation – visual acuity is best at the fovea (light from where one fixates arrives at the fovea), but gets progressively worse away from it. This means that detection performance decreases as one moves further and further out into the peripheral visual field. As the reader will see, the RV1 model also does a relatively good job of predicting detection performance across the visual field. Finally, many models of target detection do not attempt to model the underlying biological mechanisms responsible for detection performance. The RV1 model is in many ways based on known properties of the physiology and anatomy that are considered relevant for determining detection performance. It is not a fully mechanistic model, if for no other reason than that we still do not have all the relevant information about the underlying biology. Nevertheless, a serious attempt is made to ground the model in the relevant physiology and anatomy.

Despite the greater generality of the RV1 model compared to competing models, there are still serious limitations in its scope. In the sections that follow, I will describe phenomena that any truly general model of target detection must account for, but are not predicted by the RV1 model. Known neural mechanisms that might underlie accurate detection of these phenomena, but are not included in the RV1 model, are also detailed. Thus, the RV1 model only attempts to get us partway to a true unifying model of target detection. However, the reader will see that its structure lends itself to many natural extensions that might lead us even further towards this goal. Its true utility, if successful, will be in giving vision science some idea of how far the field has come (or not come) in modeling the mechanisms necessary for accurately predicting target detection performance across a wide range of conditions.

Chapter 2: Modeling the Human Visual System

OPTICS OF THE HUMAN EYE

Visual processing in humans begins when light from the external environment passes through the cornea, pupil and lens to stimulate light sensitive cells called photoreceptors at the back of the eye (photoreceptors lie in the retina – a thin layer of tissue that lines the inner surface of the eye). Photoreceptors transform the light into electrical and chemical signals. These signals are then processed by other neurons in the retina until the end result (in the retina) is a set of neural impulses sent by retinal ganglion cells to other regions of the brain, where further processing occurs. In this section, I will describe key features of the optics of the eye and processing in the retina that a general model of target detection should include. The retina-V1 model includes only a small subset of these features, but as will be seen, it provides a good foundation for adding many of the missing features (see "Extensions of the RV1 model" section).

The cornea, pupil and lens project an image on the retina that is a blurred and distorted version of the external world. Properties of this retinal image are best quantified by what is called the point-spread function (PSF). The PSF describes mathematically the distribution of light (extent of blurring) emanating from a distant point source when it passes through a given optical medium. When the PSF is applied to all point sources of light from the external world or from an image (this is done mathematically through convolution), the resulting

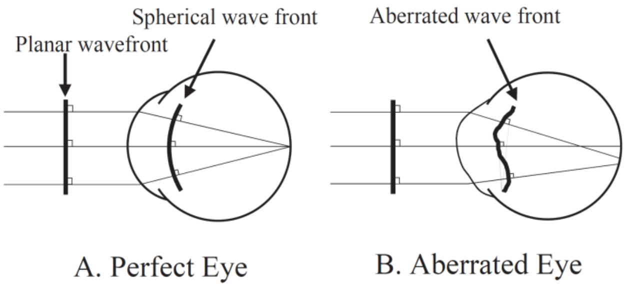


Figure 1. **A.** Aberration free eye: parallel rays converge on a single point **B.** Aberrated eye: parallel rays fail to converge on a single point. (from Williams and Hofer, 2003)

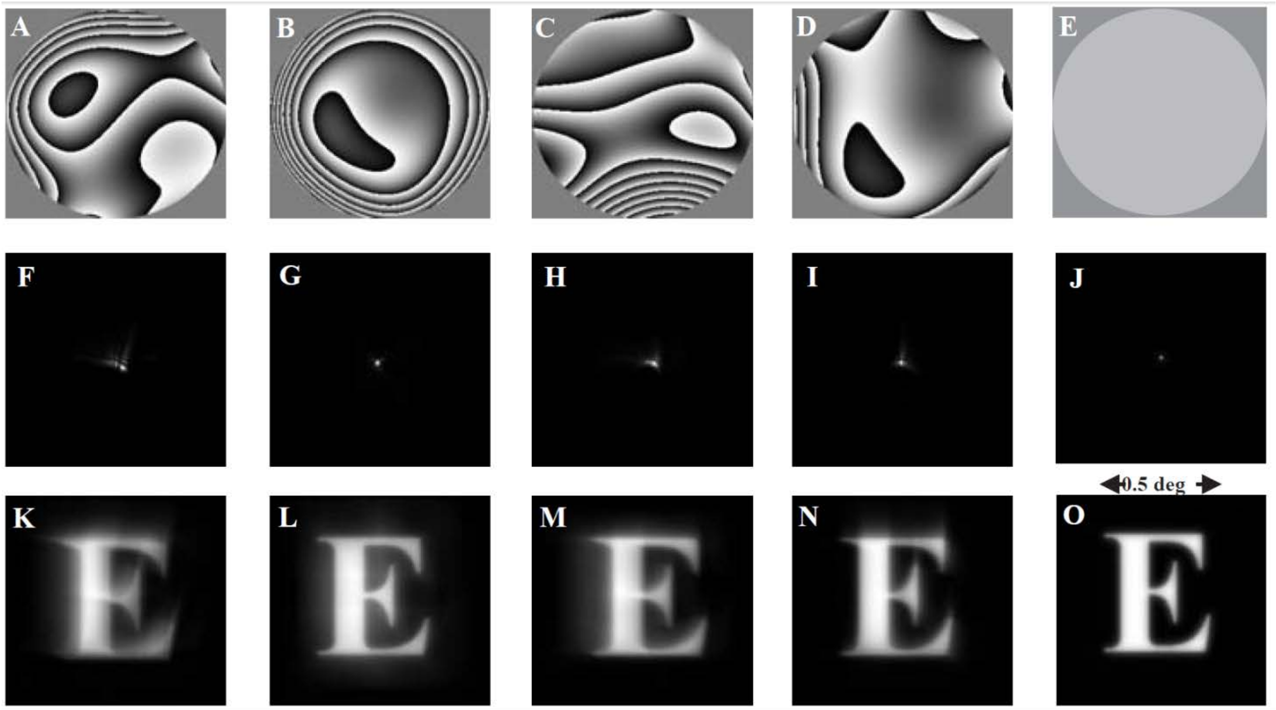


Figure 2. A-D: wave aberrations of four real eyes. E: an ideal, aberration-free eye. F-J: point spread functions (PSF's). K-O: Result of applying the PSF to the letter E. (from Williams and Hofer, 2003).

blurred image can be visualized. For the human eye, the PSF depends primarily on aberrations in the cornea and lens, and on the pupil size. Aberrations in the cornea and lens result from imperfections in their shape. This leads to different rays of light from the same point source being bent (refracted) in such a way that they fail to converge to a single point on the retina. Figure 1 illustrates this. Figure 2 shows the aberrations of four real eyes for a fixed pupil size, compared to a hypothetical aberration-free eye, and their corresponding PSF's applied to the letter E.

The effect of pupil size on the human eye's PSF is more complex to explain because the wave/particle nature of light must first be understood. Light exhibits properties of both waves and particles, but as far as physicists can tell, never both at the same time (this is called the wave-particle duality). Specifically, light is emitted as a particle (called a photon) so one can count particles of light just like one can count any other kind of particulate matter (soccer balls or pebbles on the beach). However, when a stream of photons is emitted from the same light source under (as far as one can tell) the exact same conditions, the distribution of the photons exhibits wave-like properties, and is completely unlike what we would expect if a similar experiment were done with soccer balls or pebbles. If, hypothetically, we perform the same experiment (assume conditions are identical, except the time at which the experiment is performed) on two or more identical soccer balls, the results for all such experiments will be identical. That is, their trajectories, and possible final resting place, will be identical. However, for photons, identical conditions results in a distribution of photons where the distribution resembles what we would expect if we were observing a wave. In Figure 3, one can see that each photon lands in a different location even though it was emitted under identical conditions, and the distribution over time shows a

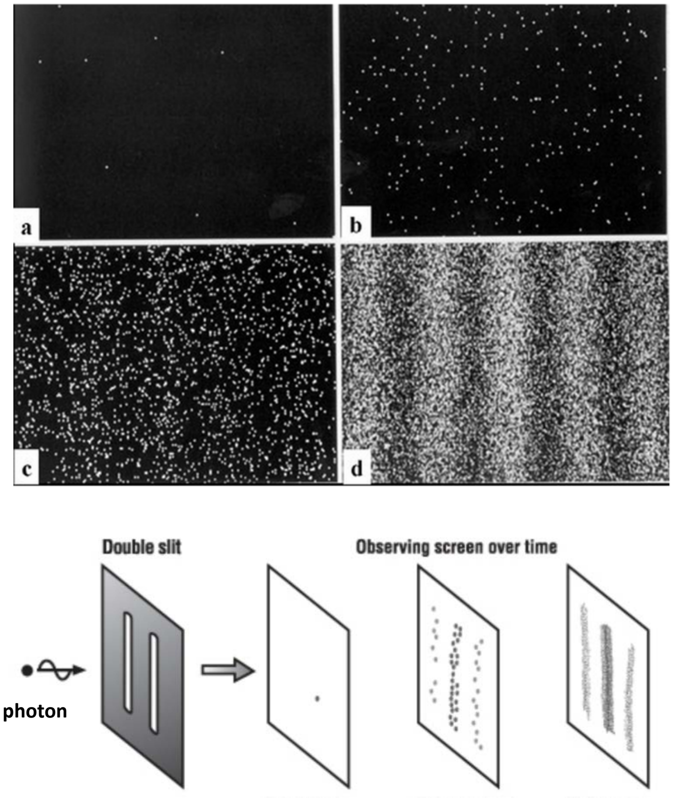


Figure 3. The double slit experiment. A photon is shot through a screen with two slits, neither of which lies directly in the path of the photon. A-D is a time lapse, showing the wave-like diffraction pattern that appears over time. This pattern appears even though the photons are shot from the same location and in the same way each time.

wave-like property called diffraction. Diffraction is observed in other types of waves (e.g. water waves) when they go around obstacles or through slits that are similar in size to their wavelength. Importantly, the wave-like property of photons is what allows each photon to seemingly pass through both slits at the same time in the famous double slit experiment shown in Figure 3 (neither slit lies directly in the path of the photon, so a particle would not pass through either slit). Pupils act like small slits, and a stream of photons will exhibit diffraction when going through it. Figure 4 shows how the magnitude of diffraction increases as the pupil size gets smaller. The combined effect of the cornea, lens and pupil are shown in Figure 4. As one can see, diffraction is the primary determinant of the PSF for small pupil sizes, while aberrations dominate for large pupils.

Two more key properties of the optics of the human eye should be briefly described before I list some reasonable limitations in the scope of the RV1 model. These properties are: chromatic aberration and accommodation. Chromatic aberration refers to the cornea and lens focusing light of different wavelengths at different locations (e.g. different distances from the lens), leading to light of one wavelength being in focus on a given image plane while light of other wavelengths are not. This occurs because any optical medium will refract light of smaller wavelengths (such as blue light) more than longer wavelength light (such as red). In axial aberration, different colors are focused at different distances from the lens. In transverse aberration, different colors are focused at different points on the same (focal) plane.

Accommodation refers to the ability of the lens to change its refractive power as needed. This allows the eye to change the distance at which objects are in focus. Accommodation allows the eye to keep a clear image of an object while it moves towards or away from it. Because the PSF depends on aberrations of the cornea and lens, on pupil

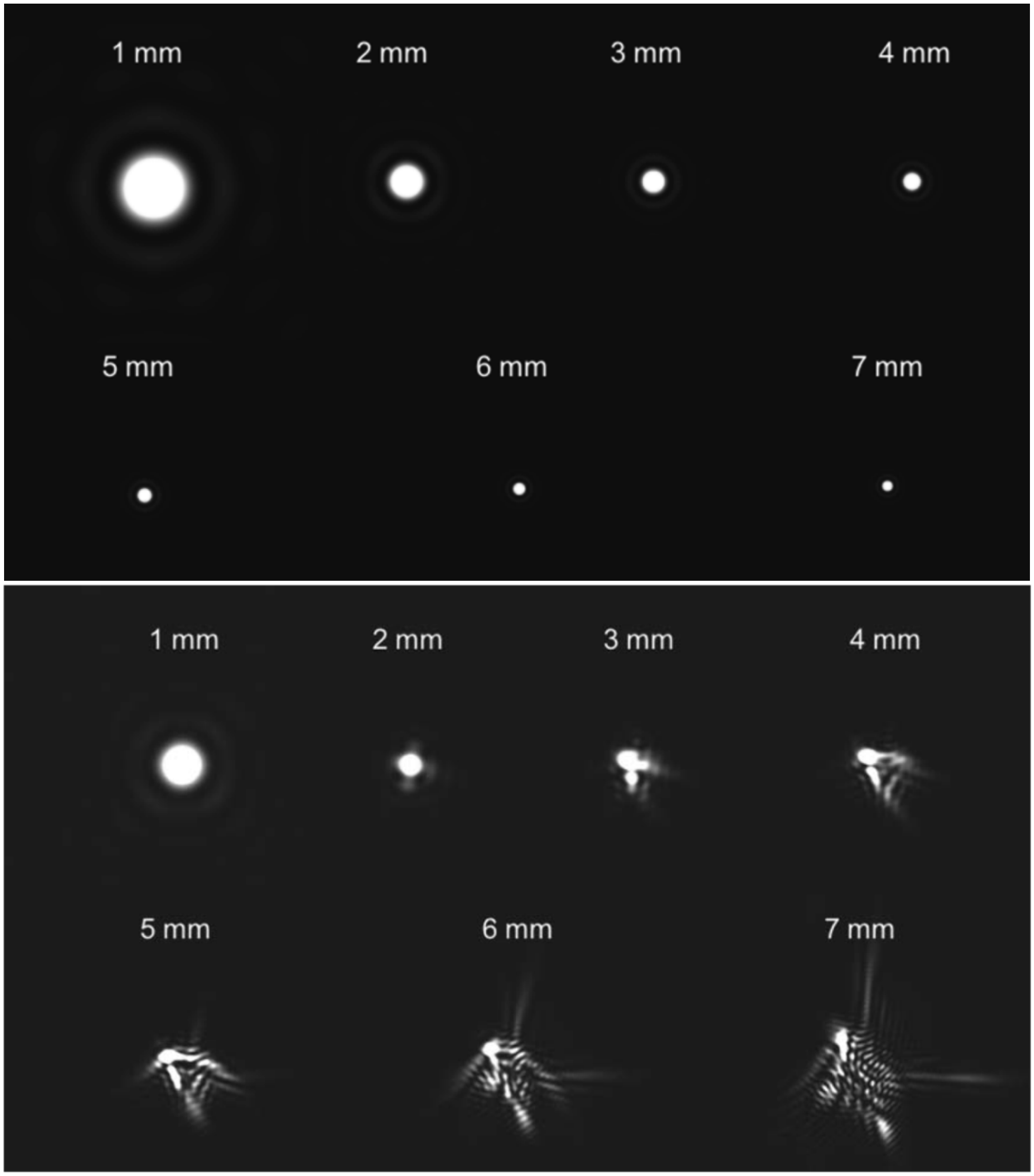


Figure 4. (**top**) PSF's for an ideal, aberration-free eye limited only by diffraction for various pupil diameters. (**bottom**) PSF's for a typical real eye for various pupil diameters. Diffraction dominates PSF shape for small pupils, while aberrations dominate for large pupils. (from Williams and Hover, 2003; courtesy of Austin Roorda).

size, on the wavelength of light, and on accommodation, we will restrict the Retina-V1 model to conditions where the PSF was measured for a fixed pupil size, for monochromatic light, and where accommodation was not a factor. Specifically, we will use the average PSF of four human observers measured by Navarro et al. (1993) in precisely these conditions as our model of the optics of the human eye.

RETINAL PROCESSING

The next step in visual processing occurs when photoreceptors transform the retinal image into a series of electrical and chemical signals through a process called phototransduction. Phototransduction begins when light changes the shape of a light-sensitive molecule called a photopigment. This shape change causes a chemical cascade of events that leads to the closure of first sodium ion (Na^{+}) channels and then calcium ion (Ca^{2+}) channels that lie at the surface (outer membrane) of the photoreceptor. In the dark, these ion channels are kept open, allowing Na^{+} and Ca^{2+} ions into the cell, which leads to the release of neurotransmitters – a chemical signal to other neurons. During phototransduction, closure of these ion channels leads to a decrease in the amount of neurotransmitter that is released. The magnitude of this decrease informs other neurons about the state of the retinal image at the location of the photoreceptor.

There are broadly speaking two types of photoreceptors: rods and cods. Rods are specialized for vision under low light conditions (scotopic vision), while cones are specialized for vision under most daylight conditions (photopic vision). The specialization of rods to low light levels is great enough that a single photon can elicit a response from a rod cell. In contrast, cones often require hundreds or thousands of photons for a response. There are several other important differences between rods and cones. First, cones respond faster to temporal changes in light levels than rods. Second, rods and cones are distributed differently across the retina. As one can see in Figure 5, the

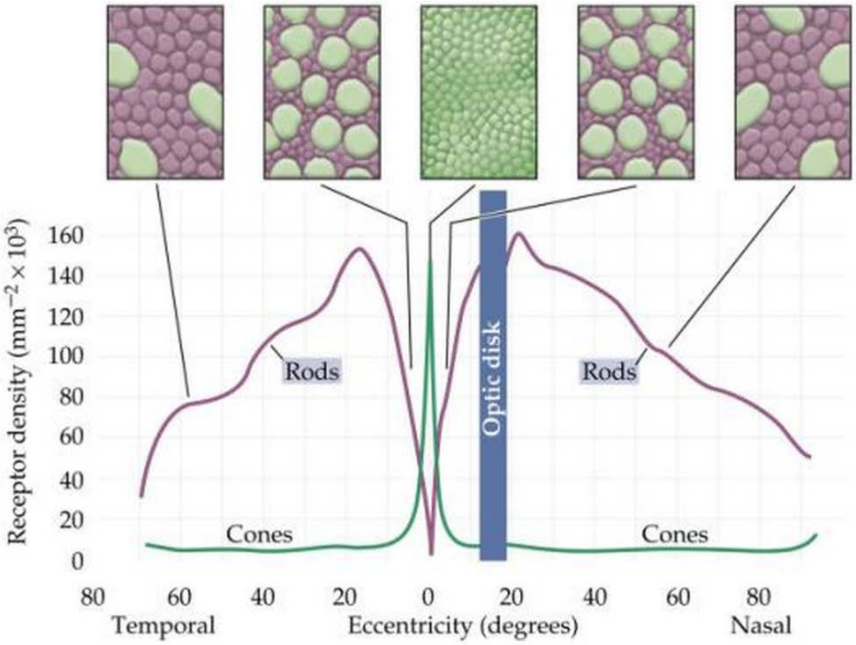


Figure 5. Distribution of rods (purple) and cones (green). Note the lack of rods in the fovea, and also the increasing size of cones in the periphery. (from Purves & Lotto, 2003)

distributions are qualitatively different. In particular, there are no rods in the central portion of the fovea (where light from the direction in which the eye is pointing falls) while the maximum density of cones occurs there. And finally, the subsequent processing of rod and cone responses differs in significant ways. Perhaps the most important difference is the degree of convergence in the outputs of rods and cones. In general, far more rod outputs converge on a single rod bipolar cell – the next cell in the retinal processing pathway – than do cone outputs on cone bipolar cells. In fact, at the fovea, a single cone bipolar cell receives input from only one cone. The greater convergence of rod outputs may be useful for its apparent role as a sensitive light detector because weak responses from many rods can be pooled together to generate a larger response in the rod bipolar cell. However, greater convergence comes with the cost of loss in acuity; indeed,

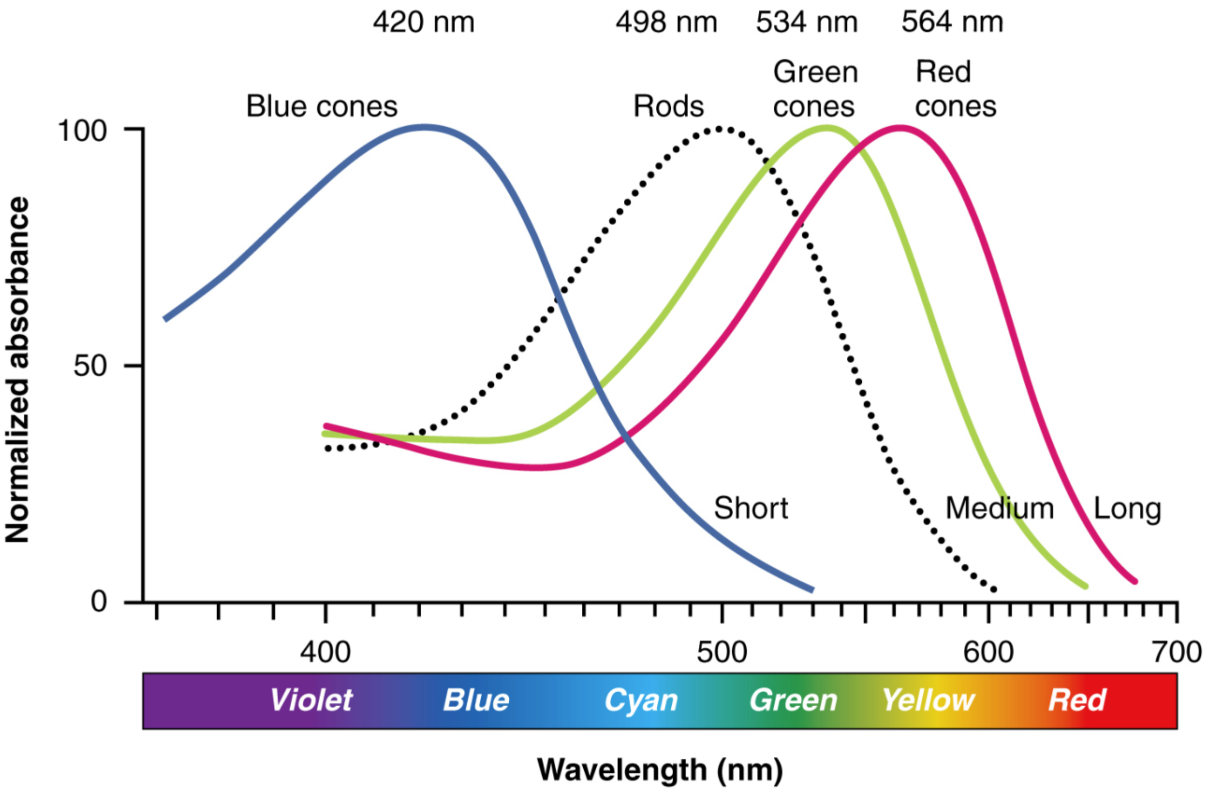


Figure 6. Normalized responsivity spectra of long, medium and short wavelength cones. Rod sensitivities are included for comparison.

the cone system provides greater acuity than the rod system despite the cones being larger than the rods and being outnumbered approximately by 20:1. Because of all these differences between rods and cones, the RV1 model will for now be restricted to detection under photopic conditions only. Nevertheless, the foundation provided by the RV1 model will allow extension to scotopic vision; this is explored in the "Extensions of the RV1 model" section.

There are further restrictions on the scope of the RV1 model that result from the existence of not one, but three types of cones: L-cones, M-cones and S-cones (long, middle and short wavelength, respectively). Each cone has a photopigment that is maximally sensitive to light of a certain wavelength, as seen in Figure 6. The human visual system combines the output of these three types of cones to create our perception of color. Specifically, every perceived color can be modeled as a set of three numbers representing the relative degree to which the L,M and S cones are stimulated. As with the distribution of rods vs. cones, the three types of cones are distributed differently across the retina. The biggest difference lies with the distribution of S-cones compared to those of the L- and M-cones. S-cones tend to be absent from the fovea and seem to be relatively non-randomly, but irregularly distributed, as compared to L-cone and M-cone mosaics (Williams et al., 1981; Ahnelt et al., 1987; Curcio et al., 1991, Roorda & Williams, 1999). In terms of numbers, L-cones are approximately twice as numerous as M-cones, while S-cones may constitute only 10% of the total (Ahnelt et al., 1987; Dacy, 1993). Another complication for including color processing at this stage in the development of a more general model of target detection is the theory of color opponency, which was first proposed by Ewald Hering in 1892. Hering pointed out two special properties of the colors red, yellow, green and blue. First, it seemed like all other colors could be created from them by mixing them together. Second, it seemed that certain pairs of these colors were impossible to perceive: specifically, reddish-green and yellowish-blue hues do not exist. Based on these observations, Hering hypothesized the existence of three distinct opponent mechanisms: a red vs. green mechanism ($\mathrm{L}-\mathrm{M}$), a blue vs. yellow mechanism ($\mathrm{S}-\left(\mathrm{L}+\mathrm{M}\right)$), and an achromatic mechanism ($\mathrm{L}+\mathrm{M}$). In 1957, Hurvich and Jameson provided evidence for the existence of such color-opponency mechanisms. Thus, to keep things simple at this stage, the RV1 model will be designed for and tested on achromatic stimuli (grayscale images). With achromatic stimuli, we need not worry about three different cone mosaics or color opponency mechanisms and can instead use a single cone mosaic, assuming implicitly that the computation on the cone outputs at this stage is $\mathrm{L}+\mathrm{M}$.

Despite now being able to treat all three cone mosaics as one by restricting ourselves to achromatic stimuli, the RV1 model does not explicitly include a model of the distribution of cones in the retina. The reasons for this are twofold: 1) the processing of the cone outputs by bipolar cells, horizontal cells and amacrine cells is still not well understood (specifically, while bipolar cells are thought to simply relay information from the photoreceptors to ganglion cells, how horizontal and amacrine cells modulate that response is not well understood), and 2) good models of ganglion cell output exist for certain classes of ganglion cells. Since ganglion cells are the output neurons of the retina, this means that we can accurately model the output of the retina to certain types of stimuli, even if we don't know the details of retinal processing that led to that output. Furthermore, since ganglion cell processing occurs after cone processing, and ganglion cells essentially form a bottleneck in human visual information processing, modeling the distribution of ganglion cell outputs across the retina makes it unnecessary to explicitly include a cone mosaic. For the RV1 model, we use data on the average ganglion cell distribution of six human retinas from Drasdo et. al (2007) to create a ganglion cell mosaic. The algorithm used, and the resulting mosaic, are shown in a later section.

Before describing our model of ganglion cell output, it is worth pointing out that including the ganglion cell mosaic in the RV1 model vastly increases its scope relative to competing models. Modeling a ganglion cell mosaic means taking into account the effects of decreasing cone and ganglion cell density as a function retinal location. Detection performance is known to deteriorate in the peripheral visual field, but few models attempt to predict such changes in performance as a function of retinal location. Most models of target detection are either models for detecting targets presented at the fovea, or they inaccurately assume equal resolution everywhere in the visual field. Furthermore, it is important to note that detection performance, as well as ganglion cell density, differ not just as a function of distance away from the fovea (retinal eccentricity), but also as a function direction away from the fovea. At any given retinal eccentricity, human visual acuity is highest to the left and right of the fovea, is a bit worse below it, and drops precipitously above it. From an evolutionary point of view, it seems that areas of the visual field where we need the greatest acuity (other than the fovea) tend to lie along the horizon or below the horizon, while the areas of the visual field where we need the least acuity lie above it (which is often the sky). The greatest acuity of course occurs at the fovea. A popular evolutionary explanation for foveation is that the human visual system needs to combine the demands of high acuity and a wide field of view, but with resource constraints prohibiting high acuity across the visual field, the human visual system instead combines a foveated system with ballistic eye movements called saccades to move the eye from one part of the visual field to another. Since the average fixation duration between saccades is 200-250 ms, we will test the RV1 model on images presented during such typical fixation durations (where the observer makes no saccades). This may potentially help in the development of models of overt visual search (overt = using eye movements) where information gathered during a fixation may be used to guide further fixations (see "Extensions of the RV1 model" section).

A very popular model of retinal ganglion cell processing is Rodieck's (1965) Difference-of-Gaussians model. To understand this model, we need to first understand the concept of a receptive field. The first use of the term receptive field was by Sherrington in 1906, who used it to describe the region of skin over which one could elicit a scratch reflex in a dog. The idea of a receptive field for an individual neuron comes from Hartline (1940), who tried to characterize the responses of individual neurons to small spots of light presented at different locations in the visual field. Hartline found that individual neurons in the frog retina responded most vigorously when stimulated with a spot of light in a very specific region of visual space (Hartline, 1938). He defined this region of visual space over which a neuron responds to photopic stimuli as the receptive field of that neuron. Further investigation showed that some retinal ganglion cells showed a burst of activity at light onset, while others showed a burst only when the light is turned off (see Figure 7). Hartline called these ON and OFF cells, respectively. He also found ON-OFF cells that responded both at light onset and offset. The next major advance in understanding ganglion cell processing came when Kuffler (1953) showed that ganglion cell receptive fields in the cat retina have a spatially concentric center-surround structure. For some ganglion cells, a small spot of light presented in the "center" region would increase their firing rate, while that same spot of light presented in the "surround" region decreased it. These ganglion cells were labeled ON-center/OFF-surround. Similarly, other ganglion cells showed the opposite behavior: light in the center region inhibited the ganglion cell, while light in the surround region excited it. These ganglion cells were called OFF-center/ON-surround cells (see Fig 8). The firing rate of the ganglion cell could then be qualitatively predicted by the amount of light that fell on the center region vs. the surround region. In particular, a spot of light in the center region of a receptive field was found to stimulate ganglion cells more than light that covered the entire

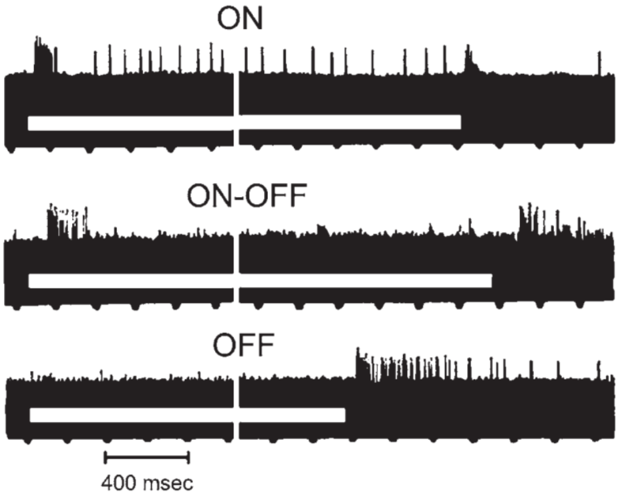


Figure 7. Hartline's recordings from retinal ganglion cell axons. ON cells give a burst of activity at the onset of a stimulus, OFF cell do likewise at the offset, and ON-OFF respond at both onset and offset.

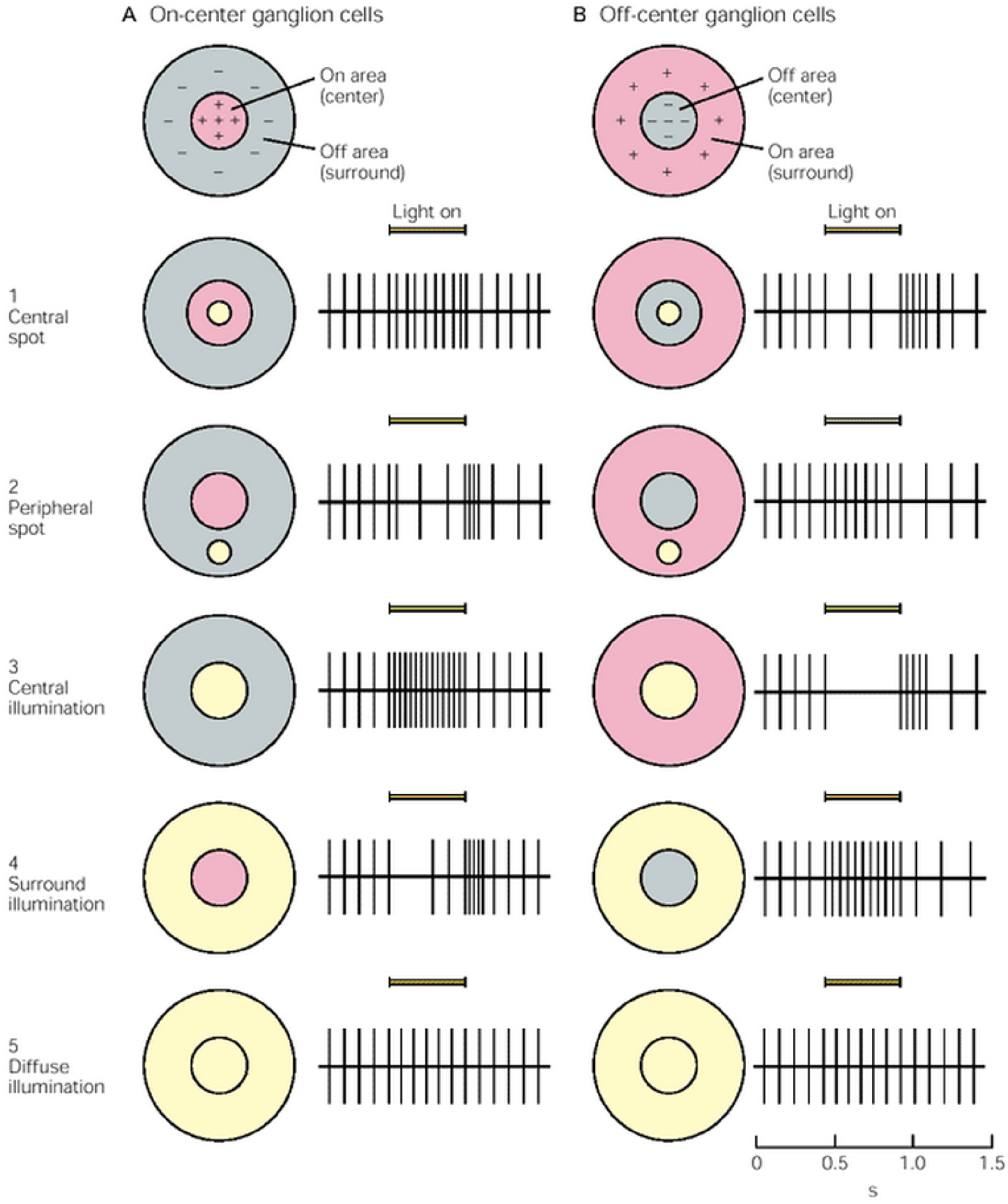


Figure 8. ON-center/OFF-surround and OFF-center/ON-surround receptive fields of ganglion cells. ON-center/OFF-surround ganglion cells respond best to illumination of the center portion of their receptive field. They are inhibited when the surround is illuminated. The exact reverse is true for ganglion cells with OFF-center/ON surround receptive fields.

receptive field (Barlow, FitzHugh & Kuffler, 1957). To allow more quantitative predictions, Rodieck (1965) modeled this receptive field structure with a difference of two Gaussian distributions, one representing the center, and the other representing the surround, with relative weights on each (the volume under the two Gaussians need not be the same). The receptive field center is represented by a Gaussian with a narrow standard deviation, while the surround is represented with a Gaussian with a larger standard deviation. Subtracting the (weighted) surround from the center gives us the Difference-of-Gaussians model (see Fig 9). The RV1 model uses this Difference-of-Gaussians model of ganglion cell processing.

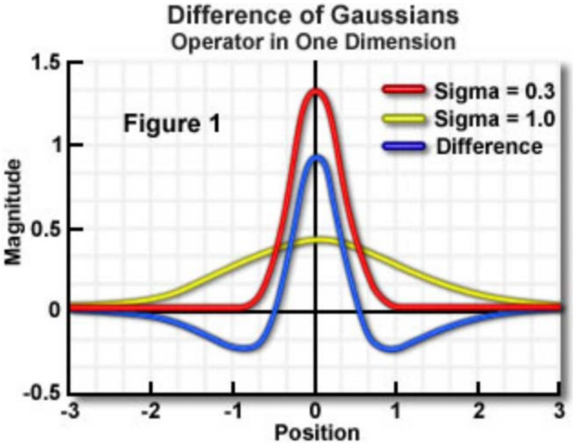


Figure 9. Example Difference of Gaussians (shown in blue). In Rodieck's model, the red Gaussian would correspond to the center portion of a receptive field, while the yellow Gaussian would correspond to the surround portion. The blue Gaussian is the red Gaussian minus the yellow Gaussian, in this toy example.

The relative sizes of the receptive field center and surround have been measured in the macaque retina (Croner & Kaplan 1995) for two types of ganglion cells, P cells and M cells, at different retinal eccentricities. P cells (parvocellular cells) tend to have smaller receptive field sizes than M cells (magnocellular cells). P cells also differ from M cells in the type of information they transmit to the visual cortex. For example, most P cells show evidence of coding red/green color opponency, while most M cells show no evidence of spectral opponency (Wiesel & Hubel, 1966; De Monasterio & Gouras, 1978). The ON center/OFF surrounds

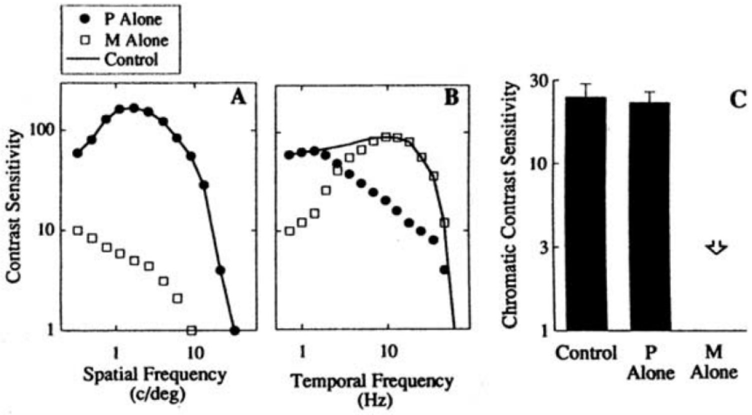


Figure 10. The effect of lesioning just the P cells or just the M cells on spatial (A), temporal (B), or chromatic (C) modulations (from Merigan and Maunsell, 1993).

of P cells are often seen to be either L-ON/M-OFF or M-ON/L-OFF. Since we have already restricted the RV1 model to achromatic stimuli, it would seem that we should model the M cells instead of the P cells. However, Merigan and Maunsell (1991; 1993) showed that lesioning M cells in the macaque preserved detection performance for achromatic gratings of varying spatial frequency (but no temporal frequency), while lesioning P cells caused a huge reduction in detection performance (see Fig. 10). This suggests P cells carry most of the information relevant for detecting stationary gratings across spatial frequencies. The effect of selective lesions of P and M cells was mixed for gratings of different temporal frequency. Lesioning the P cells preserved detection performance for high temporal frequency gratings, while lesioning M preserved performance for low temporal frequency gratings. Studies such as these have led to the view that M cells carry information about luminance, motion and coarse spatial patterns, while P cells are tuned to color and fine spatial patterns. Since we are interested in presenting relatively low temporal frequency stimuli (average fixation duration of 200-250 ms) that can vary greatly in their spatial pattern, the RV1 model will include a model of P cell output and distribution, but ignore the M cells. There are of course many more types of retinal ganglion cells than just P cells and M cells. For example, K cells are thought to process blue/yellow color opponency. In total, at least 17 physiologically different types of ganglion cells have been identified (Dacey, 2004; Field & Chichilnisky, 2007).

Restricting the RV1 model to include only P cell output still leaves us with two types of outputs: ON-center P cells and OFF-center P cells. The distributions of ON-center vs. OFF-center P cell populations are asymmetric: based on the observed differences in diameters of ON-center and OFF-center P cells, there are estimated to be approximately 1.7 times more OFF-center P cells than ON-center P cells (Dacey, 1993). While the reasons for this are not exactly known, it is thought that this is an adaptation to the asymmetric distribution of luminances in natural scenes: an average natural scene will have more "dark" patches (luminances below the mean luminance) than "light" ones (luminances above the mean). For simplicity, the RV1 model combines both ON-center and OFF-center P cells into a single "combined-response" P cell that responds either when an ON-center P cell responds, or when an OFF-center P cell responds. This simplification results in little loss of accuracy because to a good approximation, a point stimulus of light (presented at the center of the receptive field) causes either an ON-center or an OFF-center cell to respond, but not both. If the distributions of ON-center and OFF-center P cells were identical, this combined-response P cell population would provide identical information to the visual cortex as the two separate ones. However, the distributions of ON-center and OFF-center P cells are not the same; thus, there will be some discrepancy between the combined-response P cell output and the actual output of the P cells in the retina. Extending the RV1 model to include separate ON- and OFF-center P cell populations, as well as potentially including other types of ganglion cells (such as M cells or K cells), is discussed in a later section.

One more important property of retinal ganglion cell processing must be included in the RV1 model: luminance gain control. Retinal ganglion cells change the effective range of light intensities they are sensitive to depending on the level of ambient light. Across a wide range of ambient light levels, it is found that the incremental intensity of light ∂I needed to detect a photopic stimulus is related to the ambient background light level I_{A} by: $\frac{\partial I}{I_{\mathrm{A}}} = K$, where K is a constant. That is, the ratio of the light intensity needed to detect a test stimulus relative to the light intensity of the background remains relatively constant. This is called Weber's law, named after Ernst Weber, who in 1834 discovered this relation held in a study on weight lifting. In that study, Weber found that the smallest difference in weights A and B (where $A > B$) that people could detect was always a constant fraction $K = \frac{A}{B}$. Weber's law has been found to hold for a wide range of discrimination tasks, in different modalities, and across many species, though the fit is often not quite perfect. This has led to the popularity of a "near-miss" to Weber's law. For human photopic vision, there is a large amount of data on "threshold vs. intensity" functions that show Weber's law holds for most photopic conditions (Hood & Finkelstein, 1986; Hood, 1997). Functionally speaking, luminance gain control allows a neuron with a limited dynamic range (the ratio between the highest and lowest stimulus intensities over which the neuron gives discernable responses) to adjust the actual range of stimulus intensities over which it best responds. This gives the entire visual system a much higher dynamic range. Thus, if the ambient light level is low, the visual system adapts to make it easier to discriminate light intensities around that ambient light level; if the ambient light level is high, the visual system adapts to best discriminate among larger light levels. Mathematically, the RV1 model implements Weber’s law simply by dividing the Difference of Gaussians response to a stimulus by the mean luminance of the background in the neighborhood of the target.

PROCESSING IN THE VISUAL CORTEX

For predicting detection performance in uniform backgrounds, the RV1 model needs only the output of the P cells. The P cell outputs are combined using a pooling rule inspired by ideal observer analysis, which is described in the "Ideal observer analysis" section. For targets placed in non-uniform backgrounds, the RV1 model requires not just the P cell outputs, but also a model of V1 (visual area 1) neurons in the human visual cortex to predict the degree to which the non-uniform background masks (impairs detection performance of) the target. V1 neurons do not directly receive inputs from retinal ganglion cells. Retinal ganglion cells such as P and M cells project their output to the lateral geniculate nucleus (LGN), and not directly to V1. It is the LGN neurons that project their outputs to cortical areas like V1. Because the receptive fields of LGN neurons have been shown to be highly similar to those of the retinal ganglion cells they receive input from, the next important step in visual processing after the retina is thought to lie in V1.

In 1959, David Hubel and Torsten Wiesel provided neurophysiological evidence of the type of processing that occurs in V1. They showed that the cat visual cortex contained cells that preferentially respond to small bars of light presented at specific spatial orientations (Hubel & Wiesel, 1959). In many cases, these neurons were also selective for the direction of motion: large responses were observed to bars of specific orientations moving in one direction across the receptive field of the neuron but not to similar bars moving in the opposite direction. V1 receptive fields were otherwise similar to those of retinal ganglion cells in that ON/OFF and center-surround organization was

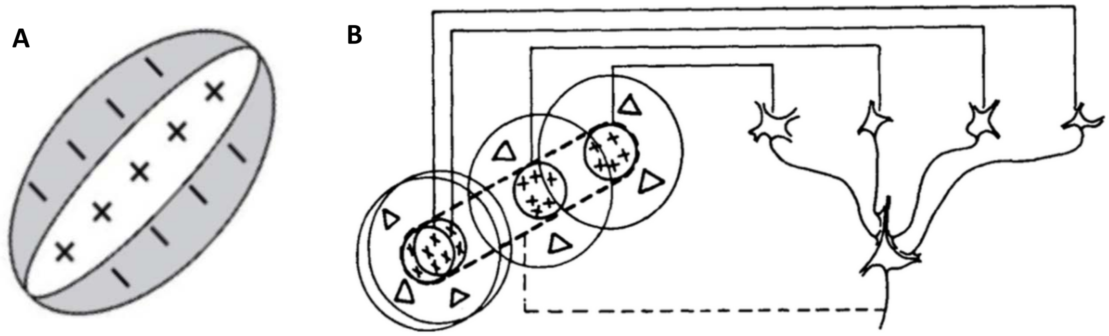


Figure 11. A. A typical V1 receptive field. **B.** Simple linear model by Hubel and Wiesel to explain how elongated V1 receptive fields can be created from the circular receptive fields of LGN neurons

observed. They were dissimilar in that the receptive fields were more elongated, with a shape better suited to detecting bars or small edge elements rather than spots of light (see Figure 11a). Summation of inputs was observed in both the center and surround regions; that is, shining two or more spots of light in the center alone (or in the surround alone) led to a predictable increase in the response of the neuron. The neuron acted as if it was simply summing the inputs to the center and surround. Furthermore, since illuminating the center canceled out illumination of the surround in an additive way, Hubel and Wiesel proposed a simple linear model of how inputs from lateral geniculate nucleus (LGN) neurons can be combined to produce the observed properties of cat V1 neurons. Their model postulated that the receptive field center of a V1 neuron was aligned with the centers of the LGN neurons it received input from (see Figure 11b); the orientation selectivity of V1 neurons arose from the locations in visual space of the LGN neurons' receptive fields. The model was not only conceptually elegant but mathematically simple in that the proposed integration mechanism of LGN outputs by V1 neurons was a simple linear weighting function. Evidence for this model was provided by experiments measuring the correlation between LGN and V1 neurons (Reid and Alonso, 1995).

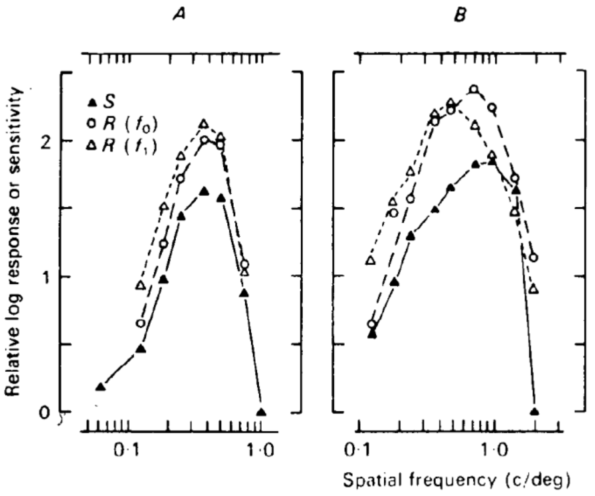


Figure 12. Spatial frequency response and sensitivity curves for (**A**) simple and (**B**) complex cells in the cat visual cortex. Responses of cells to high contrast (0.5) gratings moved across the receptive field. Open circles show mean response, open triangles show amplitude of fundamental Fourier component, and filled triangles show sensitivity. From Movshon et. al. (1978)

Several years later, Hubel and Wiesel (1962) showed that many V1 neurons had more complex receptive field properties than the ones they previously described. They called the previously described cells "simple cells" and the new ones “complex cells”. The primary difference between simple and complex cells was that simple cells responded to bars of light in a very specific region of visual space, while complex cells responded to bars of light in one of many different locations in a larger region of visual space. Thus, complex cell receptive fields showed a degree of spatial invariance not seen in simple cell receptive fields. The relatively clear distinction between center and surround areas in the receptive fields of ganglion cells and simple cells were not found in complex cells.

The next important step in understanding the receptive field properties of V1 neurons occurred when Campbell & Robson (1968) presented psychophysical evidence in humans, and Campbell & Cooper (1969) presented neurophysiological evidence in cats that many V1 neurons are highly tuned to spatial frequency (instead of just a bar of light). Figure 12 shows the spatial frequency sensitivity functions – the relative sensitivity, or strength of response, of a neuron to sinusoidal gratings of different spatial frequencies –

reported by Movshon et. al. (1978) for both simple and complex cells in the cat visual cortex. As one can see, V1 neurons show their most vigorous response to stimuli of a certain spatial frequency, with progressively weaker responses to stimuli of less similar spatial frequencies. Similar results were found in the monkey visual cortex (De Valois et. al., 1982). In both cats and monkeys, the widths of the tuning curves have been found to be similar. The widths of tuning curves are usually described in terms of the bandwidth of the tuning curve. Bandwidth is defined as the ratio between the two frequencies (higher vs. lower) where sensitivity is half that of the peak sensitivity. It is usually stated in octaves, which is the log base 2 of the bandwidth. For cats and monkeys, the average octave bandwidth is about 1.5 (Movshon et. al., 1978; De Valois et. al., 1982).

Does knowing that V1 neurons are preferentially tuned to different spatial frequencies help predict detection performance for complex spatial patterns? In 1968, Campbell and Robson showed how the human contrast sensitivity function (CSF) can predict the threshold at which one detects a complex stimulus composed of many individual sine waves. Specifically, the CSF measures the stimulus contrast at which detection performance reaches some threshold for sinusoidal gratings of different spatial frequencies. The question is whether knowing the detection performance for gratings of different spatial frequencies allows one to predict the threshold for a stimulus that is a combination of those spatial frequencies. One hypothesis was that the threshold for the more complex stimulus would be a linear combination of the thresholds for the individual sine waves. Another hypothesis, drawing on inspiration from earlier work done by others in the auditory system, suggested that the threshold for the complex stimulus would be primarily determined by the threshold of just one of the component sine waves, namely the one that reached threshold first. To test these hypotheses, Campbell and Robson made use of Fourier analysis. Fourier analysis is a mathematical technique for decomposing

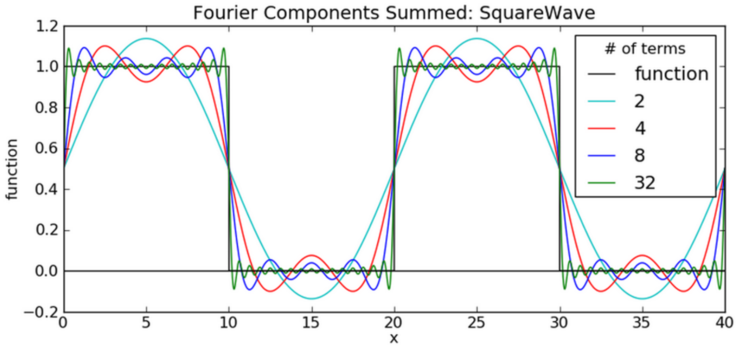


Figure 13. Fourier decomposition of a square wave. The different colors represent the result of adding up more and more components of the Fourier series. In the limit, the sum of all components (an infinite number) will precisely equal the square wave.

any signal (or image) into a sum of sines and cosines. Each sine and cosine has an amplitude and a phase associated with it. Adding all these (possibly infinite number of) sines and cosines together gives back the original signal with no error. In Figure 13, the Fourier decomposition for a square wave is shown. A square wave has a very specific Fourier decomposition that consists of a “fundamental” frequency sine wave, plus odd multiples of that fundamental frequency sine wave (integer multiples of the fundamental frequency are called harmonics). Importantly, the amplitude of the first harmonic (fundamental frequency) of a square wave is exactly $\frac{4}{\pi}$ times the amplitude of the square wave; the amplitude of the $2n+1$ st harmonic, for any positive integer n, will have $\frac{1}{2n+1}$ the amplitude of the 1st harmonic. Campbell and Robson (1968) used this property of square waves to test how the threshold for a stimulus is related to the thresholds of its sine wave components. They found that the detection threshold for almost all square waves tested was almost exactly $\frac{4}{\pi}$ times the detection threshold for a sine wave grating at the fundamental frequency (see Figure 14). The exceptions were for square waves whose fundamental was less than 0.8 cycles per degree. This meant that the threshold for a square wave was essentially being determined by the threshold of a sine wave at the fundamental frequency, as opposed to some linear combination of the thresholds of the component sine waves.

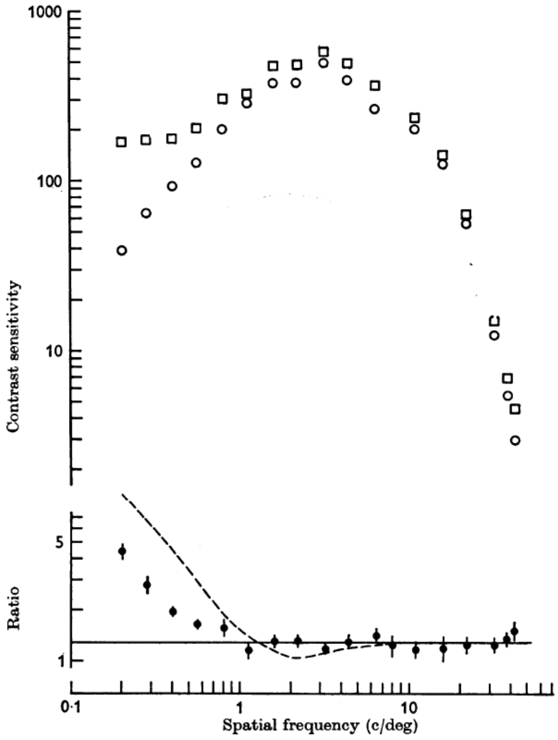


Figure 14. Contrast sensitivity for sine wave gratings (open circles) compared with square wave gratings (open squares). Their ratio is plotted below. The line is at $\frac{4}{\pi}$. From Campbell & Robson (1968).

Detection performance in certain non-uniform backgrounds was also easier to explain using the idea that target detection was mediated by neurons tuned to different spatial frequencies and orientations. In general, it was found that the detection of any spatial pattern becomes more difficult when it is presented on a background that has similar spatial frequency and orientation to the target, at the location of the target (Legge & Foley, 1980; Phillips & Wilson, 1984; Levi, Klein, et al., 2002). For example, a sine wave target presented on another sine wave background (called the mask) is generally harder to detect the more similar in frequency and orientation the target and background sine waves are. A popular explanation for why this type of masking occurs is that the mask stimulates the same feature detectors as does the target. This makes it difficult for any neuron receiving the output of that those feature detectors to determine whether it was the target or the background that caused the response. In other words, the “noise” becomes greater relative to the “signal” in cases where both the noise and the signal contain similar features. Predicting detection thresholds for more complex spatial patterns presented in more complex backgrounds requires a mathematical technique of determining which frequencies and orientations are in the region of the target. This can be done by applying the Fourier transform to both images (target and background). The Fourier transform of any image gives us two images of the same size: the amplitude spectrum and the phase spectrum of the image. The amplitude spectrum tells us which frequencies and orientations are contained in the image, as well as how much of each frequency and orientation is in the image. The phase spectrum determines the locations of the features that are created when the different frequencies and orientations interact in space. The RV1 model makes use of the Fourier transform to determine which frequencies and orientations in the background are also contained in the target; this allows the RV1 model to predict the type of masking just described.

Up until this point, I have described various linear models of how V1 neurons process information. Hubel and Wiesel’s model says simple cell receptive fields are built from a linear combination of the outputs of LGN neurons. Fourier transforms are also linear functions, meaning that (using the mathematical definition of linear function) if $f(x)$ and $g(x)$ have Fourier transforms $F(k)$ and $G(k)$, where k represents frequency, then $\mathcal{F}\left[af(x)+bg(x)\right]=a\mathcal{F}\left[f(x)\right]+b\mathcal{F}\left[bg(x)\right]=aF\left[k\right]+bF\left[k\right]$, where $\mathcal{F}$ is the Fourier transform operator, and a and b are any real numbers. Modeling any physical system as a linear function is attractive because it makes predicting the behavior of that system relatively easy. This is because knowing the response of the system to a small number of specific inputs (responses to a series of simple sine waves in the case of Fourier analysis) completely describes the response of that system to more complex inputs (e.g. weighted sum of sine waves). In the case of cortical neurons, this linear systems approach would mean that the output of that neuron to a sine wave input is also a sine wave. In fact, the frequency of the sine wave will also be the same, with only the amplitude and phase possibly being different. However, there is a great amount of evidence that V1 neurons have important non-linear components to their receptive fields. Indeed, much of the research following the discovery that V1 neurons could be accurately described with simple linear models was aimed at chipping away at this simple narrative.

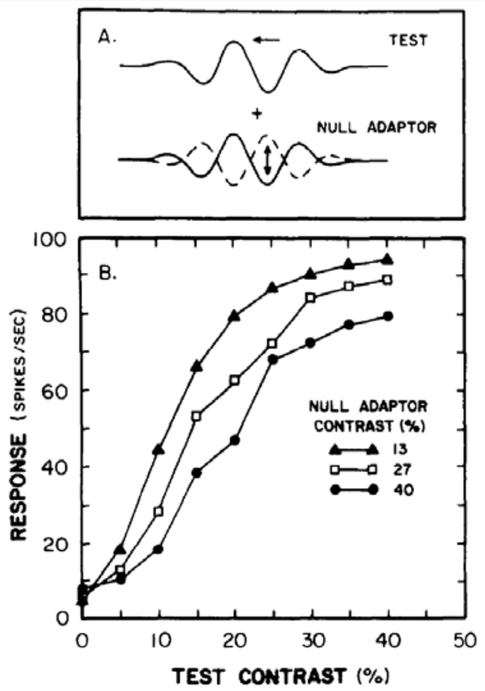


Figure 15. Response of a cortical neuron to a drifting grating after adapting to a null adapter (a stimulus that elicits almost no response from that neuron no matter the contrast). The response to the drifting grating after adaptation depends on the contrast of the null adapter. From Geisler & Albrecht (1992).

One of the first important modifications of the linear receptive field model for V1 neurons came after evidence for non-specific suppression of excitatory responses in V1 neurons was found. Bonds (1989) found that the responses of both simple cells and complex cells in the cat visual cortex decreased when a target sine wave grating was presented on top of another sine wave grating (the mask), even when the mask was presented at vastly different orientations from that of the target. The degree of response reduction varied among the cells. Simple cells with narrow orientation tuning showed more reduction in response than those with broad tuning, and simple cells were generally more susceptible to non-specific orientation suppression than complex cells.

Further evidence for non-specific suppression was provided by Albrecht and Geisler (1991; 1992) when they measured the adaptation effect of a counter-phase grating on neurons in the cat visual cortex. A counter-phase grating is one that reverses in contrast over time. For simple cells, the counter-phase grating can be positioned over the receptive field so that it produces no response (above the spontaneous activity) regardless of the contrast of the grating. Geisler and Albrecht measured the strength of adaptation/masking to this grating by varying the contrast of another, drifting grating that was superimposed on it. The results are shown in Figure 15. They found that the contrast of the counter-phase grating (also called the null adapter) changed the degree to which the neuron adapted to it, despite that counter-phase grating eliciting no response from that neuron. Presumably, this occurs because other neighboring neurons do respond to this counter-phase grating, and are sensitive to its contrast. Specifically, the results suggest that the neighboring neurons affect the gain of the target neuron, dependent on the contrast of the grating. Thus, this study provides evidence for non-specific suppression, and specifically for contrast gain control. In a separate paper, Albrecht and Geisler (1991) showed that the receptive field structure of these same cat visual cortex neurons was consistent with a model that has linear and nonlinear components. In the model, the linear spatiotemporal receptive field component is responsible for the initial selectivity to different stimulus properties such as frequency, orientation, and direction, etc… Nonlinear mechanisms play different roles. One is an exponent applied to the output of the linear component, which enhances selectivity to the stimulus. The other is non-specific suppression (in this case, contrast gain control), which adjusts the sensitivity similar to the way luminance gain control does for retinal ganglion cells. Because the gain control only depends on contrast (not on the receptive field characteristics of the target neuron), it has a broadband masking effect, unlike the narrowband (as in narrow bandwidth) masking discussed previously. It has the effect of causing response saturation (neuronal responses gradually asymptote as a function of stimulus contrast), but at the same time preserving the selectivity of the neuron along all stimulus dimensions except contrast.

Heeger (1991; 1992) popularized a model of V1 cells that took into account broadband masking as well as other observed nonlinearities. Its most recent version is found in Carandini and Heeger (2012):

$$R_j = \gamma \frac{D_j^n}{\sigma^n + \sum_k D_k^n} \tag{1}$$

In equation 1, D_j represents neuron j's driving input (such as a linear receptive field) and R_j represents neuron j's final response; the three remaining parameters: γ, σ and n are constants, with $n = 2$ a common choice. The normalization term $\sum_k D_k^n$ in the denominator runs over k neighboring neurons and is responsible for predicting gradual response saturation and broadband masking. In his early models, Heeger (1991; 1992) specified the nature of the normalization mechanism: he hypothesized that a feedback circuit was responsible for the normalization term; Albrecht and Geisler (1991; 1992) left this question open. We still do not know what the precise neural mechanisms are that produce normalization (Carandini & Heeger, 2012). There is evidence that the mechanism is feedback in nature, but there is also evidence it is feedforward. Evidence it is a feedback circuit comes from measuring normalization signals that originate from neurons whose receptive fields lie in regions of visual space relatively far from that of the target neuron. These signals show selectivity resembling that of V1 neurons. This suggests that the normalization term is summing over the output of other V1 neurons, and is thus possibly the result of a feedback mechanism. Evidence that normalization results from a feedforward circuit comes from measuring normalization signals that originate from neurons whose preferred location in visual space are near that of the target neuron. These neurons show relatively broad stimulus selectivity, similar to those seen in the receptive fields of LGN neurons. This argues possibly for a feedforward circuit as it is evidence that the inputs to the V1 neurons are driving the normalization term.

The RV1 model predicts broadband masking without explicitly encoding equation 1 (or the normalization model in Albrecht and Geisler). Instead, the broadband component of the RV1 model calculates the power in the P cell responses that contributes to masking but is not dependent on spatial frequency or orientation. This gives essentially the same result as explicitly encoding equation 1, but it allows for greater ease of computation since it can be done directly on the P cell outputs. To summarize, the RV1 model has two separate components responsible for masking: a narrowband component and a broadband component. In principle, there should be a single target-dependent cortical filter that does the job of both of these components because the cortical neurons represented in our narrowband filter are the same neurons responsible for broadband masking in the brain. How the RV1 model might be extended to use just one cortical filter that combines both narrowband and broadband masking is explored in the "Extensions of the RV1 model" section.

If the only nonlinearities of V1 neurons were the response exponent and contrast normalization, then the RV1 model would be well on its way to capturing the major masking effects of arbitrary backgrounds on target detectability. However, there is evidence of even more nonlinearities in the response properties of V1 neurons. For example, research pioneered by Polat and Sagi (1993; 1994) showed that a mask that was not at the location of the target, but a certain distance away from it, could actually facilitate the detection of the target. In their experiments with Gabor stimuli (a Gabor is just a sine wave with a Gaussian envelope), they found that flanking a target Gabor with

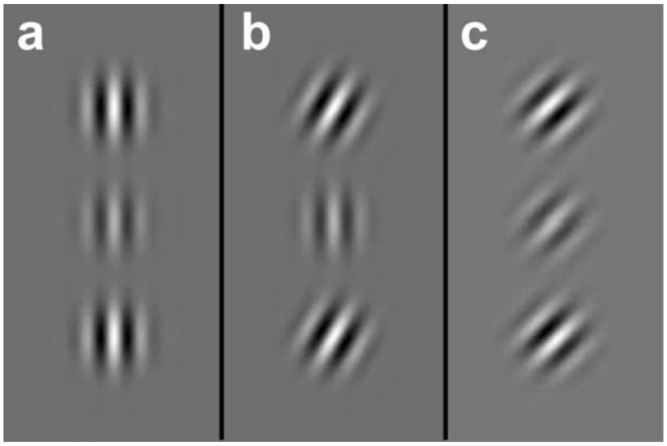


Figure 16. Thresholds were measured for the central Gabor. Significant facilitation was found in (**a**), but not in (**b**) or (**c**) where the Gabors were not collinear with each other. From Polat and Sagi (1993; 1994).

two other Gabors of the same frequency and orientation actually facilitated detection, while flanking the target Gabor with Gabors of different orientations had no such effect (see Figure 16). Control experiments showed that flankers with different frequencies but same orientation still facilitated detection, while the absolute orientation of the Gabors did not matter. Others, such as Field et. al (1993) showed that a chain of Gabors that was placed in a random field of Gabors was detected only if the Gabors were collinear to a smooth curve, or tangent to each other. These studies showed the existence of facilitation mechanisms (called collinear facilitation) that operate over areas larger than the size of simple and complex cell receptive fields. Such mechanisms have been postulated to be involved in feature integration (say for contours) that occurs possibly after processing in V1.

Another example of masking by features outside of the receptive fields of V1 neurons responding to a target is the phenomenon of crowding. Examples of crowding are shown in Figure 17 (b-d). As the reader can see for himself, the ability to identify the letter R becomes harder the more flanking letters of the same size there are (assuming one fixates at the fixation dot). No good models of crowding exist at this time, in the sense that none can predict the major properties of crowding. Evidence that crowding is

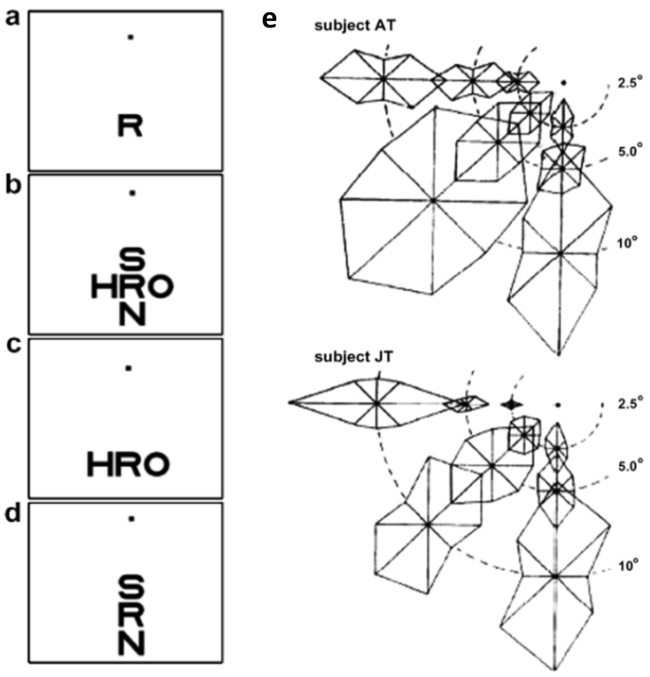


Figure 17. Examples of crowding (b-d) for the letter R. The R becomes harder to identify with more flanking letters. The peripheral extent of crowding (e) tends to scale with retinal eccentricity. The widths of the regions differ in different directions away from the fovea.

different from the type of masking discussed earlier (some call it “ordinary” masking) comes from several sources (Pelli et. al, 2004). First, the strength and extent of crowding in the peripheral visual field is greater than for ordinary masking (Andriessen & Bouma, 1976; Levi et al., 2002). Figure 17e shows the extent of crowding in parts of the peripheral visual field. In general, the width of the crowding region tends to be proportional to retinal eccentricity, while the region for ordinary masking tends to remain constant across the visual field. Second, ordinary masking and crowding seem to produce qualitatively different kinds of effects. In ordinary masking, the target ceases to be detected; in crowding, the target remains visible, but cannot be accurately identified. Thus, while masking affects detection and identification, crowding only affects identification. Contextual effects on V1 neuron responses have been shown to go even further, including differences in response properties based on whether the V1 neuron’s receptive field was within a "figure" region or a "ground" region (Lamme, 1995; Zipser et. al, 1996). In these experiments, a V1 neuron’s receptive field receives the same stimulus in two conditions. In one condition, the receptive field lies within a "figure" region, while in the other condition, it is part of the "ground". The responses are the same up till about 80 msec after stimulus onset. However, the response properties differ afterwards (greater neuronal response for the “figure” condition).

Another general category of nonlinearities in V1 neuron receptive fields is their temporal properties. One such property already mentioned was adaptation. Adaptation occurs when a high-contrast stimulus stimulates a neuron for an extended period of time (e.g 30 seconds). There are also well-known refractory effects of neurons that operate on the order of a few milliseconds, well within the time of a single fixation.

And then there is perceptual learning. Ball & Sekuler (1982) provided an early demonstration that humans improved when trained on a motion discrimination task. They found the improvement lasted over 10 weeks. That humans improve through learning is not surprising. The question was what the underlying mechanisms were. That is, how does perceptual learning change the receptive field properties of neurons? One hypothesis was that perceptual learning leads to the narrowing of bandwidths. This hypothesis was tested by Sarrinen and Levi (1995), with measurements of orientation bandwidths (using a masking paradigm) before and after observers performed a Vernier acuity task (a Vernier acuity task is a task that measures the ability to distinguish whether two stimuli are aligned or not). They found that the bandwidths after learning were narrower in some cases by a factor of 2; the degree of narrowing was also found to correlate with the improvement in the task. Thus, the receptive field of a V1 neuron not only cannot be described by a simple linear function, that function itself changes over time due to perceptual learning.

In summary, there are many known nonlinearities in the response of V1 neurons. The RV1 model only includes the narrowband and broadband masking components described earlier and restricts itself to predicting ordinary masking. No collinear facilitation, crowding, effects of context, adaptation, or perceptual learning phenomena are designed to be predicted by the RV1 model. Despite these limitations, the scope of the RV1 model is still significantly broader than competing models. The two primary areas where the RV1 is more general than other models is that it is designed to predict detection performance across the visual field, and it is designed to predict ordinary masking in arbitrary backgrounds. Experimental evidence shown later suggests it does relatively well in these tasks.

IDEAL OBSERVER ANALYSIS

How do we take the outputs of the V1 neurons (or if we choose, the P cells) and combine them to predict the probability a target will be detected at any location in the visual field on an arbitrary background? This question cannot currently be answered by neurophysiology. We do not yet understand how the outputs of V1 neurons are processed by other areas of the brain, ultimately resulting in a decision on whether the target is present or absent. Instead, we take a different approach: ideal observer analysis. In ideal observer analysis, the goal is to determine the optimal solution to a problem the human visual system must solve. For the problem of target detection, the question reduces to the one we have: given the outputs of any stage of processing already done on the stimulus, how does one use those outputs to maximize accuracy in the task (correctly determine whether a target was present or absent in the background). It is important to note here that using ideal observer analysis in the RV1 model does not imply we are assuming humans integrate information optimally. It is just a principled means of deriving the mathematical form of the pooling equation – the equation used to combine the outputs from any given stage of processing to predict detection performance. We can and will modify the optimal pooling rule to allow for suboptimal pooling; this may better describe what humans do. The important point here is that it is better to derive the form of the pooling rule from a principled method like ideal observer analysis than from just using intuition. Two questions must be answered here: 1) what is the optimal pooling rule and how does one derive it, and 2) what inputs should go into the ideal observer? First, we describe the mathematics behind ideal observer analysis.

The roots of most forms of ideal observer analysis go back to Bayes' theorem in probability theory. To illustrate Bayes' theorem, let's begin with a simple example. Suppose we have two coins, C_1 and C_2. Suppose also that only two possible events can occur if we flip either coin: heads or tails (which we'll represent as H and T, respectively). Now suppose a coin is flipped, and the result is heads, but for whatever reason we don't know which coin was flipped. The question Bayes (1702-1761) was interested in solving was: what is the probability the coin that was flipped was C_1 (or C_2) after observing H? He found that he could solve this problem if several other probabilities were already known. First, he needed to know what the prior probabilities of C_1 and C_2 are. The prior probabilities $p(C_1)$ and $p(C_2)$, of C_1 and C_2, represent the probability of C_1 and C_2 being flipped before H was observed. He also needed to know the conditional probabilities of H given C_1 and C_2. That is, he needed to know the probability $p(H|C_1)$ of H occurring given that the coin flipped was C_1. And similarly, he needed to know the probability $p(H|C_2)$ of H occurring given that the coin flipped was C_2. Once these probabilities were known, Bayes could solve the problem using what is now known as Bayes' theorem. Bayes' theorem states that for any events A and B, the following holds:

$$p(A|B)\,p(B) = p(B|A)\,p(A) \qquad (2)$$

How do we use Bayes' theorem to solve the problem of finding the probability that the flipped coin was C_1 (or C_2) given that H was observed? First, we note that in mathematical notation we want to know which of $p(C_1 | H)$ and $p(C_2 | H)$ is greater. By rearranging terms in Bayes' theorem, we find that $p(C_1 | H) = \frac{p(H | C_1) p(C_1)}{p(H)}$, and $p(C_2 | H) = \frac{p(H | C_2) p(C_2)}{p(H)}$. To compare $p(C_1 | H)$ and $p(C_2 | H)$, we first cancel the denominator $p(H)$ and then note that all the other probabilities (the prior and conditional probabilities) are already known. This solves the problem. If, in addition, we wanted to specify an optimal decision rule for correctly predicting which coin was flipped, we would say: if $p(C_1 | H) > p(C_2 | H)$, then select C_1, and if $p(C_1 | H) < p(C_2 | H)$, then select C_2. In the special case where $p(C_1 | H) = p(C_2 | H)$, then it is irrelevant which of C_1 or C_2 one picks if the goal is to maximize accuracy. This simple example can be generalized to the case where there are k possible events $E_1,...,E_k$ and n possible categories $C_1,...,C_n$.

Applying Bayes' theorem to the problem of target detection is straightforward in principle. In a target detection paradigm, there are two possible categories: target present, or target absent. Stated another way, either the stimulus (the event in this case) is target + background, or background alone. Assuming one can measure the prior and conditional probabilities (this is however often not possible), Bayes' theorem gives us a decision rule that tells us how to predict whether the target was present or absent with maximal accuracy. Thus, the ideal observer for a single "detector", such as a single retinal ganglion cell or a single cortical neuron, is in principle specified by Bayes' theorem. Of course, the question we really want the answer to is how to optimally combine the outputs of many such individual ideal observers (many neuronal outputs). This problem was solved in by Green and Swets (1966). There are several simplifying mathematical assumptions Green and Swets made. They assumed that the distributions of target + background (signal + noise) and background alone (noise alone) were both Gaussians having the same variance, differing only in their means. They also assumed that all detectors are ideal observers and their outputs are statistically independent (this simplifies things because the product of two statistically independent events occurring is the product of their individual probabilities). In the special case where the detectors received information sequentially, it was assumed no loss of information occurred over time. Under these conditions, they showed that if one defines d' (called d prime) as the difference between the means of the distributions of target + background and background alone, divided by the (average) standard deviation, then the percent correct of an ideal detector PC_{ideal} was related to d'_{ideal} by

$$PC_{\text{ideal}} = \Phi\left(\frac{d'_{\text{ideal}}}{2}\right) \tag{3}$$

where Φ is the standard normal integral function, and

$$d'_{\text{ideal}} = \sqrt{\sum_{i=1}^{n}\left(d'_i\right)^2} \tag{4}$$

Thus, the optimal way to combine the outputs (d'_i) of the individual detectors is to take the square root of the sum of the squares of the individual outputs (essentially compute a Euclidean distance, if one treats the d'_i as coordinate values). We note that the assumptions made by Green and Swets (1966) often do not hold in the strict mathematical sense. Nevertheless, the derived optimal pooling rule often approximates the actual optimal solution quite well in many cases.

The RV1 model uses equation 4 as the basis for its pooling rule. The one modification made is to allow for suboptimal pooling by letting the exponent on d'_i be a free parameter (that is, we let the exponent be different than 2). The question that still needs to be answered is at what stage of processing should we use ideal observer analysis? Should the ideal observer operate on the retinal image, the P cell outputs, or the V1 neuron outputs? We choose to let the ideal observer operate on the P cell outputs for several reasons. First, the retinal image should not be the inputs to the ideal observer because the retinal image does not include any effects of retinal eccentricity on detection performance. That is, since the brain does not have access to all the information contained in the retinal image (due to the distribution of retinal ganglion cells), and the ideal observer will use all available information, it is unwise to let the retinal image be the inputs to the ideal observer. This same type of "information loss" argument will not work in rejecting the P cells (vs. V1 neurons) as the input to the ideal observer because we assume in the RV1 model that there are more than sufficient resources in the visual cortex to fully represent the P cell outputs. We select the P cell outputs, instead of the V1 neuron outputs, as the inputs to the ideal observer primarily for two reasons: 1) we don't know enough about the processing in V1 and the distribution of different types of V1 neurons to avoid making many extra assumptions about the visual cortex that would otherwise not be made in the model, and 2) computational complexity would increase in our model because V1 has orders of magnitude times more neurons than there are P cells. Thus, the RV1 model makes use of known properties of cortical neurons to predict masking power, but it draws its inspiration for how to pool information about the target (absent masking) from ideal observer analysis.

Chapter 3: The Retina-V1 Model

SCOPE OF THE MODEL

The preceding sections described phenomena that any truly general model of target detection should include. As the reader has by now learned, the Retina-V1 model attempts to make accurate predictions only over a very limited range of conditions. To summarize from the previous sections, these include: fixed pupil size, grayscale images, conditions where accommodation is not a factor, photopic conditions, low temporal frequency stimuli, modeling only the P cells in the retina, ignoring separate distributions of ON-center and OFF-center P cells, and including only "ordinary" masking (no collinear facilitation, crowding, contextual effects, adaptation or perceptual learning, etc…). Nevertheless, while these restrictions are severe on an absolute scale, the RV1 model is far more general in its scope than competing models. For the task considered here – the signal-known-exactly task where the target and the location of the target (if it appears) is known to the observer – there are a great many models that have been proposed to account for detection and discrimination over narrow ranges of conditions (e.g., Wilson & Bergen 1979; Watson & Solomon 1997; Watt & Morgan 1985; Morrone & Burr 1988; Foley 1994; Foley et al. 2007; Watson & Ahumada 2005; Arnow & Geisler 1996; Goris, et al. 2013). None of these models are as general in scope as the RV1 model. Specifically, the RV1 model is designed to predict detectability over the entire visual field (not just the fovea), and it is also designed to predict detectability in natural backgrounds (not just uniform or artificial noise backgrounds). In developing this model, it is important to note that our goal was not to incorporate all existing knowledge into a grand model, nor to compete with existing models designed for a narrow range of conditions, but rather to combine what appear to be the most important factors identified in the spatial vision literature into a streamlined model, for which it is practical to generate predictions rapidly for arbitrary backgrounds and targets at arbitrary retinal locations.

The proposed model is based largely on known anatomical and physiological factors, and hence there are relatively few free parameters. To estimate some of these parameters, and test the model for foveal detection on uniform backgrounds, we fitted the ModelFest data set, which consists of detection thresholds measured in 16 observers, for 43 different targets (see, Watson & Ahumda 2005). To estimate the remaining parameters and test the model in the more general case, we measured and then fitted detection thresholds in 3 observers, for 1/f noise backgrounds (which have the power spectrum of natural images; Burton & Moorehead 1987; Field 1987) and natural image backgrounds, for three ModelFest targets, at several eccentricities. In what follows, we first describe the model, then the psychophysical measurements, and finally the predictions for the psychophysical measurements.

BASIC OUTLINE OF THE MODEL

There are orders of magnitude more photoreceptors and primary visual cortex neurons than there are retinal ganglion cells. This fact alone suggests that the optic nerve (the population of retinal ganglion cells) may be the main bottleneck for spatial pattern detection information in the human visual system. In other words, there are likely to be sufficient neural resources in early cortical areas to encode the ganglion cell responses from the target and the background. Also, the target and background are most compactly represented in the ganglion cell responses (fewest numbers of neurons), and much is known about the anatomy and physiology of the retina. These observations motivated us to anchor a model of target detectability on a model of the retinal ganglion cell responses.

In the model, cortical mechanisms also play important roles. One is to limit the spatial-frequency and orientation content in the ganglion cell responses that mask detectability of the target. The other is to pool the neural responses, in order to make perceptual decisions.

The RV1 model has two major components: a “retinal” component and a “cortical” component (Figure 18). The retinal component is grounded in the known anatomy and physiology of the eye, while the cortical component is grounded in known properties of neurons in primary visual cortex as well as empirical relations from the psychophysical literature. The retinal component simulates the responses of the midget ganglion cells (P cells) in the human/primate retina. It includes the average optical point spread function (*psf*) of the human eye (Navarro et al. 1993), local luminance gain control (G_L), which enforces Weber’s law for detection on uniform backgrounds (for reviews see Hood & Finkelstein 1986; Hood 1997), the average spatial sampling density of midget retinal ganglion cells in the human retina (Curcio & Allen 1990; Dacey 1993; Drasdo et al. 2007), and the receptive field properties of midget ganglion cells in the non-human primate retina (Croner & Kaplan 1995; Derrington & Lennie 1984). We focus on the midget ganglion cell pathway because of evidence that it is responsible for detection performance under conditions of low to moderate temporal frequency (Merigan et al. 1991; Merigan & Maunsell 1993). These conditions include the case of interest here: static stimuli presented for the duration of a typical eye fixation (150-400 ms).

The cortical component simulates the spatial pattern masking effect of the background, as well as the final pooling of responses that determines the predicted detectability of the target (d'). The spatial pattern masking is represented by an effective total contrast power (P_{eff}) that is the sum of three components: a baseline component, a narrowband component, and a broadband component. The narrowband component is

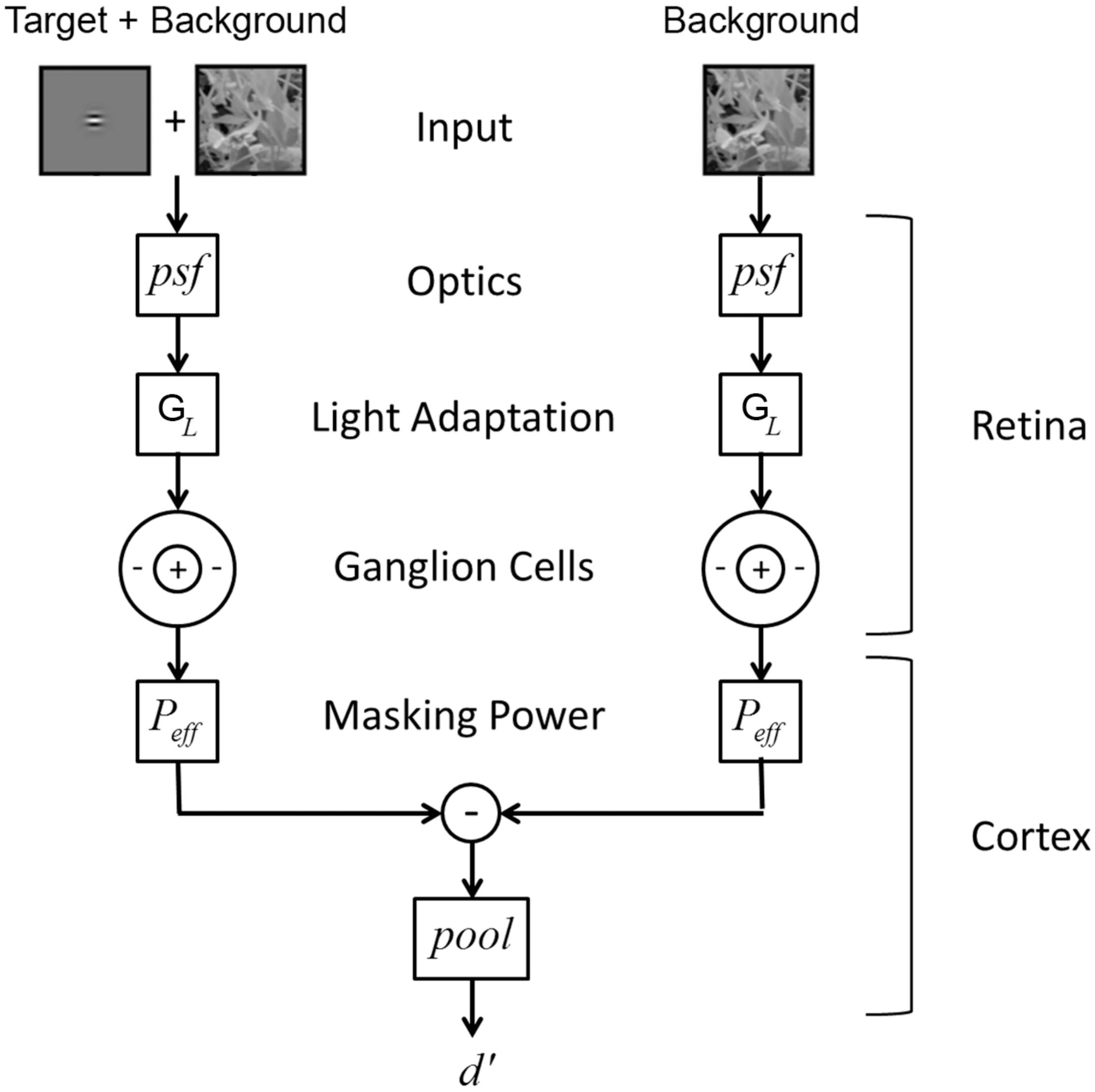


Figure 18. Schematic of the Retina-V1 (RV1) model of detection.

computed assuming filtering matched to the average spatial frequency and orientation bandwidth of neurons in monkey primary visual cortex (for reviews see, De Valois & De Valois 1988; Geisler & Albrecht 1997; Palmer et al. 1991; Shapley & Lennie 1985), which are generally consistent with estimates from the psychophysical literature (for reviews see De Valois & De Valois 1988; Graham 1989; 2011). The broadband component is consistent with the contrast normalization effects observed in cortical neurons (Albrecht & Geisler 1991; Geisler & Albrecht 1997; Heeger 1991; 1992; Carandini et al. 1997; Sit et al. 2009; Carandini & Heeger 2012) and evidenced in the psychophysical literature (Foley 1994; Goris et al. 2013; Watson & Solomon 1997). We assume that the effective total contrast power acts as an equivalent noise power in the computation of d' (Burgess & Colborne 1988; Lu & Dosher 1999; 2008; Eckstein et al. 1997a). This enforces the psychophysical rule that threshold contrast power increases linearly with background contrast power for white-noise backgrounds (Burgess, et al. 1981; Legge, et al. 1987), and for 1/f-noise backgrounds (Najemnik & Geisler 2005). This concludes a brief summary of the model; we now provide more details.

RETINAL IMAGE

The input image is either the background image alone $I_B(\mathbf{x})$ or the sum of the target and background images $I_T(\mathbf{x}) + I_T(\mathbf{x})$, where we have simplified the notation by letting $\mathbf{x} = (x, y)$. Until the final steps of the model the operations are effectively linear, and hence the target and background can be processed separately. The retinal images of the target and background are computed by convolving the target and background images with an appropriate optical point spread function:

$$T(\mathbf{x}) = I_T(\mathbf{x}) * psf(\mathbf{x}) \tag{5}$$

$$B(\mathbf{x}) = I_B(\mathbf{x}) * psf(\mathbf{x}) \tag{6}$$

In the current implementation, we use the average human point-spread function in the fovea reported in Navarro et al. (1993). The convolution is computed in the Fourier domain using their reported modulation transfer function:

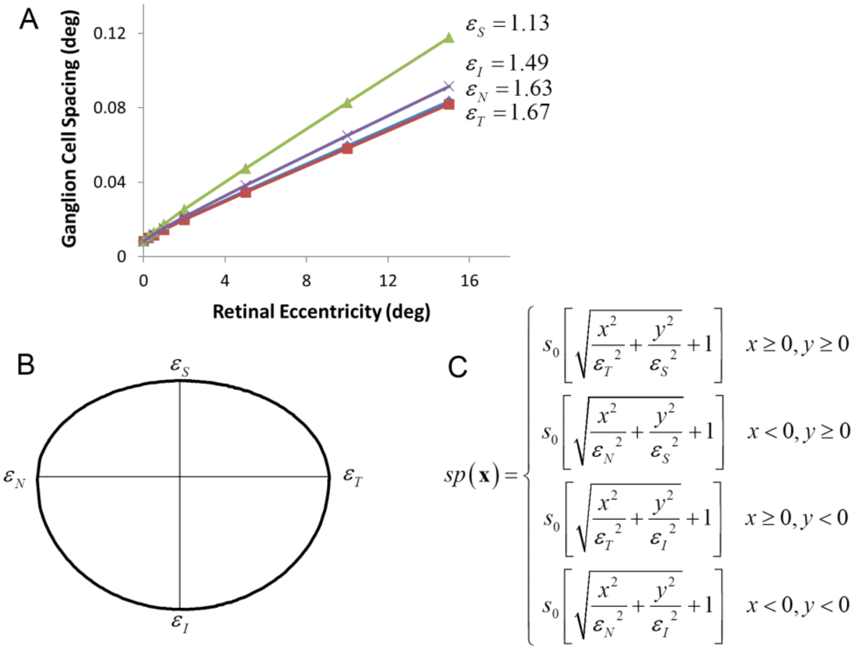
$MTF(f) = 0.78\exp(-0.172f) + 0.22\exp(-0.037f)$. The optical point spread function broadens (blur increases) with retinal eccentricity, but is relatively constant out to 10 deg eccentricity, the largest eccentricity measured in the present study. This component of the model could be easily adjusted for greater eccentricities, or to take into account individual differences in optics.

GANGLION CELL SAMPLING AND THE MAGNIFICATION PRINCIPLE

There is strong evidence that each different type of retinal ganglion cell forms a mosaic such that the dendritic branches and the receptive fields of the cells in the mosaic tile the retinal image with no gaps. Furthermore, for each cell type the percent overlap of the receptive fields is approximately constant and independent of retinal eccentricity (for a recent review, see Field & Chichilinsky 2007). This result suggests a tight link between the anatomical spacing of retinal ganglion cells and the size of their receptive fields. We exploit this fact to reduce the number of parameters in the model.

We use the average ganglion cell density reported by Drasdo et al. (2007) (6 human eyes) to generate a mosaic of midget ganglion cells. The results reported by Drasdo et al. (which are based on a reanalysis of the data in Curcio & Allen, 1990) describe the combined falloff for all types of ganglion cells. However, Dacey (1993) reports that the falloff in midget ganglion cell density in humans tracks that reported by Curcio and Allen over the first 15 deg eccentricity, the range of interest here. Thus, we assume that human midget ganglion cell density from 0 to 15 deg is proportional to the ganglion cell density reported by Drasdo et al. (2007).

The symbols in Figure 19A plot one over the square root of the density (the linear spacing) of ganglion cells as function of eccentricity in the four cardinal directions (nasal, temporal, inferior, superior), assuming that there is one midget ganglion cell (with a



$$sp(\mathbf{x}) = \begin{cases} s_0\left[\sqrt{\dfrac{x^2}{\varepsilon_T{}^2} + \dfrac{y^2}{\varepsilon_S{}^2}} + 1\right] & x \ge 0, y \ge 0 \\ s_0\left[\sqrt{\dfrac{x^2}{\varepsilon_N{}^2} + \dfrac{y^2}{\varepsilon_S{}^2}} + 1\right] & x < 0, y \ge 0 \\ s_0\left[\sqrt{\dfrac{x^2}{\varepsilon_T{}^2} + \dfrac{y^2}{\varepsilon_I{}^2}} + 1\right] & x \ge 0, y < 0 \\ s_0\left[\sqrt{\dfrac{x^2}{\varepsilon_N{}^2} + \dfrac{y^2}{\varepsilon_I{}^2}} + 1\right] & x < 0, y < 0 \end{cases}$$

Figure 19. Midget ganglion cell spacing (in degrees) in the human retina. **A**. Ganglion cell spacing (1/square-root of density) in the four cardinal directions of the visual field assuming one midget ganglion cell for each cone in the center of the fovea (which sets the y intercept). This one “midget ganglion cell,” which can respond positively and negatively, represents an on and off pair of ganglion cells. (Data from Drasdo et al. 2007.) **B**. To generate a ganglion cell mosaic we assume that in each quadrant, the contours of constant spacing fall on an ellipse. This specific contour shows the retinal locations where the spacing between midget ganglion cells is twice what it is in the center of the fovea. Thus, in the upper vertical direction the spacing doubles at about 1.1 deg of eccentricity, but in the horizontal directions it does not double until about 1.6 deg. **C**. Equation that defines the spacing function: s_0 is the spacing in the center of the fovea and $\varepsilon_N, \varepsilon_T, \varepsilon_I, \varepsilon_S$ are the eccentricities in the four cardinal directions where spacing reaches twice s_0.

linear receptive field) for each cone in the center of the fovea. The spacing of cones in the center of the fovea, s_0, is approximately 30 arc sec (0.0083 deg), and thus the

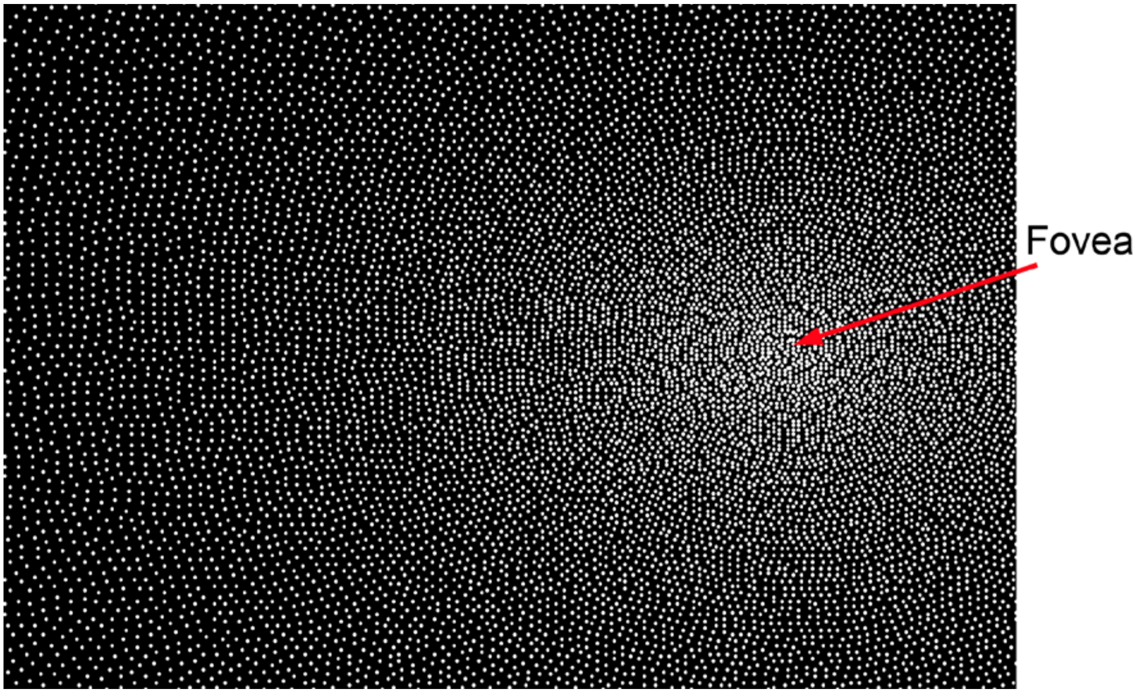


Figure 20. Part of midget ganglion cell mosaic generated from the human anatomical data in Figure 19. Each dot represents the location of the center of a ganglion cell receptive field. This mosaic is generated from the equation in Figure 19C, using the algorithm given in the Appendix.

assumed density of midget ganglion cells in the center of the fovea is 120 cells/deg. In reality there is one on and one off midget ganglion cell for each cone, and hence the actual density is approximately 240 cells/deg. However, with little loss of precision we represent the pair of on and off cells by a single linear receptive field that produces positive and negative responses. [In the current model, we ignore differences in the density and receptive field sizes of on and off ganglion cells (Dacey & Peterson 1992).] As can be seen, midget ganglion cell spacing increases approximately linearly with a slope that depends on direction in the visual field. We use these data to generate a ganglion cell mosaic. In particular, we assume that the contours of constant spacing in each quadrant of the visual field fall on an ellipse (Figure 19B). Thus, the spacing function is given by the equation in Figure 19C, where $\varepsilon_N, \varepsilon_T, \varepsilon_I, \varepsilon_S$ are the eccentricities in the four cardinal directions at which the spacing between ganglion cells reaches twice what it is in the center of the fovea. This spacing is then used to generate the ganglion cell mosaic, a portion of which is shown in Figure 20. The specific algorithm used to generate the mosaic is given in the figure caption. The algorithm produces a mosaic that satisfies the spacing function and does not have any observable artifacts. We represent the mosaic by the function $samp(\mathbf{x})$. Once the ganglion cell mosaic is specified, we then enforce the magnification principle by assuming that the receptive field properties (center and surround size) of the simulated ganglion cells scale with the spacing between ganglion cells in the mosaic. Thus, for each property, there is only a single free parameter, a scale factor, that applies to all eccentricities.

LIGHT ADAPTATION

Retinal light adaptation mechanisms maintain pattern detection and discrimination sensitivity by keeping the responses of neurons within their limited dynamic ranges. The primary effects of light adaptation can be summarized as a multiplicative luminance gain control (the signal is scaled by the inverse of the average luminance). An important perceptual effect of retinal light adaptation is Weber’s law: contrast threshold on uniform backgrounds is approximately constant independent of background luminance. To include luminance gain control we compute the local average luminance at each retinal location. Let $g_a(\mathbf{y};\mathbf{x})$ be a 2D Gaussian (with a volume of 1.0) centered on retinal location $\mathbf{x}$. Then the local average luminance at $\mathbf{x}$ is

$$L(\mathbf{x}) = \sum_{\mathbf{y}} B(\mathbf{y}) g_a(\mathbf{y};\mathbf{x}) \tag{7}$$

It is plausible that local retinal luminance gain is set by neural populations having receptive fields that increase in size with retinal eccentricity, but less is known about these populations in primates, and hence for simplicity we assume the standard deviation of the Gaussian is fixed: $\sigma_L(\mathbf{x}) = \sigma_L$. Thus, the effect of light adaptation is represented by a single parameter. The local luminance gain is $G_L(\mathbf{x}) = 1/L(\mathbf{x})$. Note that when the background is uniform then luminance gain is the same at all retinal locations (because the Gaussian has a volume of 1.0). To handle low light levels where Weber's law fails, a constant L_0 can be added to the denominator; but for the conditions of interest here that was not necessary. We note that there is also global light adaption due to slower mechanisms (pupil response, photoreceptor adaptation) that adjust the retina to the overall ambient light level in the environment. However, here we focus on stimuli where the global average luminance is fixed (displays where the average luminance is fixed across conditions), and hence we ignore the effects of global light adaptation.

GANGLION CELL RESPONSES

The spatial receptive fields of midget ganglion cells (and the corresponding P cells in the lateral geniculate nucleus) are often approximated by a difference of 2D Gaussians (Croner & Kaplan 1995; Derrington & Lennie 1984; Rodieck 1965). Using this approximation, let $g_c(\mathbf{y};\mathbf{x})$ and $g_s(\mathbf{y};\mathbf{x})$ be 2D Gaussians representing the center and surround mechanisms of a midget ganglion cell at retinal location $\mathbf{x}$ (equations are in the Appendix). The response of ganglion cells to the background alone is given by

$$r_B(\mathbf{x}) = samp(\mathbf{x}) \sum_{\mathbf{y}} G_L(\mathbf{y}) B(\mathbf{y}) D(\mathbf{y};\mathbf{x}) \tag{8}$$

where $D(\mathbf{y};\mathbf{x})$ is a difference of Gaussians: $D(\mathbf{y};\mathbf{x}) = w_c g_c(\mathbf{y};\mathbf{x}) - (1-w_c) g_s(\mathbf{y};\mathbf{x})$. For the conditions of interest here, the target contributes little to the local luminance and hence the response to the target plus background is simply the sum of the responses to the target and background, where the response of the ganglion cells to the target is given by

$$r_T(\mathbf{x}) = samp(\mathbf{x}) \sum_{\mathbf{y}} G_L(\mathbf{y}) T(\mathbf{y}) D(\mathbf{y};\mathbf{x}) \tag{9}$$

We assume that the magnification principle holds, and thus the standard deviation of the center mechanism is given by $\sigma_c(\mathbf{x}) = k_c sp(\mathbf{x})$ and the surround mechanism by $\sigma_s(\mathbf{x}) = k_s sp(\mathbf{x})$. We see then that three parameters, w_c, k_c and k_s, describe the receptive field properties of all the ganglion cells.

EFFECTIVE MASKING POWER

Masking in the RV1 model is represented by an effective contrast power P_{eff} that is the weighted sum of three components (see Figure 21); a baseline component P_0 (masking power of a uniform background), a narrowband component P_{nb}, and a broadband component P_{bb}:

$$P_{eff} = P_0 + k_b w_b P_{nb} + k_b(1 - w_b) P_{bb} \tag{10}$$

where k_b sets the overall strength of pattern masking, and w_b sets the relative strength of the narrowband and broadband components. The baseline component is a constant that represents the masking when the background is uniform. This component includes the effect of spontaneous ganglion cell activity, decision noise, and other factors not dependent on the spatial pattern of the background.

The narrowband component is the power in the ganglion cell response to the background that drives the population of primary visual cortex (V1) neurons responding to the target, and thus it is target dependent. In other words, the narrowband component represents the fact that neurons in V1 are simultaneously selective to spatial frequency and orientation, and thus will filter out background power in the ganglion cell responses that does not activate the population of V1 neurons activated by the target. In computing the narrowband component we assume that the spatial frequency selectivity of V1 neurons is approximately Gaussian in log frequency (a log Gabor function) with a bandwidth that averages 1.5 octaves (De Valois, Albrecht & Thorell 1982; Geisler &

Albrecht 1997), and that the orientation selectivity is approximately Gaussian on a circle with a bandwidth that averages 40 deg (De Valois, Yund & Hepler 1982). The log Gabor and Gaussian functions are defined in the Appendix.

The first step in computing the narrowband component is to obtain the filtered ganglion cell responses which are given by

$$r_{nb}(\mathbf{x}) = samp(\mathbf{x}) \sum_{\mathbf{y}} G_{bc}(\mathbf{y};\mathbf{x}) f_T(\mathbf{y};\mathbf{x}) \tag{11}$$

where $G_{bc}(\mathbf{y};\mathbf{x})$ is the continuous (unsampled) ganglion cell center response to the background, and $f_T(\mathbf{y};\mathbf{x})$ is the target specific filter that removes the background power in the ganglion cell responses that does not drive the cortical neurons that encode the target. To determine the target specific filter we (i) take the Fourier transform of the ganglion cell center response to the target alone, (ii) convert to log polar coordinates (log frequency vs. orientation), (iii) convolve (in the frequency domain) with a function that is the product of the amplitude spectrum of a log Gabor (bandwidth 1.5 octaves) and a Gaussian function in orientation (bandwidth 40 deg), (iv) convert back into standard spatial frequency axes and take the inverse Fourier transform. We convert to log polar coordinates so that the cortical filters at all log frequencies and orientations have the same shape, allowing simple convolution in step (iii) (Watson & Solomon 1995 use a

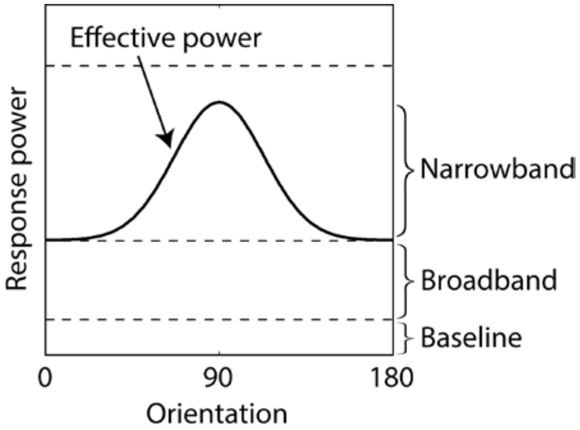


Figure 21. Effective masking power of responses to a 1/f noise background, for a Gabor target. Shown is a cross-section of the average masking power as a function of orientation, at one spatial frequency (solid curve).

similar trick). In this version of target-dependent filtering, we did not include the effect of the ganglion cell surround, because the lowest frequency (DC) is automatically removed by the log-Gabor cortical filtering. However, we have preliminary results that include both center and surround, and the quality of the model predictions is similar.

The narrowband component is given by

$$P_{nb}(\mathbf{x}) = \sum_{\mathbf{y}} r_{nb}^2(\mathbf{y}) E_T(\mathbf{y};\mathbf{x}) \tag{12}$$

where $E_T(\mathbf{y};\mathbf{x})$ is the blurred spatial envelope of the target, where the blurring depends on retinal location (i.e., envelope size increases with eccentricity). The filtered ganglion cell responses are weighted by the blurred envelope of the target under the plausible assumption that only the background power falling within some spatial neighborhood of the target will have a masking effect. The envelope is defined to be the 2D Gaussian (with arbitrary covariance matrix) that best fits the absolute value of the target (see Appendix). The blurred envelope is obtained by convolving the envelope with a 2D Gaussian having a standard deviation of the ganglion cell center σ_c (see Appendix);

The broadband component is the power in the ganglion cell responses that contributes to masking but is not spatial-frequency and orientation dependent. Such a broadband component is consistent with divisive contrast gain control (normalization) observed in cortical neurons (Albrecht & Geisler 1991; Geisler & Albrecht 1997; Heeger 1991; 1992; Carandini et al. 1997; Sit et al. 2009; Carandini & Heeger 2012) and with the psychophysical literature (Foley 1994; Goris et al. 2013; Watson & Solomon 1997). The broadband component is given by

$$P_{bb}(\mathbf{x}) = \sum_{\mathbf{y}} \left[r_B(\mathbf{y}) - r_0 \right]^2 E_T(\mathbf{y};\mathbf{x}) \tag{13}$$

where r_0 is the response of a ganglion cell to a uniform background, which is a constant that depends only on the relative weight of center and surround: $r_0 = 2w_c - 1$. Subtraction of r_0 guarantees that P_{bb} is zero for uniform backgrounds. Note that P_{nb} is also zero for uniform backgrounds, because the spatial frequency tuning of the cortical neurons is log Gabor (which goes to zero at zero spatial frequency). Finally, note that $r_{nb}\left(\mathbf{x}\right) \le r_B\left(\mathbf{x}\right) - r_0$ at all eccentricities, because the target-dependent filter can only remove background power. This implies that the masking power of the background will be greatest when the weight on the narrowband component is zero (upper dashed line in Figure 21) and least when the weight on the broadband component is zero (i.e., when the solid curve touches the baseline in Figure 21).

POOLING

We assume that the pooled response is given by the following formula:

$$r_{pooled} = \frac{\sqrt[\rho]{\sum_{\mathbf{x}} \left| r_T\left(\mathbf{x}\right)\right|^{\rho}}}{\sigma_{eff}} \tag{14}$$

where ρ is a pooling exponent, and $\sigma_{eff} = \sqrt{P_{eff}}$ is the effective masking contrast. If one regards the effective masking contrast as an equivalent noise (Burgess & Colborne 1988; Lu & Dosher 1999; 2008; Eckstein et al. 1997), then r_{pooled} can be regarded a signal-to-noise ratio. In this case, if the pooling exponent is 2.0, then equation (14) is the standard formula for optimal pooling of statistically independent Gaussian signals ("d' summation"). Following others (Quick 1974; Watson 1979; Graham 1977; Watson & Ahumada 2005), we allow the pooling exponent to be greater than 2.0 (which is suboptimal), although for the current model the estimated exponent is only slightly larger, 2.4 (see later).

DETECTABILITY AND CONTRAST THRESHOLD

The last step is to specify the relationship between the pooled response, detection threshold, and detectability. For the purpose of predicting detection performance, we define the contrast of the target in the standard way, as the target amplitude (peak gray level) divided by the mean background gray level of the whole screen. We define detection threshold c_t to be the contrast of the target at which the signal-to-noise ratio given by equation (14) is equal to 1.0, which corresponds to 69% correct. In other words, the predicted contrast threshold is the solution to the equation

$$r_{pooled}\left(c_t\right)=1 \tag{15}$$

Although this equation gives the threshold, another parameter β is required to predict the steepness of the psychometric function. Specifically, we assume that detectability has the form:

$$d'\left(c\right)=r_{pooled}^{\beta}\left(c\right) \tag{16}$$

Note that at threshold: $d'\left(c_t\right)=r_{pooled}\left(c_t\right)=1$. In the model, for a given target and background, the pooled response is linear with target contrast and hence it is easy to show that

$$d'\left(c\right)=\left(c/c_t\right)^{\beta} \tag{17}$$

Using the usual formula from signal detection theory, the predicted psychometric function is given by:

$$pcorr\left(c\right)=\Phi\left[\frac{1}{2}d'\left(c\right)\right]=\Phi\left[\frac{1}{2}\left(c/c_t\right)^{\beta}\right] \tag{18}$$

where $\Phi\left(z\right)$ is the standard normal integral function (this assumes optimal criterion placement). Note that for expository purposes we regard σ_{eff} as an equivalent noise. However, it could also be regarded as a deterministic gain control, which would make $r_{pooled}^{\beta}\left(c\right)$ a deterministic signal. A constant late decision noise would then also give equations (16)-(18).

In sum, the contrast thresholds predicted by the model are determined by only eight parameters. Five of these parameters, k_c, k_s, w_c, P_0 and ρ, determine the predicted contrast thresholds for uniform backgrounds at all retinal locations. The additional three parameters, σ_L, k_b, and w_b, determine predicted thresholds for more complex backgrounds. A ninth parameter β is needed for predicting values of detectability (d') that do not correspond to the 69% correct threshold.

IMPLEMENTATION

While the RV1 model is relatively simple conceptually, programming an efficient implementation is non-trivial, especially if one would like to rapidly compute detectability for all possible target locations and/or fixation locations, for a wide range of targets and backgrounds. The primary difficulty is that all the linear weighted summations (except the optics) are shift variant (they change with location relative to the point of fixation). To make the computations efficient we use multi-resolution stacks. Specifically, we fix the target and background images at a canonical location, centered at $\mathbf{x} = (0,0)$, and then convolve each image separately with a series of Gaussians having standard deviations that incrementally increase in powers of two. This set of images forms a stack of successively blurred images, each corresponding to a particular discrete standard deviation (resolution). We pre-compute and save these stacks for each target and background image to be processed. For each target image we also pre-compute and save the target-specific spatial-frequency filter corresponding to each level of the target stack. Once these stacks are computed and stored, they can be interpolated to rapidly determine the local luminance function, the ganglion cell target response function, and the ganglion cell effective background response function for any target location and fixation location. Specifically, each fixation location and target location specifies a spatial region of the background, as well as the spatial coordinates of the samples (ganglion cells) covering that region. The location of a sample specifies a particular continuous standard deviation (resolution). That resolution will fall between two neighboring resolutions in the stack. The value at the sample location is obtained by linearly interpolating between the two values in these neighboring resolution images. This procedure provides a close approximation to the exact calculations.

PARAMETER ESTIMATION

To estimate model parameters we minimized the squared error between the measured and predicted contrast thresholds expressed in log units (dB). Let c_i be the observed contrast threshold (in dB) for condition *i* and let $\hat{c}_i(\boldsymbol{\theta})$ be the predicted contrast threshold for parameters $\boldsymbol{\theta}$. We minimize the sum of the squared errors $S(\boldsymbol{\theta})$, and thus, $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} S(\boldsymbol{\theta})$. When the background is fixed (e.g., a uniform background) this minimization is straight forward. However, when the background randomly varies from trial to trial (the 1/f noise and natural backgrounds), it is not practical to generate a predicted model response for each trial, for each vector of parameter values evaluated during the parameter search.

To handle the case of variable backgrounds, we use the following procedure. First, we pick a random background patch for each background condition and then obtain estimates $\hat{\boldsymbol{\theta}}_1$. Once these estimates are obtained we generate the predicted threshold $\hat{c}_{ij}(\hat{\boldsymbol{\theta}}_1)$ for each specific background patch *j* in each condition *i*. Then, for each condition we rank order the thresholds and select the patch having the median threshold. Let this patch be j_i. We then estimate the parameters again, where the fixed patch for condition *i* is j_i. These estimates are $\hat{\boldsymbol{\theta}}_2$. We repeat this process until the estimated parameters converge (usually just a couple of iterations). Simulations show that this procedure is effective in finding the optimal parameters. Once the optimal parameters are estimated, a predicted threshold is computed for every background patch in every condition. The predicted threshold for a particular condition is the average of the predicted thresholds for all the background patches in that condition.

Chapter 4: Testing the Retina-V1 Model

EXPERIMENT: MEASURING DETECTABILITY IN 1/F NOISE AND NATURAL BACKGROUNDS

The goal of our model is to accurately predict detection thresholds for localized targets in arbitrary natural backgrounds at arbitrary locations in the visual field. To test the accuracy of the predictions, we measured contrast detection thresholds in a single-interval forced choice (yes-no) paradigm for three target stimuli (Gabor, Gaussian, and Edge) presented at four retinal eccentricities (0, 2.5, 5 and 10 degrees) along the horizontal meridian in the right visual field, in three different types of background (uniform, 1/f noise, and natural image). The 1/f-noise and natural-image backgrounds were presented at RMS contrast levels of 7.5% and 15%. The yes-no task was used because it more typical of real world tasks where one is not given the opportunity to compare the image with and without the target present. Also more like natural tasks, the sample of 1/f noise and natural image background was different on each trial. Thresholds were measured for three observers (two were authors in the corresponding paper).

The targets were chosen because they represent three broad categories of targets: narrowband in frequency and orientation (Gabor), broadband in spatial frequency and orientation (Gaussian), and narrowband in orientation and broadband in frequency (Edge). The background types were chosen to vary in the degree of similarity to natural backgrounds. Natural backgrounds are extraordinarily complex, differing from uniform backgrounds along a number of different dimensions. One dimension is the shape of the average amplitude spectrum, which typically falls off inversely with spatial frequency (Field 1987). Thus, as a first approximation to natural backgrounds, we used random noise backgrounds that have a 1/f amplitude spectrum. The 1/f-noise backgrounds are isotropic and stochastically stationary across space. However, natural backgrounds tend to vary across space in luminance, contrast, spatial frequency, orientation, and phase structure. Our second (closer) approximation to natural backgrounds was to include the spatial frequency, orientation and phase structure, but to control the variations in local luminance and contrast. To do this we adjusted the grayscale histograms of natural images to match those of 1/f noise with 7.5% and 15% contrast (see Stimuli). These "Gaussianized" natural images appear remarkably naturalistic (see Figure 30A in Discussion), and comparing detection performance in 1/f noise with that in Gaussianized natural images allows us to isolate the effects of spatial frequency, orientation and phase structure.

In future studies we plan to measure detection thresholds in unaltered natural backgrounds, but we focused first on Gaussianized backgrounds because they are more useful for testing our model. Because of the large variations in local luminance and contrast in natural images, there are many trials, even for a fixed amplitude target, where the target will be either trivially detectable or trivially impossible to detect. Performance on such trials is easier for a model to predict, making unaltered natural images less useful.

STIMULI

Eight-bit gray-scale images were displayed on a calibrated monitor (Sony Trinitron, GDM-FW900) at a resolution of 1920 x 1080 pixels and a frame rate of 60 Hz non-interlaced. The monitor was placed 168 cm from the eyes, and all stimuli were displayed at 120 pixels per degree. The graphics card lookup table was set to produce 256 linear steps in luminance with a mean luminance of 18 cd/m^2. There were three target stimuli in our experiment: Gabor, Gaussian, and Edge. The Gabor was horizontal, at 4 cycles per degree, in cosine phase, and had a bandwidth of one octave. The Gaussian had a standard deviation of 8.43 arc min. The Edge was horizontal and windowed with a

Gaussian having a standard deviation of 0.5 degrees. These three targets were taken from the ModelFest stimulus set (ModelFest stimuli #12, #27 and #30, see Watson & Ahumada 2005). Targets were presented at the center of a 512 x 512 background located within a larger mean luminance background (18 cd/m^2, 1920 x 1080). Depending on the background condition, the 512 x 512 background was either set to mean luminance, or randomly selected from either large 1/f noise images (1280 x 1280) or from one of 10 large (4284 x 2844) "Gaussianized" natural images. In all conditions, the pixels on the edge of the 512 x 512 background were set to black; this created a 1-pixel wide box that cued the location of the background under all conditions. Detection measurements were obtained for uniform backgrounds, and for 1/f-noise and natural backgrounds of 7.5% and 15% RMS contrast (i.e., 5 background conditions).

The natural images were randomly selected from a set of 1200 calibrated natural images (available at www.cps.utexas.edu/natural_scenes), and both the 1/f-noise and natural images were converted to 8-bit gray scale. The natural images were "Gaussianized" by matching their grayscale histograms to a large 1/f noise image. The first step was to rank-order the pixels in each image according to gray level from smallest to largest. Note that for each specific gray level, the fraction of pixels having that gray level will differ between the two images. The goal was to make the fraction of pixels at each gray level in the natural image the same as that in the 1/f noise image. This was done in the second step by applying the following mapping: $g_i = f_j$, where g_i is the gray level of the natural-image pixel having rank order *i* out of a total of N pixels, f_j is the gray level of the 1/f noise pixel having rank order *j* out of a total of M pixels, with $j = \lceil iM/N \rceil$. (Note that $N > M$, and $\lceil x \rceil$ is the “ceiling” function.) This mapping preserved the spatial frequency, orientation, and phase structure of natural images, but allowed us to select patches from Gaussianized natural images with similar mean luminance and contrast as patches selected from our large 1/f noise images. Specifically, for each randomly selected 512x512 patch of 1/f noise used in the experiment, we randomly selected a patch of Gaussianized natural image having approximately the same mean luminance (the mean luminance differed by a maximum of 1.45 cd/m^2); the RMS contrasts of the two patches were set to the same value (i.e., 7.5% or 15%).

PROCEDURE

Psychometric functions were measured in a single-interval, blocked, forced-choice paradigm where the observer judged whether a target was present or absent at the center of the background. Each psychometric function was based on at least 240 trials, collected in separate sessions of 120 trials each. For a given condition in a session, four blocks of 30 trials each were run in descending order of target contrast. Eye position was monitored using an Eyelink 1000 eye tracker. If eye position deviated by more than 1 deg from the fixation dot, the trial was discarded and another trial added to the block. Each 30-trial block began with a standard 9-point calibration procedure for the eye tracking. After the calibration procedure, the observer was required to hold fixation on a fixation dot for each of the 30 trials in the block. Each trial began with a 500 ms interval in which the background location was cued with a one-pixel wide black square outlining the background area. In conditions where the target location was the center of the fovea, the fixation dot was extinguished 100 ms before onset of the test stimulus. The test stimulus consisted of a 250 ms presentation of either background or background-plus-target. At the end of this interval, there was a 2 sec response window (mean luminance background) during which the observer could signal "target present" or "target absent" by pressing one of two buttons. Failure to respond led to the trial being replaced with a new one; this occurred less than 1% of the time. Feedback was given at the end of the 2 sec response window with a high tone representing "correct" and a low tone representing "incorrect". The next trial began immediately after feedback was given. Psychometric functions were measured for 60 separate conditions (3 stimuli x 4 eccentricities x 5 background conditions). The psychometric functions with uniform and 1/f noise backgrounds were measured in a random order. Then the psychometric functions for the Gaussianized natural backgrounds were measured in a random order.

FITTING PSYCHOMETRIC FUNCTIONS AND THRESHOLDS

As mentioned earlier, we used a yes-no task because it is more typical of natural conditions. Performance in all forced choice tasks can be influenced by criterion bias, but yes-no tasks are often thought to be more susceptible. Therefore, for each condition (comprising at least 240 trials), we obtained maximum-likelihood estimates of the threshold (c_t), steepness parameter (β), and criterion (γ). Consistent with equation (18), the probability of a hit is given by

$$P_h = \Phi\left(\frac{1}{2}\left(\frac{c}{c_t}\right)^{\beta} - \gamma\right) \tag{19}$$

and the probably of a false alarm by

$$P_{fa} = \Phi\left(-\frac{1}{2}\left(\frac{c}{c_t}\right)^{\beta} - \gamma\right) \tag{20}$$

Thus, the log likelihood of all the responses from a condition is

$$\begin{aligned} &\ln L\left(c_t, \beta, \gamma\right) = \\ &\sum_{i=1}^{n} N_h\left(c_i\right)\ln P_h\left(c_i\right) + N_m\left(c_i\right)\ln\left(1 - P_h\left(c_i\right)\right) + N_{fa}\left(c_i\right)\ln P_{fa}\left(c_i\right) + N_{cr}\left(c_i\right)\ln\left(1 - P_{fa}\left(c_i\right)\right) \end{aligned} \tag{21}$$

where n is the number of contrast levels of the target, and $N_h(c_i)$, $N_m(c_i)$, $\mathrm{N}_{\mathrm{fa}}(c_i)$, $\mathrm{N}_{\mathrm{cr}}(c_i)$ are the numbers of hits, misses, false alarms, and correct rejections, for contrast level c_i. We first estimated the parameters by maximizing equation (21). We found that the values of the steepness parameter were consistent across conditions (see Results) and that there were no systematic variations in the criterion across conditions for a given observer.

Thus, the final thresholds for each observer were obtained by setting the steepness parameter to the average across all subjects and conditions, setting the criterion to the average across conditions for that subject, and then finding the maximum likelihood estimate of the thresholds using equation (21). Importantly, the pattern of thresholds was robust across different versions of this analysis (including ignoring criterion effects and only analyzing percent correct).

Results

Maximum likelihood fits of equations (19) and (20) to the psychometric data were used to obtain the estimated contrast threshold c_t for each of the 60 conditions. Figure 22

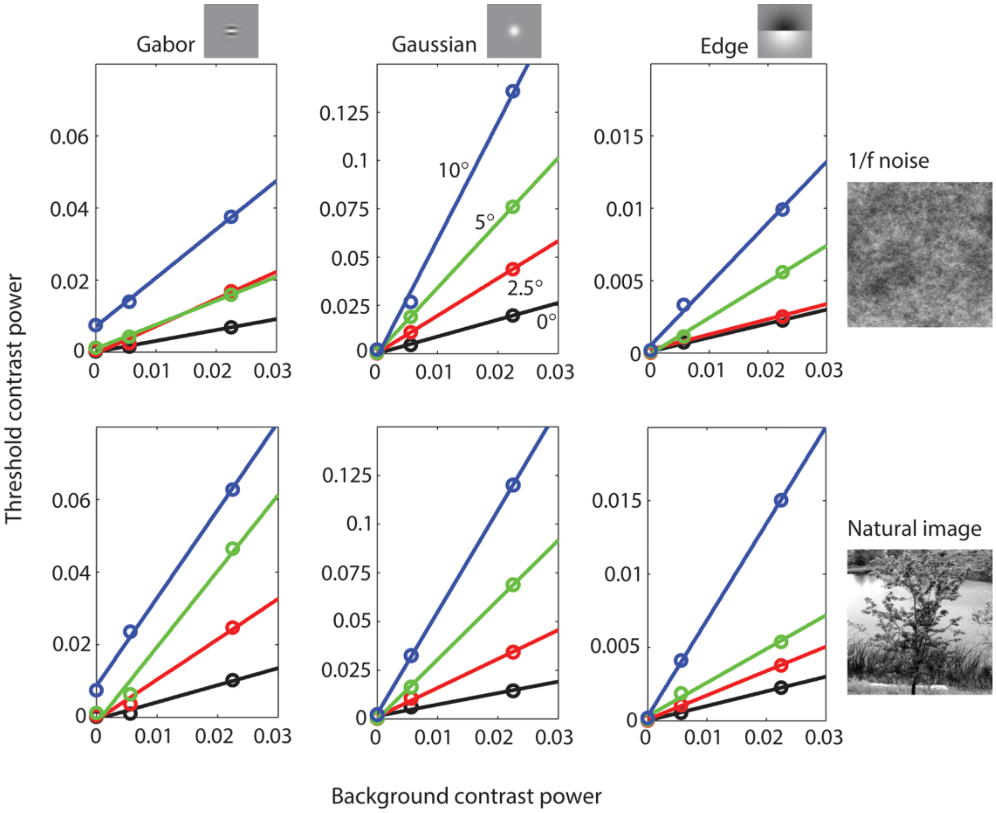


Figure 22. Detection threshold measurements for three different targets, at four different eccentricities as function of background contrast power, for 1/f noise and natural backgrounds. Data points are the average of three observers. The solid lines are best fitting linear functions.

plots the square of the estimated contrast thresholds (threshold power) as a function of square of the background contrast (background power). The open circles represent the average thresholds of three observers for three target stimuli (columns) presented at 4 retinal eccentricities (colors) in 1/f noise and Gaussianized natural backgrounds (rows). The colored lines are linear fits to the data (not model predictions). Note the thresholds measured in uniform backgrounds (background contrast of zero) are the same in both rows of plots, and that the vertical scales are different for the different targets. The estimated criterion (bias) for the three observers in units of d-prime were: 0.362 (JSA), 0.228 (CKB), 0.277 (SPS).

Two principles of masking are suggested by these plots: 1) threshold contrast power increases linearly as a function of background contrast power, and 2) the slope of the best fitting line increases as a function of retinal eccentricity. Figure 23 shows more clearly how well our data are described by linear masking functions. In this figure, the

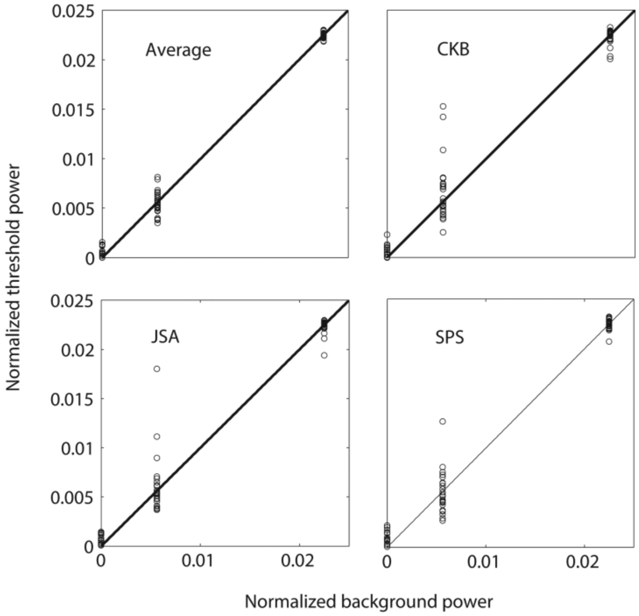


Figure 23. Normalized contrast threshold power as a function of background contrast power for all experimental conditions.

data from each subject has been normalized so that the linear fits have an intercept of 0 and a slope of 1; also shown are the average thresholds (Fig. 23a). If threshold contrast power is a linear function of background contrast power then the normalized data points should fall on a line of slope 1 through the origin (black line).

Although not easily seen in Figure 22, the intercepts of the masking function also increase with retinal eccentricity. Figure 24 plots the intercept as a function of retinal eccentricity for each type of target. The intercepts tend to increase exponentially with eccentricity (solid curves). In general, the thresholds in 1/f noise and in natural backgrounds are similar (see Fig. 22). However, for the Gabor and Edge targets masking was somewhat greater in the natural backgrounds. This can be seen clearly in Figure 25, which plots threshold in 1/f noise as a function of threshold in natural backgrounds separately for each target. The points for the Gaussian target fall near or slightly below

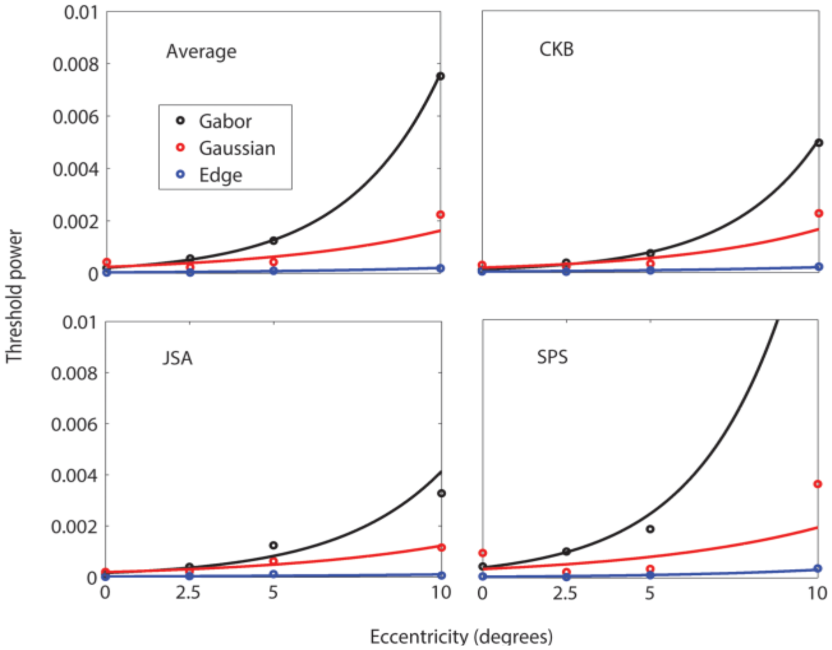


Figure 24. Threshold contrast power as function of eccentricity when background is uniform. Solid curves are best fitting exponential.

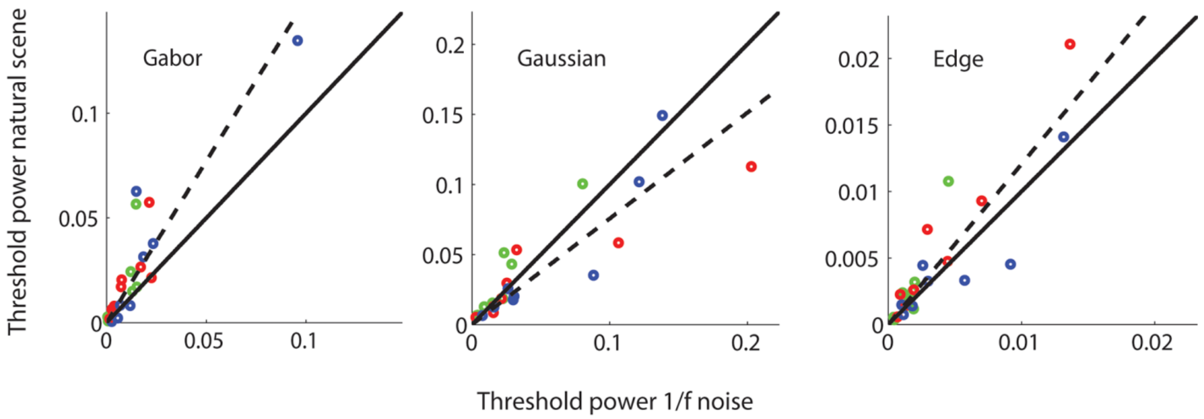


Figure 25. Threshold in Gaussianized natural images as a function threshold in 1/f noise, for all conditions, for the three subjects: CKB (green symbols), SPS (blue symbols), JSA (red symbols). Dashed curve is best fitting line through the origin.

the diagonal, but the points for the Gabor and Edge target fall above the diagonal. Even though the points do fall off the diagonal, they fall roughly on straight lines, indicating that thresholds in 1/f noise and natural backgrounds differ approximately by a fixed proportionality constant that depends on the target.

The slopes of the psychometric functions were fairly constant over the 60 conditions. Figure 26a shows that the steepness parameter varies little across the three

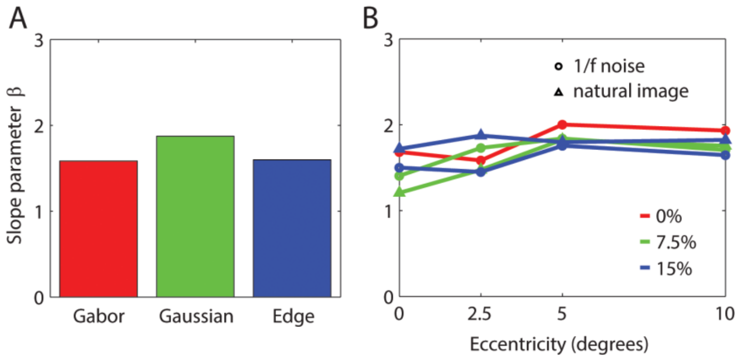


Figure 26. Psychometric function slope parameter. **A**. Average slope parameter values for three types of target. **B**. Average slope parameter as function of eccentricity for type of background and background contrast.

types of target. Figure 26b plots the average steepness parameters of the three targets for all background and retinal eccentricity conditions. On average there is a slight trend for the parameter to increase with eccentricity. Overall the average steepness parameter is 1.685. We take this average parameter value to be the estimate of β in the RV1 model.

MODEL PREDICTIONS

In fitting the model to the estimated thresholds we use equation (15), and then equation (18) if we need to predict thresholds for a different criterion percent correct (e.g., 82% rather than 69%). In what follows, we estimate a subset (five) of the parameters by fitting the model to the average thresholds (based on 16 observers) reported in the ModelFest study. Keeping these parameters fixed (except for P_0 that scales all thresholds up and down), we estimate the remaining three parameters by fitting the average thresholds from the present experiment. Finally, in the discussion section we keep the parameters fixed (except for P_0) and generate predictions for the results of Foley et al. (2007). Figure 27 shows the predictions of the RV1 model (red line) to the ModelFest data set (black points), which is the average contrast thresholds of 16 observers across 27 labs for foveal detection of 43 target stimuli in a uniform background. The thresholds in the ModelFest set are based on an 82% correct criterion. The thresholds are plotted in dB units. Thus, a 6 dB difference in threshold corresponds to a factor of two in contrast threshold. The RMS error of the model is comparable to the RMS error of the better models tested by Watson and Ahumada (2005). Recall these predictions depend on five parameters: three that control the relative size and strength of the center and surround, a pooling exponent, and a baseline noise parameter. The values of the parameters are given in the figure caption.

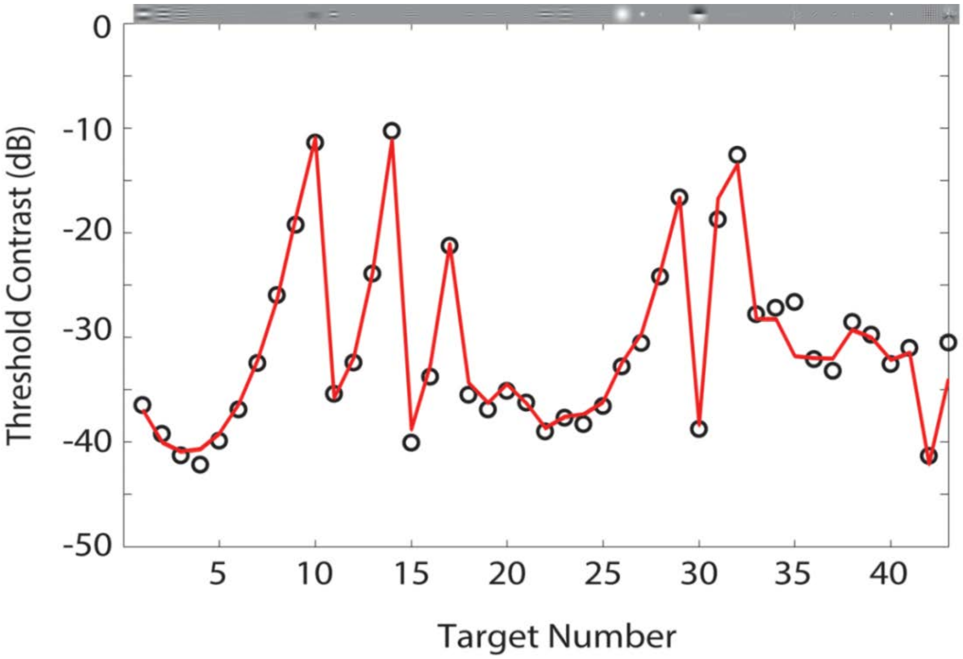


Figure 27. Predictions for ModelFest data set. Data points are average contrast thresholds of 16 observers in 10 labs, for 43 different targets. The solid curve is the prediction of the RV1 model. Parameter values: $k_c = 1$, $k_s = 10.1$, $w_c = 0.53$, $\rho = 2.4$, $\beta = 1.685$, $P_0 = 1.4E - 3$. Note that changing P_0 translates the entire predicted curve vertically on the logarithmic (dB) scale. Threshold contrast in dB = $20\log(c_t)$, where here c_t is the 82%-correct contrast threshold (RMS error = 1.09 dB).

Figure 28 shows the predictions (solid curves) of the model for the present experiment (plotted in dB units rather than contrast power as in Fig. 5). The open circles show the average contrast thresholds of the 3 observers for three target stimuli (columns) as a function of retinal eccentricity, for three background contrasts (colors), in 1/f noise and Gaussianized natural backgrounds (rows). The plotted thresholds for detection in the uniform background (black open circles) are the same in both rows. In order to maximize compatibility with the ModelFest data, these thresholds are also based on the 82% correct criterion. In fitting these data we kept the parameters values obtained from fitting the

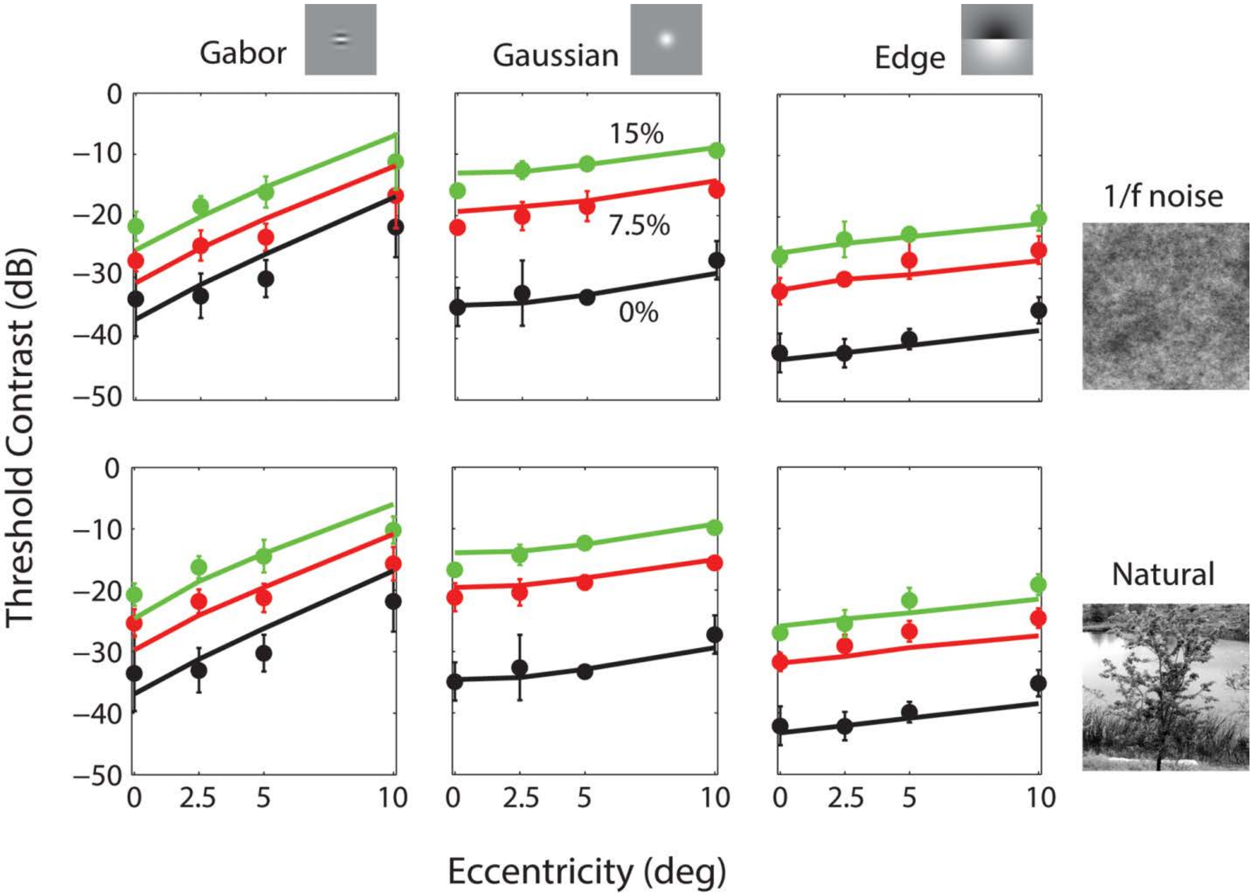


Figure 28. Predictions for the data from the present experiment. Data points are the average contrast thresholds of 3 observers for 60 different conditions (3 targets x 4 eccentricities x 3 background contrast levels, for 2 kinds of background; the 0% background contrast is a uniform field, and hence the black points are the same in the upper and lower plots). Error bars represent ± 2 standard errors (across observers). Parameter values: $k_c = 1$, $k_s = 10.1$, $w_c = 0.53$, $\rho = 2.4$, $\beta = 1.685$, $P_0 = 4.45E - 4$, $\sigma_l = 1°$, $k_b = 25$, $w_b = 0.962$. The first six parameters are the same as in Figure 27, except for P_0, which accounts for (modest) differences in overall sensitivity between groups of observers. (RMS error = 2.27 dB)

ModelFest data, with the exception that we allowed the baseline noise parameter P_0 to change. The only effect of the baseline noise parameter is to shift the predictions for uniform backgrounds vertically on the dB axis. Although we allowed the baseline noise parameter to vary, the estimated value was well within the range of individual differences for that parameter in the ModelFest dataset. Recall that the model has three additional parameters for predicting thresholds in non-uniform backgrounds: overall pattern masking strength, the relative weight of the narrowband and broadband components, and the spatial area for local luminance gain control. Again, the values of the estimated parameters are given in the figure caption. As can be seen, the model captures most of the variance in the thresholds, but is qualitatively more accurate for the Gaussian and Edge targets than for the Gabor target. Note that the foveal thresholds for the three targets on the uniform background are similar to those in ModelFest data set (our targets correspond to ModelFest stimuli #12, #27 and #30 in Figure 27).

UNIFORM BACKGROUNDS

The predictions of the RV1 model for uniform backgrounds are determined by six parameters. One of these parameters, β, was determined from the average steepness of the psychometric functions in the experiment reported here. We find $\beta = 1.685$, which we note corresponds to a Weibull slope parameter of 2.13. The remaining five parameters can be estimated from the detection thresholds measured on uniform backgrounds in the fovea. To estimate these parameters we fit the ModelFest dataset which consists of foveal detection thresholds measured for 43 different targets on 16 observers in 10 different laboratories. The fit of the model to the ModelFest data is good; comparable to (slightly worse than) the best non-physiologically based models (see Watson & Ahumada 2005).

The estimated parameters for the midget ganglion cell receptive fields are reasonably consistent with the anatomy and physiology of the primate retina. Our psychophysical estimate of the standard deviation of the ganglion cell center mechanism σ_c is almost exactly equal to the spacing between the (on or off) midget ganglion cells, which in the central visual field is approximately equal to the spacing between the photoreceptors (about a half minute of arc). This is consistent with the anatomical finding that in the central visual field a midget ganglion cell synapses with one midget bipolar cell, which synapses with one cone photoreceptor. The measured width of center mechanisms with single-unit recording is larger than a single cone, but the larger size is expected because of the effect of the optical point-spread function; the measured center mechanism should be the convolution of the physiological center mechanism and the optical point spread function. Croner & Kaplan (1995) report that in the central 5 degrees the median standard deviation of the center mechanism is 0.03 deg, and of the surround mechanism is 0.18 deg (about 6 times larger than the center). We computed the effective center standard deviations for our model and find that they range from 0.021 deg at 0 deg eccentricity to 0.038 at 5 deg eccentricity, spanning the value reported by Croner & Kaplan. Similarly, the effective surround standard deviation for the model ranges from 0.077 (3.6 times larger than center) at 0 deg eccentricity to 0.3 (7.9 times larger) at 5 deg eccentricity. Finally, Croner & Kaplan report that the relative weight on the center mechanism w_c is about 0.64, whereas our estimate is 0.53. Thus, we also find greater weight for the center mechanism, but not by as large a factor.

The ModelFest dataset only contains measurements made in the center of the fovea. In the present experiment we made measurements for three of the ModelFest targets at four eccentricities (black circles in Figure 28), and obtained reasonable predictions (solid curves) without altering parameters, except that we allowed the

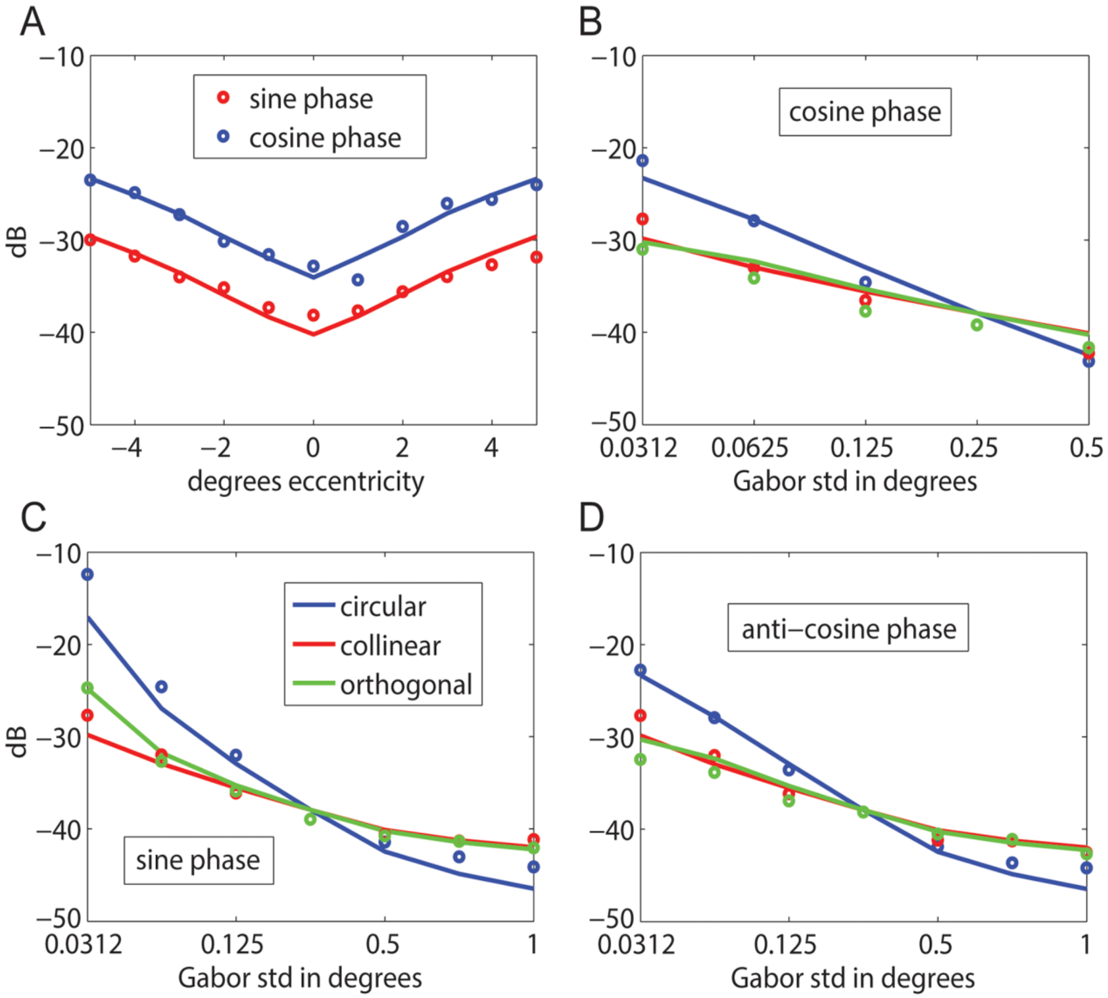


Figure 29. Predictions for data from Foley et al. (2007). **A**. Threshold as a function of eccentricity for 4 cpd radially-symmetric Gabor targets in sine phase (envelope SD = 0.25 deg; 2 observers; $P_0 = 5.4E-4$), and cosine phase (envelope SD = 0.18 deg; 3 observers; $P_0 = 1.3E-3$). **B-D** Threshold as a function of envelope standard deviation for 4 cpd Gabor targets with a circular envelope, an envelope elongated collinear with the grating, and an envelope elongated orthogonal to the grating (2 observers; $P_0 = 9.2E-4$). Solid curves are the predictions of the RV1 model with same parameters as in Figure 10, except for P_0**,** which accounts for (small) differences in overall sensitivity between groups of observers. (RMS error = 1.28 dB).

baseline masking power P_0 to change from 1.4E-3 to 4.5E-4 to account for modest differences in overall sensitivity among different groups of observers. As a further test of the model, we generated predictions for the detection thresholds reported in Foley et al. (2007). In their Experiment 1, Foley et al. measured thresholds for vertical 4 cpd Gabor targets at retinal eccentricities ranging from -5 to 5 deg along the horizontal meridian. In three observers thresholds were measured for a cosine-phase Gabor having an envelope standard deviation (SD) of 0.25 deg. In two other observers, thresholds were measured for a sine-phase Gabor having an envelope SD of 0.18 deg. The symbols in Figure 12A show the average thresholds. The solid curve shows the prediction of the RV1 model without altering parameters, except for the baseline masking power (see figure caption). In their Experiment 2, Foley et al. measured thresholds in the fovea for 4 cpd Gabor targets in cosine phase (Figure 29B), sine phase (Figure 29C), and anti-cosine phase (Figure 29D), for various areas and aspect ratios, in two observers. The blue symbols show the thresholds for Gabor targets with a radially symmetric envelope. In this case, the horizontal axis gives the standard deviation of the envelope in all directions. The red symbols show the thresholds for Gabor targets that are elongated parallel to the orientation of the grating. In this case, the horizontal axis gives the standard deviation of the envelope in the parallel direction, where the standard deviation in the perpendicular direction is fixed at 0.25 deg. The green symbols show the thresholds for Gabor targets that are elongated perpendicular to the orientation of the grating. In this case, the horizontal axis gives the standard deviation in the perpendicular direction, where the standard deviation in the parallel direction is fixed at 0.25 deg. The solid curves show the predictions of the RV1 model. We have only evaluated the predictions of the model out to 10 deg eccentricity. However, if it works well over this range then the literature suggests that it would apply over a wider range (Peli, et al. 1991).

The relatively good fit of the model to all the uniform background data, and the reasonable agreement of the estimated parameters with retinal anatomy and physiology, suggest that optical and retinal factors may be the primary factors causing the variation in detection thresholds across different targets on uniform backgrounds. This is not implausible given that the optic nerve is arguably the major information transmission bottleneck in the visual pathway, making it possible for cortical circuits to process the ganglion cell responses with relatively constant efficiency across the different targets. The largest errors (underestimates) of the thresholds in Figure 27 occur for the two spatially complex targets (binary noise, #34, and cityscape, #43), for which it is reasonable to expect reduced central efficiency in pooling all the relevant features.

NON-UNIFORM BACKGROUNDS

The predictions of the RV1 model for patterned backgrounds depend on three additional parameters. To estimate these remaining parameters and provide a further test of the model we measured psychometric functions for Gabor, Gaussian and Edge targets at four different eccentricities in uniform backgrounds, in 1/f-noise backgrounds, and in natural backgrounds whose gray-scale histogram has been adjusted to match that of 1/f noise. The predictions are good, but slightly poorer for the Gabor target than for the Gaussian and Edge targets (see Figure 28). It is interesting to note, however, that the average thresholds reported by Foley et al. (2007) for the Gabor target (Figure 29a) increase slightly faster with eccentricity, in better agreement with the RV1 model.

Perhaps the most remarkable result is that the model does about as well predicting detection thresholds in Gaussianized natural backgrounds as it does in 1/f-noise backgrounds, and that the thresholds for the two kinds of background are similar. The background masking effects in the model are entirely based on the narrowband and broadband power in the ganglion cell responses, not on the specific phase structure, which differs greatly between the natural-image and 1/f-noise backgrounds. Perhaps the trial-to-trial variation in the backgrounds is hiding the effect of the phase structure. That is, thresholds may be similar in the two types of background only because on some trials the phase structure helps detection and on other trials it hurts detection. However, if this were true then one might expect shallower psychometric functions for natural backgrounds. In fact, the slope parameter of the psychometric functions is similar for uniform, 1/f-noise, and Gaussianized natural backgrounds (see Figure 26). It would appear that for Gaussianized natural backgrounds, the complex phase structure of natural backgrounds has, practically speaking, a relatively minor effect on detection thresholds.

A limitation of our test of the RV1 model for patterned backgrounds is that it is based on data for only three different targets. However, note that the pattern-masked thresholds for these three targets tend to parallel (on a log scale) the thresholds obtained on a uniform background (see Figure 28). This suggests that the pattern-masked thresholds for other ModelFest targets would also tend to parallel those obtained on a uniform background. Thus, it seems likely that the predictions of the RV1 model would be of similar accuracy for the other ModelFest targets, given the accuracy of its predictions for the other ModelFest targets on uniform backgrounds.

Chapter 5: Discussion

COMPONENTS IN THE MODEL AND COMPONENTS NOT IN THE MODEL

The RV1 model contains a number of different components, and they each play an important role in the predictions. The optical point spread function has a substantial effect on the shape of the contrast sensitivity function (especially the high-frequency falloff) and on how rapidly thresholds rise with eccentricity – thresholds for high-frequency targets would rise more rapidly without the effect of the optics, because the effective ganglion cell center size would grow more rapidly. Obviously, the discrete sampling function has a big effect. The number of samples declines rapidly with eccentricity, and hence the maximum amount of retinal image information transmitted by the ganglion cells for high-frequency and broadband (e.g., natural or 1/f noise) images drops rapidly. The continuous variation in ganglion-cell receptive field size with the sample spacing is also important. For example, consider the contrast sensitivity function (CSF) in the fovea. In Figure 27, the thresholds for stimuli 1-10 give the CSF for targets with a fixed spatial extent, and the thresholds for stimuli 11-15 give the CSF for targets with a fixed numbers of cycles. These CSFs are not well approximated by a difference of Gaussians (Watson & Ahumada, 2005), which is the shape of the ganglion cell receptive fields. The relatively accurate prediction of the RV1 model is due in part to the fact that there is a distribution of ganglion-cell receptive field sizes falling under the stimuli.

In agreement with the masking literature (Foley 1994; Solomon & Watson 1997; Eckstein et al. 1997b) we find that both the narrowband and the broadband masking components are important. If parameters are estimated with the weight on the narrowband component set to zero ($w_b = 0.0$, see eq. 6), then the predictions are substantially worse. Conversely, if the parameters are estimated with the weight on the broadband component set to zero, then predictions are also substantially worse. Although the estimated weight is higher on the narrowband component ($w_b = 0.962$) than the broadband component ($1 - w_b = 0.038$), they both play an important role. In fact, the average total masking power due to the broadband and narrowband components is about equal across the three targets: $w_b P_{nb} \cong (1 - w_b) P_{bb}$. More specifically, the average ratio of narrowband to broadband masking power is smallest for the Gabor target (0.165), intermediate for the Edge target (0.98), and largest for the Gaussian target (2.27). Although we find that both components are important in the current version of the model, the result may depend on how the target-dependent filter is computed. It is perhaps worth emphasizing that broadband and narrowband components have no effect on the predictions for uniform backgrounds.

There are some well-known components that are not included in the RV1 model. One is a component that would produce the oblique effect—foveal detection thresholds tend to be higher for gratings oriented along the diagonals (Campbell et al. 1966; McMahon & MacLeod 2003). This effect is most likely cortical in origin (McMahon & MacLeod 2003). We left out this component because the underlying anatomy and neurophysiology are not well understood, and because including the oblique effect produces only minor improvements in prediction accuracy for the stimuli tested here (Watson & Ahumada 2005). However, it would not be difficult to include in the model. A second missing component is one that would produce the dipper effect – when the masker has the same (or nearly the same) shape as the target, then detection threshold reaches a minimum (dips) when the contrast of the masker is itself at or near detection threshold (Legge & Foley 1980). The dipper effect has been modeled with an accelerating (or threshold) nonlinearity prior to late noise (e.g., Legge & Foley 1980; Foley 1994; Solomon & Watson 1997; Goris et al. 2013). We left out this component because it would reduce the computational efficiency of the RV1 model (which depends on linearity), and because the dipper effect is likely to occur very infrequently under natural conditions. The dipper effect reduces or disappears if the target and masker differ just a little in shape (e.g., spatial-frequency bandwidth of the dipper effect is about 0.25 octaves, ref), which they generally do in 1/f noise or natural images. A third missing component is one that would produce some of the stronger crowding effects – identification of targets can be strongly suppressed by the presence of surrounding objects/textures that are sufficiently similar to the target (for a review see Levi 2008). We did not try to include explicit crowding mechanisms because the current aim is to predict detection rather than identification performance (detection is a special case of identification). However, it is interesting that the RV1 model is able to predict detection in natural backgrounds without including the kinds of mechanisms (extended feature/texture integration) thought to underlie crowding. For example, natural backgrounds are filled with edges of various scales and orientations, yet threshold for the edge target across the visual field is accurately predicted from only the background power falling under the envelope of the target (note the envelope expands slightly with eccentricity, see Appendix). Like crowding paradigms, doesn't detection in this case involve identifying whether the specific target is present as opposed to whatever other edge shape or object might be at that location? (Recall that in the present Yes/No task, the observer does not get to compare target + background with background alone.) Perhaps the success of the model is because in natural scenes (and in 1/f noise) the target is on average not very similar to the background surrounding the target. This raises the question: How important are crowding effects when looking for specific targets in natural scenes? If one takes an arbitrary target and adds it at a random location in a natural image, then does the target tend to be sufficiently similar to the surrounding background for crowding effects to be strong relative to the more local masking effects? It may be possible to answer this question by analysis of natural image statistics. Of course, in some natural cases crowding effects are known to be very important (e.g., reading), and in other cases are likely to be very important (e.g., detecting animals, which often mimic the backgrounds in their natural habitat).

COMPARISON TO PREVIOUS MODELS

The RV1 model borrows heavily from previous models of pattern detection (as indicated by the references in earlier sections), but has several unique features. First, the model directly incorporates physical measurements of the optics of the eye and of the anatomy and physiology of retinal ganglion cells, extending an earlier attempt to do this (Arnow & Geisler 1996). This approach exploits known physical and physiological constraints and hence reduces the number of free parameters. Second, there are few models of pattern detection that explicitly model the variation in spatial resolution across the visual field. Indeed most models focus exclusively on detection in the fovea, which reduces their generality and utility. Third, the model takes into account the spatial frequency and orientation selectivity of cortical populations (channels) by applying a target-dependent filter (that varies with eccentricity) to the modeled ganglion cell responses. This approach allows for very efficient computation, while still representing the information processing carried out by the (very large) cortical population. Fourth, the implementation of the model makes extensive use of Gaussian stacks, which make it possible to rapidly generate predictions for arbitrary locations across the visual field even though the visual system is highly inhomogeneous (shift variant).

There are a number of previous pattern masking models that include narrowband spatial-frequency channel masking together with broadband contrast masking (e.g., Foley 1994; Eckstein et al. 1997b; Solomon & Watson 1997; Rohaly et al. 1997; Goris, et al.

2013), and some of these have been applied to detection of targets in natural backgrounds (e.g., Eckstein et al. 1997b; Rohaly et al. 1997). These models differ from the present model in that they explicitly represent the spatial frequency channels rather than implicitly with a target-dependent filter, and they do not explicitly represent the variation is spatial resolution across the visual field. As mentioned above the models of Foley (1994), Solomon & Watson (1997), and Goris et al. (2013) include an accelerating nonlinearity to account for the dipper effect, which the present model does not.

Another class of pattern masking model is related to ideal detection in noise (Burgess, et al. 1981; Legge, et al. 1987; Myers & Barrett 1987; Eckstein et al. 1997a; Zhang et al. 2006; Burgess 2011). These models typically involve first characterizing the statistical properties of the backgrounds, plus any assumed internal noise properties or constraints, and then deriving a model (ideal) observer that is optimal given those statistical properties and constraints. This is a more principled approach that often yields nearly parameter free predictions and can provide deeper insight into neural computation (Geisler 2011). This approach has been extensively developed in the area medical imaging perception (e.g., see Samei 2010; Zhang et al. 2006; Burgess 2011). The RV1 model does not directly consider the statistical properties of backgrounds and hence is not an ideal observer model; however, it borrows from this approach by regarding the baseline, narrowband, and broadband masking effects as a combined equivalent noise in a signal detection framework. In the future an ideal observer analysis that includes the biological constraints represented in the RV1 model may provide deeper insights into, and new predictions for, the neural computations underlying pattern detection in natural scenes. Nonetheless, the RV1 model may prove useful because it is: (i) based directly on known biological constraints, (ii) contains few parameters, (iii) is extensible, (iv) takes images of the background and target as input and produces a predicted performance (d’) or predicted response (Yes/No) as output, (v) can generate predictions across the visual field for arbitrary backgrounds, and (vi) is computationally efficient.

DETECTABILITY MAPS

The computational efficiency of the RV1 model makes it possible to generate maps of target detectability across the visual field for arbitrary backgrounds. Figure 30A shows a Gaussianized natural image that is 24 deg across. Figures 30b-d illustrate three

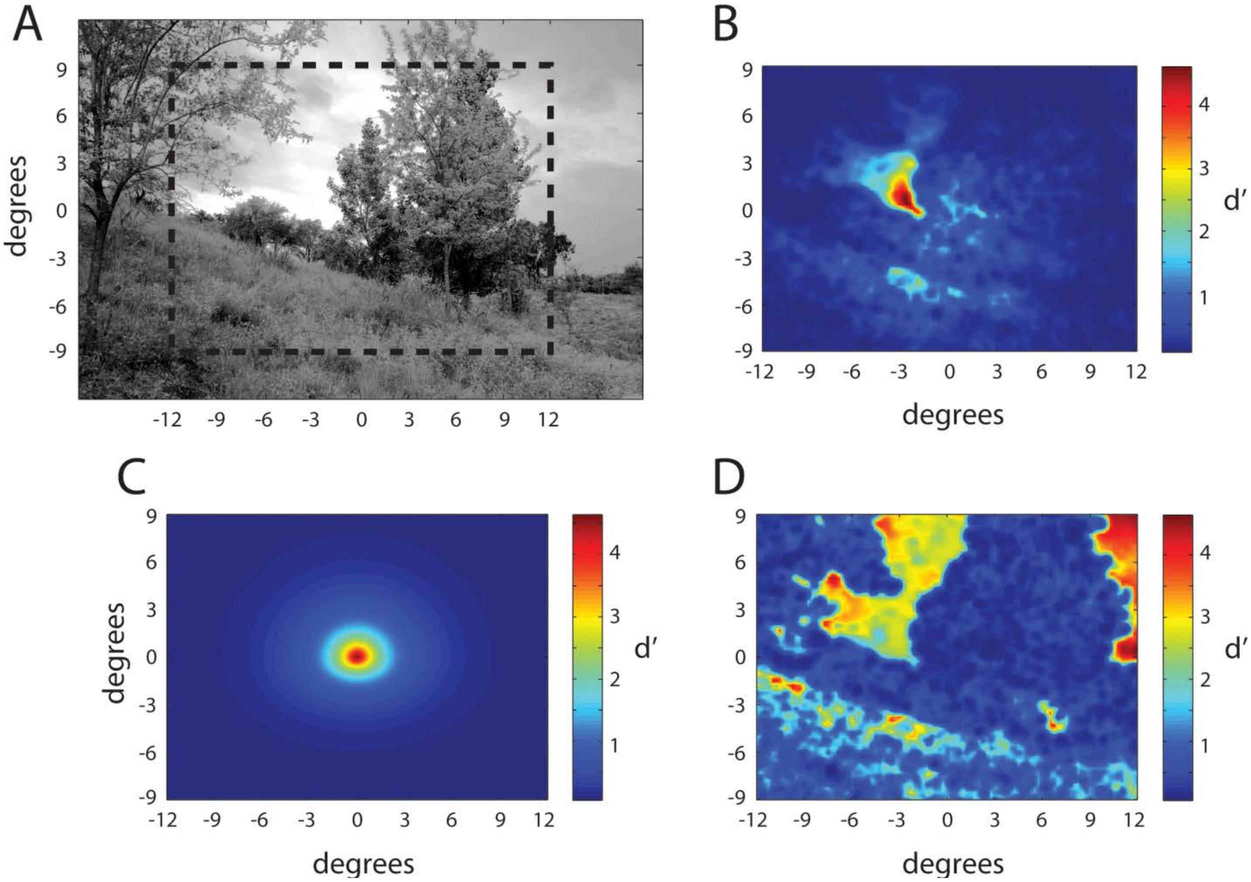


Figure 30. Detectability maps for a 4 cpd Gabor target. **A**. Gaussianized natural image. **B**. Detectability (d') of the target at all locations within the dashed box in A, given fixation at the center of the scene. **C**. Detectability of the target presented in the center of the scene, for all possible fixation locations within the dash box in A. **D**. Detectability of the target at all locations within the dash box in A, given fixation at the location of the target. The target contrast was fixed within each map, but was set so that d' reached a maximum of 4.5.

different types of detectability (d’) map. Figure 30B shows the d’ map for all possible locations of a fixed-contrast 4 cpd Gabor target, given fixation in the center of the image (0,0). As can be seen, d’ is predicted to be higher near the fixation point, but also to vary greatly depending on the background content at the target location. Such d’-maps for target location may be useful in making predictions for single fixation search, where the observer’s task is to detect the target during a single brief presentation when the location of the target is uncertain. For example, these *location d’ maps* could be used to determine the best possible search performance, assuming perfect parallel processing of all potential target locations. This is a critical baseline analysis for interpreting the results of visual search and attention experiments (e.g., see Eckstein 2011; Geisler & Cormack 2011). Figure 30C shows the d’ map for all possible fixation locations given that the target location is at the center of the image. In this case, d’ falls smoothly away from the target location. This *fixation d’ map* is closely related to the conspicuity area—the spatial region around a target where it can be detected in the background (Engel 1971; Bloomfield 1972; Geisler & Chou 1995; Toet et al. 1998). The conspicuity area can be defined as the area of the region where d’ exceeds some fixed criterion. Previous studies (Geisler & Chou 1995; Toet et al. 1998) have shown that there is strong negative correlation (on the order of -0.8 to -0.9) between the conspicuity area and the time it takes humans to locate the target, even in natural scenes (Toet et al. 1998). This is a powerful result of theoretical importance and of potential practical value. But to be of practical value one must know the conspicuity area for the particular target at its particular location in the background. The RV1 model might prove useful for estimating conspicuity areas without having to directly measure them in preliminary psychophysical experiments. Finally, Figure 30D shows the d’ map for all possible target locations when the observer is directly fixating the target. Such *foveal d’ maps* could be used to determine the best possible search accuracy for a given target in a given background, given unlimited search time.

EXTENSIONS OF THE MODEL

There are many ways of extending the RV1 model, either to new tasks or to a larger range of stimuli for the same task. However, for most of these extensions, it seems like the extension is not trivial and will require more careful thought. Therefore, this section is mostly exploratory in nature. Only for a few of the ideas presented have we done any preliminary research.

One of the possible extensions of the RV1 model is towards discrimination and identification tasks. Formally, there is no need to discriminate between discrimination and detection if one defines the “background” in a discrimination task as the (unmodified) target + background and the “target + background” as the (modified) target + background. The target is then simply the “target + background” minus the “background”. To give an example, suppose we have an orientation discrimination task and the observer has to decide whether a horizontal Gabor or a slightly off-horizontal Gabor appeared. Then the horizontal Gabor could be called the background and whatever needs to be added to the horizontal Gabor in order to produce the slightly off-horizontal Gabor would be the target. Seen from this point of view, the RV1 model can be used in its current form and tested on the large discrimination literature. To extend the RV1 model to identification, we simply use the RV1 model n times (once for each of the possible n targets, using naturally a different target-dependent filter for each target) and then do a maximum likelihood calculation to identify the target. Exactly what the effects of similarities of the possible targets are on the psychometric functions will have to be investigated.

One possible way of extending the RV1 model to a larger range of stimuli is to extend the model towards the detection of targets that occlude the background. In nature, it is almost always the case that an object will occlude all objects directly behind it (the exceptions are with transparent objects or organisms). The RV1 model is currently set up to predict detection of targets in a relatively unnatural situation, where the target is added to the background. To extend the RV1 model to target-occluding-the-background conditions, we cannot just use the trick described above for discrimination of targets and say: assume the background behind the target is uniform. If we did this, the model would fail because a target that occludes the background usually becomes far more visible than one added to the background. This is because the boundary of the target is likely to create edges, making the target more visible than if there were no occlusion. One option would be to make the target very small and estimate the masking effect of the background surrounding it (possibly over a much larger area the target envelope). Another option is to modify the task to a "similarity masking" task, where masking power depends on the similarity of the target to the background surrounding it. This would become more like a camouflage detection task. The initial strategy in both cases would be to estimate the masking power of the background using the same narrowband and broadband components as in the current version of the RV1 model, but change the inputs to the model (instead of the background behind the target, we select samples of background directly surrounding the target). Nevertheless, as stated before, a lot more thought and experimentation must be done to see how well the RV1 model can be extended to this more natural detection task of a target occluding the background immediately behind it.

Another possible extension of the RV1 model is more of a modification than an extension, and the idea came from a conceptual problem we had with the narrowband filter in the current version of the RV1 model. The narrowband filter is created by first passing the target through the optics and the receptive field center of a ganglion cell. After space-variant convolution with a log Gabor in frequency and a Gaussian in orientation, we obtain the target-dependent filter. This target-dependent filter is now applied to the result of passing a background through the optics and the receptive field center. In both cases, the target and background were passed through the optics and center mechanisms, but neither was passed through the surround. Conceptually, this seems a bit odd, or at least it seems harder to justify than creating the target-dependent filter from the full ganglion cell output to the target (which includes the surround), and then applying that filter to the full ganglion cell output to the background. If we did that, it could basically be interpreted as the cortex filtering the retinal ganglion cell output to the background based on what it thinks the retinal ganglion cell output to the target should look like. We tried this without altering anything else in the model (thus, this is only a very preliminary analysis) and two things happened. First, the predictions were slightly worse (after parameter optimization) than with the current version of the RV1 model: the RMS error on our yes-no experiment (for 3 targets, in noise and natural scenes, etc…) was 2.5 dB, which is greater than the 2.27 dB we have with the current version of the model. Second, the weight on the narrowband filter went to 1. That is, the optimization routine led to all weight being placed on the narrowband component and no weight on the broadband component (this weight on the narrowband component was varied in increments of 0.001 and it chose 1!). From this, the idea of possibly using a single cortical filter emerged. The cortical filter would be target-dependent, but more broadband than the narrowband component in our current version of the model. The question is where the neurophysiological justification would come from for such a change. In the present version, the orientation bandwidth (40 degrees) and frequency bandwidth (1.5 octaves) come from measurements on actual cortical neurons. However, it is worth noting that these are average values. Measures of the distribution of such values can be found in Geisler & Albrecht (1997). Geisler and Albrecht summarized the distributions of orientation and frequency bandwidths for cat and monkey cortical neurons with Gaussians. These bandwidth distributions determine the distribution of tuning curve shapes. For example, the Gaussian distribution of orientation bandwidths determines how often the tuning curve of cortical neuron will have an orientation bandwidth of 40 degrees or 35 degrees. The average shape of the tuning curve turns out not to be a Gaussian, but instead something that is better described as a weighted sum of a Gaussian and a Laplacian (exponential decay away from the mean in both directions). Such a distribution is more heavy-tailed than a Gaussian, and thus is more broadband in orientation than a Gaussian. This means that a more accurate use of the data in Geisler & Albrecht (1997) leads naturally to a narrowband filter that is more broadband than the one being used currently. Whether implementing a single cortical filter based on the measured distribution of tuning curve bandwidths will work well with using the full ganglion cell output to the target (that is, with the surround) as the basis for the target-dependent filter remains to be seen. We note that in this modified model, there would be one less free parameter because there is only a single cortical filter and the bandwidth distributions are specified by the neurophysiology.

A general strategy for extending the RV1 model to larger sets of stimuli is through adding different types of ganglion cells or cortical cells to the model. For example, we could add the M cells or the K cells and expand the set of stimuli this model accounts for. The seemingly easiest-to-add extension would be to explicitly include the ON-center and OFF-center P cell populations into the model. In our current version, we model a "combined-response" P cell that responds each time either as an ON-center or an OFF-center P cell. The modeled cell is a true combined-response P cell because the receptive field center of the ON-center P cells responds when the receptive field center of the OFF-center doesn't, and vice versa. The problem is that we must assume the distributions of ON-center and OFF-center P cells are identical. Their distributions are not identical (Dacey, 1993). It is estimated that there are 1.7 times more OFF-center P cells than ON-center P cells in the periphery; near the fovea their densities are the same. It is not difficult to create the mosaics for both the ON- and OFF-center P cells. The same algorithm that creates our combined-response P cell mosaic can be used to create both: we just use a different spacing function. However, potential problems may emerge once one actually tries to incorporate both mosaics into the model. For example, in equations 8, 9, and 11, we sample with the retinal ganglion cell mosaic (currently, that of the combined-response P cells). Now, we would have to sample separately with the ON-center and OFF-center mosaics. The problem is that the images we are sampling from in those three equations are for combined-response P cells, not ON-center or OFF-center P cells. This could be fixed by rectification (above and below the local mean luminance) before sampling, but the important point is that including two or more populations of cells is not necessarily a trivial extension.

With such potential difficulties in mind, we can list several other possible extensions of the RV1 model. One extension would be to color processing. Other than the problems outlined above, the method for incorporating color into the RV1 model seems straightforward. The P cells already carry color information. Some have a center-surround receptive field structure with a red center, while others have a green center. These naturally come in both ON and OFF varieties. The surround is not simply a green surround for a red center, or a red surround for a green center. Instead, the surround receives input from both red and green cones, and different distributions of the relative amounts of red vs. green input to these surrounds determines the degree of color opponency in the red/green sensitive P cells. The distributions of red-center and green-center ganglion cells can be modeled, and nothing in principle needs to be done on their outputs before feeding them into the RV1 model: the model will become naturally selective for color as a result. Other than the sampling issue described earlier, one problem with extending the model to color processing is that blue-yellow opponency information is carried by the K cells. This means we may have to include another population of ganglion cells if we want to incorporate color processing. Finally, what about an extension of the RV1 model that includes the M cells? M cells respond more transiently and better to high temporal frequency stimuli than do P cells. They are thought to carry much of the information from the rods during scotopic viewing conditions. Their inclusion might lead to having to model temporal properties of P cells or V1 neurons, but would also allow the potential addition of area MT (where cells are highly sensitive to directions of motion) to create a Retina-V1-MT model. Such a model would need to reduce to the Retina-V1 model when low temporal frequency stimuli are presented. This is likely one of the more difficult extensions, and should be attempted only after success with simpler problems (like adding ON-center and OFF-center P cells). Nevertheless, these examples do highlight the potential of the RV1 model for becoming a more general model of target detection.

APPLICATIONS OF THE MODEL

There are several potential applications of the RV1 model. One area where the RV1 model may be useful is in the diagnosis of vision problems. For example, if it is known that one group of observers has a particular vision problem while another group has normal vision, then one could use the RV1 model to make predictions about differences in detection performance between the two groups for many types of targets and backgrounds. The logic behind how this could help diagnose vision problems is as follows: if one can find a stimulus, or a small set of stimuli, where detection performance between people with normal vision and those with a particular vision disorder is vastly different, then one can use that stimulus as a diagnostic tool for determining whether a person has the disorder or not, and perhaps maximize the likelihood of early detection of the disorder. Currently, there is no way to find such a stimulus except through trial and error on actual human subjects. The RV1 model may be especially suited for this task because one can modify different components of the model and see what their effects on detection performance are. For example, one could put an extra optical filter in the model that represents a cataract, or one could simulate knocking out a subset of the retinal ganglion cells and see what effect these changes have on detection performance. A model that predicts detection performance well with such simulated impairments may be useful in finding good diagnostic stimuli for visual impairments. Given how well the RV1 model predicts data from our experiment (important here is that we tested across backgrounds and in the periphery of the visual field), there is a good chance the RV1 model will prove useful in this task. Theoretically, success in finding better diagnostic stimuli could also lead to improved classification of vision disorders.

Another area where the RV1 could prove to be useful is when an engineer wants to display something in a perceptually salient way, or when he wants to do the opposite and camouflage something. A simple example would be a typical head-up display (HUD) that one might see in military aircraft cockpits. The purpose of a HUD is so that the pilot does not need to keep looking down at instruments while also trying to focus on what is in front of him (possibly targeting another fighter in a dogfight). The ability to accurately predict what is visible and identifiable given a fixation point can potentially minimize the number of saccades, and thus the amount of time wasted on eye movements, that are needed to perform a task. Importantly, the background in a HUD display is often a complex natural scene, and the target is projected (added) onto the display, which matches in some ways the conditions that the RV1 model was developed for. Of course, factors not yet incorporated in the RV1 model will come into play. Information is often presented on HUD's that needs to be read and interpreted, not just detected. This means object recognition mechanisms not yet in the model, and not yet envisioned to be incorporated in the model, are probably very important (perhaps including effects such as crowding). The same is likely true for webpages, or for information communicated through other forms of media. For webpages, factors like how useful the design is, or how easily it can be navigated, are likely to be more important than pure detection probability at a fixation location. Nevertheless, because detection is necessary in all these tasks, it is hard to imagine the RV1 model not being useful in some sense towards their design.

At the other end of this spectrum is camouflage: the attempt to make something less visible in an environment. As described earlier in the "Extensions of the RV1 model" section, it may be possible to extend the RV1 model to "similarity masking" tasks where the properties of the surrounding background are compared to the target. As stated earlier, there are potential problems in predicting the detectability of a target that occludes the background because the boundaries of the target are often highly visible. But what if the boundaries are not very visible? In this case, the RV1 model in its current

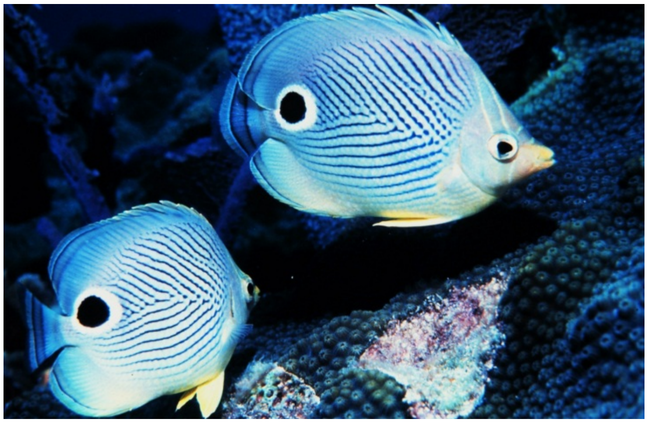


Figure 31. A type of camouflage the RV1 will fail to predict. The feature is highly detectable, but it confuses or distracts (e.g., tail marking looks like the eye of a large fish).

form can already be tested in a camouflage task. That is, we can test how well does the RV1 model does in predicting the least detectable patterns in a particular background. If it can do even reasonably well at this task, it would be an improvement over the status quo. There is a lot of research on camouflage (for the military, if for no other reason), but for the most part it is a trial and error endeavor based on inspiration from nature, ideas artists have had, or just intuitions people have about what might work best. This doesn't mean that effective camouflage doesn't exist. It just means that finding good camouflage for a novel environment, or finding a better camouflage for a known environment, is a resource intensive operation. There are clearly some types of camouflage that are effective, but where the RV1 model will incorrectly classify as ineffective. One such example is shown in Figure 31. The RV1 model will correctly determine that the camouflage (fake eye) is highly detectable, but the model has no mechanisms to determine whether a highly detectable pattern is distracting or confusing (the fake eye might confuse a predator, but other highly detectable features might not). Thus, the RV1 model cannot go farther than to say how detectable the feature is, and will thus fail at classifying the fake eye as effective camouflage. However, for types of camouflage that blend in with the environment, (as long as object recognition is not required) there is a decent chance that the RV1 model will make accurate predictions. Perhaps the area of greatest potential utility with respect to camouflage is where one has to design camouflage that is most effective across multiple background patterns. If one needs camouflage for just a single type of background pattern, an artist is likely to do a very good job. But what works on average best across multiple backgrounds is harder to intuit.

Another way in which the RV1 model could be useful is in reducing interpretation errors when clinicians look at medical images (Berlin, 2005; 2007). Percentage estimates in radiology are around 30% for both miss rates and false positives. Many of these errors are due to visual errors – errors that occur because the clinician either does an incomplete search or because the clinician failed to recognize an abnormality even though he fixated at its location (Giger, 1988). Visual errors have been estimated to comprise about 55% of all interpretation errors in interpreting medical images. Of visual errors, approximately 65% were due to the clinician failing to look near the region of the abnormality (Kundel, 1975; 1978), while 35% were due to the failure to detect the abnormality despite fixating at it (Carmody, 1980). The explanation given for this 35% is that fixation time was insufficient to properly interpret the features as an abnormality. Masking of lesions by normal anatomical structures is also estimated to increase lesion detection thresholds by an order of magnitude (Samei, 1997). In this context, it is understandable why computer aided diagnosis (CAD) – an algorithm automatically searches for features likely to be part of an abnormality and then notifies a clinician to take a closer look – is becoming more popular (Doi, 2007). Despite its increasing importance, it is not yet precisely clear how best to integrate CAD with diagnosis by humans. Machines still make many errors, cannot yet replace humans in this task, and many clinicians ignore CAD when it starts making too many errors (either false positives or misses). This is one area where any detection model that can predict human performance better could be useful. For example, one could better evaluate existing CAD by predicting how well they predict the difficult-to-detect abnormalities, which are really the ones CAD should be useful for. One of the technical details that would need to be fleshed out is what the target is. It's possible that for something like tumor detection one would have to use a signal-known-probabilistically paradigm instead of the signal-known-exactly case used for the detection tasks the model was tested on. That might mean the target is either a weighted average of many targets, or the model looks for one of many different targets as it would in an identification task.

Finally, a very important potential application of the RV1 model is for creating better models of visual search. The human visual system combines high-speed, ballistic eye movements called saccades with a foveated retina as a solution to the demand of providing high spatial resolution across a wide field of view given limited resources (one example of such a resource constraint is size of the optic nerve. If the retina had foveal acuity, the optic nerve would not fit in the space provided). Modeling overt visual search (search using eye movements) is a tough problem, even if one restricts the task to simple detection of a target in a larger background where eye movements are necessary to find the target. To properly model visual search, one needs to know how information obtained during one fixation (or information still retained from all previous fixations) is used to decide the next fixation location. In order to do this, one needs an accurate detectability map for each fixation made. These are precisely the types of d' maps generated from the RV1 model shown in Figure 30b. Najemnik and Geisler (2005) showed how such d' maps could be useful for models of visual search. They derived the optimal eye movement strategy (the Bayesian ideal observer) for a visual search task in 1/f noise background. Importantly, this derivation first requires knowing a d' map. In their case, the d' map was the same for all locations in the background because the background was statistically speaking homogeneous. However, in the more general case of search in natural scenes, the d' map would have to be separately estimated for each fixation location. This is because natural scenes are statistically speaking inhomogeneous from region to region (the sky has very different statistical properties than the ground or a tree). Generating such d' maps is precisely where the RV1 model has utility. One potential problem with using the RV1 model to generate d' maps for every possible fixation location is that doing so becomes computationally intractable. Even generating a single d' map can take a good deal of time (the RV1 model is practical when it just needs to generate a single predicted d' or a smaller number of d's, but it cannot quickly generate an entire d' map). However, several studies have shown that one does not need to have a complete d' map to predict many aspects of visual search performance: the concept of a conspicuity area will do. A conspicuity area is simply the region over which d' exceeds some fixed criterion (Engel 1971; Bloomfield 1972; Geisler & Chou 1995; Toet et al. 1998). Studies have shown that conspicuity area predicts the time it takes humans to find a target, even for something like a vehicle hidden in a natural scene (Toet et al. 1998). This means that the RV1 model may have potential value even by predicting an iso-d' contour, or the conspicuity area. This is far more computationally tractable than entire d' maps. In summary, there are many possible applications of the RV1 model that should be explored; these range from improved medical diagnosis to camouflage to visual search.

Chapter 6: Concluding remarks

In this paper, the Retina-V1 model was introduced. It is a model of target detection designed to predict the detectability of any known localized spatial pattern (a target) in any arbitrary background at any point in the visual field. The model is practical, needing no appreciable time to generate a predicted detection threshold, and it is testable across a wide range of conditions. As has been described, there are good opportunities to extend the model to even more general conditions and also good opportunities for it to produce practical benefits. In its current form, it is restricted to what may seem to be a highly restrictive set of conditions: photopic conditions, grayscale images, static images, no eye movements allowed, nothing that requires object recognition (there are no grouping mechanisms in the model), etc… Nevertheless, it is far more general in its scope than competing models. Its primary areas of generality are: 1) the ability to predict detectability in natural scenes, and 2) the ability to predict detectability across the visual field. The RV1 model was fit to two datasets. For uniform background conditions, we fit the model to the ModelFest dataset. There are 43 ModelFest stimuli, and the RV1 model predicts the data with an RMS error comparable to competing models. However, unlike competing models, the RV1 model is grounded in the known optics, physiology and anatomy of the eye. Only the final stage where the retinal ganglion cell outputs are pooled to predict d' is the model based on ideal observer analysis. The optimal parameters for fitting the ModelFest dataset were fixed and used unaltered (except for one observer-dependent free parameter) to predict Foley's dataset, which consists of different types of Gabors presented at different retinal eccentricities (ModelFest is for foveal detection only), in uniform backgrounds. The RV1 model performed approximately as well on Foley's dataset as on the ModelFest dataset. The masking components of the RV1 model were tested by fitting the model to data from our own experiment. Our experiment tested detectability of three ModelFest targets (one of the Gabors, one of the Gaussians and the Edge) in different types of backgrounds (uniform, 1/f noise, and Gaussianized natural scenes), at 4 different retinal eccentricities. We found that threshold contrast power increases as a linear function of background contrast power, not just for 1/f noise (previously known), but also for Gaussianized natural scenes. The slope of this linear function was found to increase as a function of retinal eccentricity. In general, thresholds in 1/f noise and natural backgrounds were similar, though thresholds were somewhat greater in natural backgrounds. The RV1 model does a decent job of predicting the data. Of all the conditions tested, the only ones where the model did not do well was for Gabors presented at 10 degrees eccentricity in all types of backgrounds; the predictions are pretty good up to 5 degrees. It remains to be seen how well the model does at predicting detectability for many other stimuli presented in various conditions. In conclusion, the purpose of this study was to build a foundation for a more general model of one small but important area of vision science: target detection. Given the relative success of the model on the data tested, and given the relatively clear potential for extending the model to include a much wider range of stimuli, one can argue that this project was a success.

Appendix

GENERATION OF THE GANGLION CELL MOSAIC

First, place a ganglion cell at the center of the fovea. The algorithm then creates successive rings of ganglion cells around that location. Begin by defining $g_{k,n}$ to be the k^{th} ganglion cell on ring n. The ganglion cell at the fovea will be $g_{1,1}$ (the only ganglion cell on ring 1). Also, define $C\left(g_{k,n}\right)$ to be the circle centered at $g_{k,n}$ with a radius (spacing) specified by the equation in Figure 2C (the radius depends on the retinal location of $g_{k,n}$). Two rules specify how all ganglion cells on ring n are created given that ring n-1 has been completed. For each rule, there is a special case when the fovea is the previously created ring.

Rule 1: The first ganglion cell on ring n, $g_{1,n}$, is placed at the intersection furthest from the fovea between $C\left(g_{k-1,n-1}\right)$ and $C\left(g_{k,n-1}\right)$, where k is a randomly chosen positive integer at most as large as the total number of ganglion cells on ring n-1. The special case where n-1=1 (the central ganglion cell) is handled by placing $g_{1,2}$ at any randomly chosen point on $C\left(g_{1,1}\right)$.

Rule 2: The *k*th ganglion cell on ring n, for k>1, is found by first identifying all intersections between $C\left(g_{k-1,n}\right)$ and the circles of all ganglion cells on ring n-1. In the special case where n-1=1, there will be only two such intersections, one clockwise and the other counter-clockwise from $g_{k-1,n}$. In this case, choose the intersection that is clockwise from $g_{k-1,n}$ as the location for $g_{k,n}$. In the more general case where n-1>1, we first find the subset of intersections between $C\left(g_{k-1,n}\right)$ and the circles of all ganglion cells on ring n-1 that lie counter-clockwise from $g_{k-1,n}$. The location of $g_{k,n}$ is at the intersection (within this subset) that is furthest from the fovea.

GANGLION CELL RECEPTIVE FIELDS

The center and surround mechanisms at retinal location $\mathbf{x}$ are given by:

$$g_c(\mathbf{y};\mathbf{x}) = \frac{1}{2\pi\sigma_c^2(\mathbf{x})}\exp\left(-0.5\frac{\|\mathbf{y}-\mathbf{x}\|^2}{\sigma_c^2(\mathbf{x})}\right) \tag{22}$$

$$g_s(\mathbf{y};\mathbf{x}) = \frac{1}{2\pi\sigma_s^2(\mathbf{x})}\exp\left(-0.5\frac{\|\mathbf{y}-\mathbf{x}\|^2}{\sigma_s^2(\mathbf{x})}\right) \tag{23}$$

where $\|\cdot\|$ is the Euclidean norm (vector length).

CORTICAL FILTERS

The filtering characteristics of cortical neurons are assumed to be described by the product of a log Gabor function in spatial frequency (Gaussian on a log spatial frequency axis) and a Von Mises function in orientation, where the log Gabor has a spatial frequency bandwidth (at half height) of 1.5 octaves and the Von Mises function has an orientation bandwidth (at half height) of 40 deg. The form of the functions is as follows:

$$G_{\log}(u; f,\sigma) = \exp\left[-0.5\frac{(\log u - \log f)^2}{\sigma^2}\right] \tag{24}$$

$$V(\theta;\sigma_V) = \exp\left[k\cos(\theta-\theta_0)\right] \tag{25}$$

TARGET ENVELOPE

The envelope of the target was computed by first finding the parameters of a scaled two-dimensional Gaussian function $g(\mathbf{y};\mathbf{u},\mathbf{\Sigma})$ that best fits the absolute value of the target:

$$\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{k} = \underset{\boldsymbol{\mu},\boldsymbol{\Sigma},k}{\arg\min} \sum \left[\left| T\left(\mathbf{y}\right) \right| - kg\left(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}\right) \right]^2 \qquad (26)$$

where $\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\Sigma}$ is the covariance matrix, and $g\left(\mathbf{y};\mathbf{u},\boldsymbol{\Sigma}\right)$ has a volume of 1.0. To obtain the envelope for a given retinal location, we then blurred this Gaussian with another Gaussian (of volume 1.0) having the size of the ganglion cell center at that retinal location $\sigma_c\left(\mathbf{x}\right)$. Thus the envelope (which also has a volume of 1.0) is given by

$$E_T\left(\mathbf{y};\mathbf{x}\right) = g\left(\mathbf{y};\hat{\mathbf{u}},\hat{\boldsymbol{\Sigma}} + \begin{bmatrix} \sigma_c\left(\mathbf{x}\right) & 0 \\ 0 & \sigma_c\left(\mathbf{x}\right) \end{bmatrix} \right) \qquad (27)$$

References

Ahnelt, P. K., Kolb, H., & Pflug, R. (1987). Identification of a subtype of cone photoreceptor, likely to be blue sensitive, in the human retina, *J. Comp. Neurol*., 255, 18-34

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

Andriessen, J. J., & Bouma, H. (1976). Eccentric vision: Adverse interactions between line segments. *Vision Research*, 16(1), 71-78.

Arnow, T. L., & Geisler, W. S. (1996). Visual detection following retinal damage: Predictions of an inhomgeneous retino-cortical model. *SPIE Proceedings: Human Vision and Electronic Imaging, 2674*, 119-130.

Ball, K., & Sekuler, R. (1982). A specific and enduring improvement in visual motion discrimination. *Science*, 218, 697-698.

Barlow, H. B., FizHugh, R., & Kuffler, S. W. (1957). Dark adaptation absolute threshold and Purkinje shift in single units of the cat's retina. *J. Physiol*., 137, 327-337.

Berlin, L. (2005). Errors of omission. *AJR*, 185, 1416-1421

Berlin, L. (2007). Accuracy of diagnostic procedures: has it improved over the past five decades? *AJR*, 188, 1173-1178

Bloomfield, J.R., (1972). Visual search in complex fields: size differences between target disc and surrounding discs. *Human Factors* 14, 139–148.

Bonds, A. B., (1989). Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. *Vis. Neurosci*., 2, 45-55.

Burgess, A.E. (2011). Visual perception studies and observer models in medical imaging. *Seminars in Nuclear Medicine*, 41, 419-436.

Burgess, A.E. & Colborne B. (1988). Visual signal detection: IV. Observer inconsistency. *Journal of the Optical Society of America A*, 2, 617-627.

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

Burton, G. J., & Moorehead, I. R. (1987). Color and spatial structure in natural scenes. *Applied Optics, 26*, 157-170.

Campbell, F. W., Cooper, G. F., & Enroth Cugell, C. (1969). The spatial selectivity of the visual cells of the cat. *J. Physiol. (Lond.)*, 203, 223-235.

Campbell, F. W., Kulikowski, J. J., & Levinson, J. Z. (1966). The effect of orientation on the visual resolution of gratings. *Journal of Physiology*, 187, 427-436.

Campbell, F. W., & Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. *J. Physiol*., 197, 551-566.

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

Carandini, M., Heeger, D. J., & Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience, 17,* 8621–8644.

Carmody, D. P., Nodine, C. F., Kundel, H. L. (1980). An analysis of perceptual and cognitive factors in radiographic interpretation. *Perception*, 9, 339-344.

Croner, L. J., & Kaplan, E. (1995). Receptive fields of P and M ganglion cells across the primate retina. *Vision Research, 35*(1), 7-24.

Curcio, C. A., & Allen, K. A. (1990). Topography of ganglion cells in human retina. *The Journal of Comparative Neurology, 300*, 5-25.

Curcio, C. A., Allen, K. A., Sloan, K. R., Lerea, C. L., Hurley, J. B., Klock, I. B., & Milam, A. H., (1991). Distribution and morphology of human cone photoreceptors stained with anti-blue opsin. *J. Comp. Neurol*., 312, 610-624.

Dacey, D. M. (1993). The mosaic of midget ganglion cells in the human retina. *Journal of Neuroscience, 13,* 5334-5355.

Dacey, D. M. (2004). Origins of perception: retinal ganglion cell diversity and the creation of parallel visual pathways. In *The Cognitive Neurosciences*, ed. MS Gazzaniga, 281-301. Cambridge, MA: MIT Press.

Dacey, D. M., & Peterson, M. R. (1992). Dendritic field size and morphology of midget and parasol ganglion cells of the human retina. *Proceedings of the National Academy of Sciences,* 89, 9666-9670.

De Monasterio, F. M., & Gouras, P. (1975). Functional properties of ganglion cells of the rhesus monkey retina. *J. Physiol. (Lond.),* 251, 167-195.

Derrington, A. M., & Lennie, P. (1984). Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. *Journal of Physiology, 357,* 219-240.

De Valois, R. L., Albrecht, D. G., & Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research,* 22, 545-559.

De Valois, R. L., & De Valois, K. K. D. (1988). *Spatial vision*. Oxford: Oxford University Press.

De Valois, R. L., Yund, E. W., & Hepler, N. (1982). The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, 22, 531-544.

Doi, K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput Med Imag & Graphics*, 31, 198-211.

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,* 2901-2911.

Eckstein, M. (2011). Visual search: A retrospective. *Journal of Vision, 11*(5), 1-36.

Eckstein, M. P., Ahumada, A. J., & Watson, A. B. (1997a). Visual signal detection in structured backgrounds. II. Effects of contrast gain control, background variations, and white noise. *Journal of the Optical Society of America A,* 14, 2406-2419.

Eckstein, M. P., Ahumada, Jr., A. J., & Watson, A. B. (1997b). Image discrimination models predict signal detection in natural medical image backgrounds. In *Human Vision, Visual Processing, and Digital Display VIII, ed. B.E. Rogowitz and T.N. Pappas, Proc. Vol. 3016,* pp. 44-56, SPIE, Bellingham, WA

Engel, F. (1971). Visual conspicuity, directed attention and retinal locus. *Vision Research*, 11, 563–576.

Field, D. J. (1987). Relations between the statistics of natural images and the repsonse properties of cortical cells. *Journal of the Optical Society of America A,* 4, 2379-2394.

Field, D. J., Hayes, A., & Hess, F. (1993). Contour integration by the human visual system: evidence for a local "association field". *Vis. Res*., 33, 173-193.

Field, G. D., & Chichilnisky, E. J. (2007). Information processing in the primate retina: Circuitry and Coding. *Annual Review of Neuroscience,* 30, 1-30.

Foley, J. M. (1994). Human luminance pattern-vision mechanisms: Masking experiments require a new model. *Journal of the Optical Society of America A, 11*(6), 1710-1719.

Foley, J. M., Varadharajan, S., Koh, C. C., & Farias, M. C. (2007). Detection of Gabor patterns of different sizes, shapes, phases and eccentricities. *Vision Research*, 47(1), 85–107.

Geisler W.S. & Albrecht D.G. (1997). Visual cortex neurons in monkeys and cats: Detection, discrimination and identification**,** *Visual Neuroscience*, 14, 897-919.

Geisler, W. S. & Chou, K. (1995). Separation of low-level and high-level factors in complex tasks: Visual search. *Psychological Review*, 102(2), 356-1378.

Geisler, W.S. & Cormack, L. (2011). Models of overt attention. In S.P. Liversedge, I.D, Gilchrist & S. Everling (Ed.) *Oxford Handbook of Eye Movements*. New York: Oxford University Press.

Giger, M.S., Doi, K., MacMahon, H. (1988). Image feature analysis and computer-aided diagnosis in digital radiography. 3. Automated detection of nodules in peripheral lung fields. *Med Phys*, 15, 158-166

Goris, R.L.T., Putzes, T., Wagemans, J., & Wichmann F.A. (2013). A neural population model for visual pattern detection. *Psychological Review*, 120 (3), 472–496.

Graham, N.V. (1977). Visual detection of aperiodic spatial stimuli by probability summation among narrowband channels. *Vision Research*, 17, 637-652.

Graham N.V. (1989). *Visual Pattern Analyzers*. New York: Oxford.

Graham N.V. (2011). Beyond multiple pattern analyzers as linear filters (as classical V1 simple cells): Useful additions of the last 25 years. *Vision Research*, 51, 1397-1430.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Hartline, H. K. (1938). The response of single optic nerve fibres of the vertebrate eye to illumination of the retina. *Am. J. Physiol.*, 121, 400-415.

Hartline, H. K. (1940). The receptive fields of optic nerve fibres. *Am. J. Physiol.*, 130, 690-699.

Heeger, D. J. (1991) Nonlinear model of neural responses in cat visual cortex. In M. S. Landy & J. A. Movshon (Eds.), *Computational Models of Visual Perception* (pp. 119-133). Cambridge: The MIT press.

Heeger, D. J. (1992) Normalization of cell responses in cat striate cortex. *Visual Neuroscience,* 9, 191-197.

Hood, D. C. (1997) Lower-level visual processing and models of light adaptation. *Annual Review of Psychology*, 1998, 49, 503-535.

Hood, D. C., & Finkelstein, M. A. (1986) Sensitivity to light. In K. R. Boff, L. Kaufman & J. P. Thomas (Eds.), *Handbook of Perception and Human Performance* (Vol. 1). New York: John Wiley and Sons.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurons in the cat's striate cortex. *J. Physiol*., 148, 574-591.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol*., 160, 106-154.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *J. Neurophysiol.*, 16, 37-68.

Kundel, H. L. (1975). Peripheral vision, structured noise and film reader error. *Radiol*, 114, 269-273.

Kundel, H. L., Nodine, C. F., Carmody, D. (1978). Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Invest. Radiol,* 13, 175-181.

Lamme, V. A. F. (1995). The neurophysiology of figure-ground segregation in primary visual cortex. *J. Neurosci*., 15, 1605-1615.

Legge, G. E., & Foley, J. M. (1980). Contrast masking of human vision. *J. Opt. Soc. Am*., 70, 1458-1471.

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

Levi, D. M., & Klein, S. A. (2002). Classification images for detection and position discrimination in the fovea and parafovea. *Journal of Vision*, 2(1):4, 46-65.

Lu, Z.L. &, Dosher, BA. (1999) Characterizing human perceptual inefficiencies with equivalent internal noise. *Journal of the Optical Society of America A*, 16, 764-778.

Lu, Z.L. &, Dosher, BA. (2008) Characterizing observers using external noise and observer models: Assessing internal representations with external noise. *Psychological Review*, 115, 44-82.

McMahon, M. J., & MacLeod, D. I. (2003). The origin of the oblique effect examined with pattern adaptation and masking. *Journal of Vision*, 3(3), 230-239.

Merigan, W. H., Katz, L. M., & Maunsell, J. H. R. (1991). The effects of parvocellular lateral geniculate lesions on the acuity and contrast sensitivity of macaque monkeys. *The Journal of Neuroscience, 11*, 994-1001.

Merigan, W.H. & Maunsell, J.H.R. (1993) How parallel are the primate visual pathways? *Annual Review of Neuroscience*, 16, 369-402.

Morrone, M. C., & Burr, D. C. (1988) Feature detection in human vision - a phase-dependent energy model. *Proceedings of the Royal Society of London Series B - Biological Sciences*, 235, 221–245.

Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978). Spatial and temporal contrast sensitivity of neurons in areas 17 and 18 of the cat's visual cortex. *J. Physiol*., 283, 101-20

Myers K.J. & H. H. Barrett, H.H. (1987) Addition of a channel mechanism to the ideal-observer model, *J. Opt. Soc. Am. A* **4**, 2447–2457.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*, 387-391.

Navarro, R., Artal, P., & Williams, D. R. (1993). Modulation transfer function of the human eye as a function of retinal eccentricity. *Journal of the Optical Society of America A, 10,* 201-212.

Palmer, L. A., Jones, J. P., & Stepnoski, R. A. (1991). Striate receptive fields as linear filters: Characterization in two dimensions of space. In A. G. Leventhal (Ed.), *The Neural Basis of Visual Function* (pp. 246-265). Boston: CRC Press.

Peli E, Yang J, Goldstein RB (1991). Image invariance with changes in size: The role of peripheral contrast thresholds. *J Opt Soc Am A*, **8**, 1762–1774.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision*, 4(12):12, 1136-1169.

Phillips, G. C., & Wilson, H. R. (1984). Orientation bandwidths of spatial mechanisms measured by masking. *J. Opt. Soc. Am.*, 1(2), 226-232.

Polat, U., & Sagi, D. (1993). Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments. *Vis. Res.*, 33, 993-999.

Polat, U., & Sagi, D. (1994). The architecture of perceptual spatial interactions. *Vis. Res.*, 34, 73-78.

Quick, R. (1974) A vector-magnitude model of contrast detection. *Biological Cybernetics*, 16, 65-67.

Reid, C. R., & Alonso, J. M. (1995). Specificity of monosynaptic connections from thalamus to visual cortex. *Nature*, 378 (6554), 281-284.

Rodieck, R. W. (1965). Quantitative analysis of cat retinal ganglion cell response to visual stimuli. *Vision Research, 5,* 583-601.

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, 3225-3235.

Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology*, 88(1), 455–463.

Roorda, A., & Williams, D. R., (1999). The arrangement of the three cone classes in the living human eye. *Nature*, 397, 520-522.

Saarinen, J., & Levi, D. M. (1995). Perceptual learning in vernier acuity: What is learned? *Vision Research*, 35, 519-527.

Samei, E. (Ed.). (2010). *The handbook of medical image perception and techniques*. Cambridge University Press.

Samei, E., Flynn, M. J., Kearfott, K. J. (1997). Patient dose and detectability of subtle lung nodules in digital chest radiographs, *Health Physics*, 72(6S).

Shapley, R. M., & Lennie, P. (1985) Spatial frequency analysis in the visual system. *Annual Review of Neuroscience, 8,* 547-583.

Sit Y.F., Chen Y., Geisler W.S., Miikkulainen R., & Seidemann E. (2009). Complex dynamics of V1 population responses explained by a simple gain-control model. *Neuron*, 64, 943-956.

Toet A, Kooi FL, Bijl P & Valeton JM (1998). Visual conspicuity determines human target acquisition performance. *Optical Engineering, 37,* 1969-1975.

Watson, A.B. (1979) Probability summation over time *Vision Research*, 19, 515-522.

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

Watson, A. B., & Solomon, J. A. (1997). Model of visual contrast gain control and pattern masking. *Journal of the Optical Society of America A, 14*, 2379-2391.

Watt, R. J., & Morgan, M. J. (1985). A theory of the primitive spatial code in human vision. *Vision Research*, 25, 1661–1674.

Wiesel, T. N., & Hubel, D. H. (1966). Spatial and chromatic interactions in the lateral geniculate body of the rhesus monkey. *J. Neurophysiol*., 29, 1115-1156.

Williams, D. R., Hofer, H. (2003). *The Visual Neurosciences, Vol. 1*, pp 795-810, MIT press.

Williams, D. R., MacLeod, D. I. A., and Hayhoe, M. M., (1981). Punctuate sensitivity of the blue sensitive mechanism. *Vision Research*, 21, 1357-1376.

Wilson, H. R., & Bergen, J. R. (1979). A four mechanism model for threshold spatial vision. *Vision Research, 19*, 19-32.

Zhang Y.L., Pham B.T. & Eckstein M.P. (2006). *IEEE Trans Med Imaging,* 25(10), 1348-62.

Zipser, K., Lamme, V. A. F., & Schiller, P. H. (1996). Contextual modulation in primary visual cortex. *J. Neurosci.*, 16, 7376-7389.