Detection of Occluding Targets in Natural Backgrounds

R. Calen Walshe & Wilson S. Geisler
University of Texas at Austin

Abstract

Detection of target objects in the surrounding environment is a common visual task. There is a vast psychophysical and modeling literature concerning the detection of targets in artificial and natural backgrounds. Most studies involve detection of additive targets or of some form of image distortion. While much has been learned from these studies, the targets that most often occur under natural conditions are neither additive nor distorting; rather, they are opaque targets that occlude the backgrounds behind them. Here we describe our efforts to measure and model detection of occluding targets in natural backgrounds. To systematically vary the properties of the backgrounds, we used the constrained sampling approach of Sebastian et al. (2017). Specifically, millions of calibrated gray-scale natural-image patches were sorted into a 3D histogram along the dimensions of luminance, contrast, and phase-invariant similarity to the target. Eccentricity psychometric functions (accuracy as a function of retinal eccentricity) were measured for 4 different occluding targets and 15 different combinations of background luminance, contrast and similarity, with a different randomly-sampled background on each trial. The complex pattern of results was consistent across the three subjects, and was largely explained by a principled model observer (with only a single efficiency parameter) that combines three image cues (pattern, silhouette, edge) and four well-known properties of the human visual system (optical blur, blurring and downsampling by the ganglion cells, divisive normalization, intrinsic position uncertainty). The model also explains the thresholds for additive foveal targets in natural backgrounds reported in Sebastian et al. (2017).

Introduction

Natural selection pushes perceptual mechanisms to match the natural tasks an organism performs in the environments where it evolved, and thus the computational and experimental study of natural tasks and stimuli is important for developing and testing principled hypotheses (e.g., see Geisler 2008). A fundamental and ubiquitous natural task is identifying the presence or absence of specific target objects in visual scenes. Most studies of identification performance have been directed at the case where targets are added to non-natural backgrounds, including uniform backgrounds (König & Brodhun 1889; Mueller 1951; Hood 1998), grating backgrounds (Stromeyer & Julesz 1974; Legge & Foley 1981; Wilson et al. 1983), and noise backgrounds (Burgess et al. 1981). More recently there have been studies directed at detection of additive targets in natural backgrounds (Caelli & Moragila 1986; Rohaly et al. 1997; Bex et al. 2009; Alam et al. 2014; Bradley et al. 2014; Sebastian et al. 2017; Sebastian et al. 2020). There have also been recent studies directed at the related task of detecting specific kinds of distortions in natural images (Nadenau et al. 2002; Bex 2010; Freeman & Simoncelli 2011).

Identification tasks with additive targets (and with distortion targets) have the practical advantage that it is relatively easy to measure detection thresholds for any target, anywhere in the visual field, by varying the target's amplitude (or the level of distortion). Another specific advantage of additive targets is that it is relatively easy to develop formal models, including ideal-observer models (for reviews see Geisler 2011; Burgess 2018), which provide principled hypotheses for neural computations and an appropriate benchmark against which to compare the organism's behavioral or neural performance.

Although much has been learned from studies with additive (and distortion) targets , they are relatively rare under natural conditions. Most real-world targets are composed of opaque surfaces that occlude the background rather than add to the background. Detection of occluding targets is fundamentally different from additive targets because (i) occluding targets almost always create a sharp boundary with the background and this boundary is an important part of the signal, (ii) most occluding targets are trivially detectable in the center of the fovea (unless they are extremely small) and only become difficult to detect in the periphery, and (iii) there are currently no well-developed ideal-observer or other principled models for the detection of occluding targets. It should be noted that non-overlapping opaque elements (e.g., letters) on a uniform background (as in most search, memory, and crowding studies) are mathematically additive. It is only when the target occludes background features (or is partially occluded by background features) that additivity is violated.

There have been very few studies of detection of occluding targets in natural backgrounds, although Wallis & Bex (2012) studied detection of occluding random textures (see Discussion). Here we describe a principled observer for detection of specific (known) occluding targets in gray-scale natural backgrounds, and compare its performance to that of human observers. Our approach builds upon the "constrained-sampling" approach recently described in Sebastian et al. (2017). The logic of the approach is to identify, from the existing literature, multiple stimulus dimensions known to have a major effect on task performance, and then bin natural stimuli jointly along those dimensions. The effect of those dimensions on task performance can then be determined by measuring performance for natural stimuli randomly selected from a sparse subset of the bins covering the space of natural stimuli.

Here, as in Sebastian et al., three dimensions of natural backgrounds are considered: the luminance (L), contrast (C), and the spatial similarity (defined later) to the target (S). Natural target objects often have sharp boundaries and contain one or a few dominant orientations; therefore, we measured detection performance in natural backgrounds for four targets—vertical edge, horizontal edge, bowtie (oriented to have only horizontal and vertical edges), and spot (center-surround)—all having the same mean luminance (see Figure 1).

For each target, eccentricity psychometric functions were measured for natural backgrounds that varied along each of the cardinal dimensions (L, C and S), while the other two dimensions were held at their median values. An eccentricity psychometric function describes accuracy (hits and false alarms) as function of the retinal eccentricity of the target. We chose to measure eccentricity psychometric functions for two reasons. First, for occluding targets it is often not possible to measure thresholds by varying target luminance or contrast, because performance is always well above chance. Second, when looking for a target under natural conditions, its

detectability only varies when the fixation location is varied. In other words, eccentricity psychometric functions are an appropriate measure of detectability under natural conditions.

To interpret the results, we propose and evaluate a principled model observer. This model observer uses three cues that together are likely to capture most of the available information for detecting occluding targets: the response of a template matched to the target pattern, the response of a template matched to the silhouette of the target, and the response of an edge-energy measure of the target boundary. The model observer also includes four factors known to limit human detection performance: the optics of the eye, retinal ganglion-cell sampling, divisive normalization, and intrinsic position uncertainty. These factors were taken directly from existing optical, anatomical, neurophysiological and psychophysical measurements. We find that this model observer predicts most of the variance in the human data, with a single free parameter: an overall efficiency scalar.

Methods

Natural Background Statistics: LCS Space

The natural background statistics were computed from 1,204 high-resolution calibrated natural images taken around the Austin area. The image set excluded human-made structures such as buildings and roads. Each image (4,284 x 2,844 pixels) was taken with a calibrated camera that was linear in luminance and had 14-bit resolution per color channel. The calibration procedure and natural-image database are available online at natural-scenes.cps.utexas.edu. The RGB images were converted to grayscale by transforming the images to XYZ space and storing only the Y (luminance) channel. The top 1% of pixels were clipped to a maximum value, and then all pixels were normalized by that maximum value.

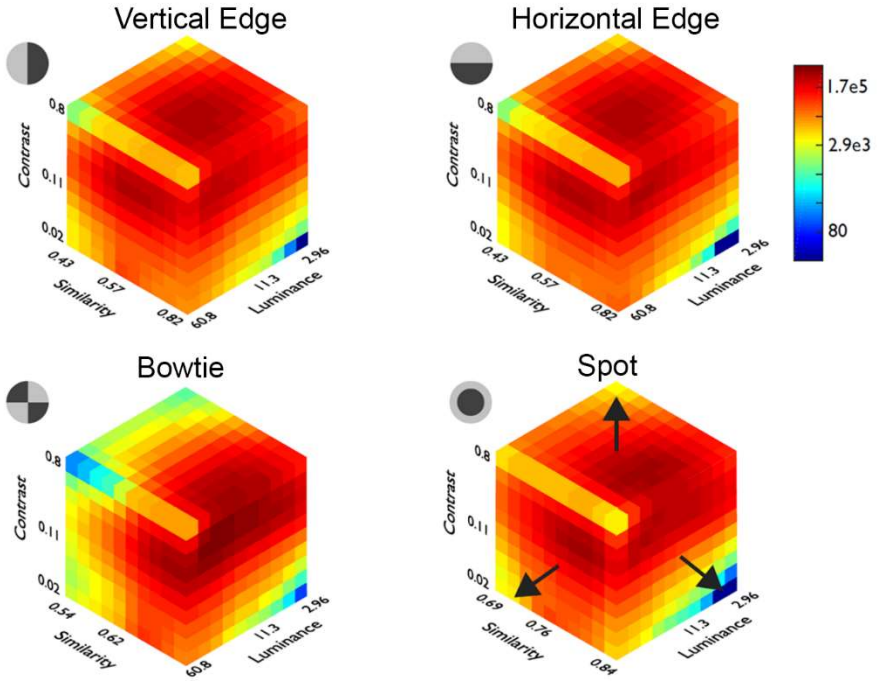

Figure 1. Joint histograms of background patches sorted along the dimensions of luminance L, contrast C and similarity S (see text for definitions of these dimensions). The bin color represents the number of patches in the bin. Because similarity depends on the specific target, there are separate histograms for each of the four targets in the study (vertical edge, horizontal edge, bowtie, spot).

Statistics were computed on ~ 150 million small natural-image patches (21 pixels diameter) that were the size of the targets used in the experiment. Patches were extracted from the large natural images at a stride of 10 pixels (half the patch width). Thus, the center of each patch was at a minimum of 10 pixels from neighboring patches. For each patch three statistics were computed, the mean luminance L of the patch, the root-mean-squared (RMS) contrast C of the patch, and the cosine similarity S between the amplitude spectrum of the patch and that of the target. The cosine similarity is the dot product of the normalized amplitude spectra of the patch and target. This measure of similarity captures similarity in orientation and spatial frequency and is independent of the phase spectra of the patch and target. Because the similarity measure depends on the specific target, similarity was computed separately for each of the four targets. The mathematical definitions of the patch statistics are given in the Appendix.

The statistics measured from the natural scene database were sorted into 3D histograms, one for each target (see Figure 1). Each dimension of the histogram contained 10 bins for a total of 1000 bins. The definitions of the bins and a table of the bin centers are given in the Appendix. Bin limits were defined after excluding the lower and upper 5% of patches along each dimension. Most bins in each histogram contain hundreds to thousands of patches.

Figure 2 shows example patches for the vertical-edge target when the luminance is at the median value. Note that as the similarity increases the background patches become more like vertical edges.

In the experiment, psychometric functions were measured for a subset of bins in the 3D histogram for the target being tested. In other words, the background selected in a trial was characterized by the statistics of a 21-pixel-diameter patch. However, the background patch presented in the trial was larger; it included the background pixels surrounding the central 21-pixel-diameter patch where the target might appear.

Psychophysical Detection Task

The experiment was conducted on three experienced psychophysical observers, including one of the authors. The experimental procedures were approved by the University of Texas Institutional Review Board and informed consent was obtained from all participants.

Stimuli

For each target, eccentricity psychometric functions were measured for 15 bins in the 3D histogram. Specifically, these consisted of either 5 or 6 bins along each of the cardinal dimensions passing through the middle of histogram (see arrows in lower right of Figure 1). Because the bin in the center of the histogram is shared by all three dimensions there were 6 bins along the dimensions of luminance and similarity and 5 for contrast. The stimuli for each target type, background bin, and eccentricity were created in the following way. A background was randomly drawn (without replacement) from the bin. This background sample subtended 4^o (241 x 241 pixels), and included both the central region where the statistics for the bin were

calculated and the surrounding context region. The background outside the 4° natural background, and entire background between trials, was set the mean luminance of the bin being tested. On target-present trials, the pixels in the central region were replaced by the target, which subtended 0.33° (21 pixels). Target definitions are provided in the Appendix. For all conditions, the mean luminance of the target was fixed at 17.8 cd/m^2, and the contrast at 33% RMS. On target absent trials, no change was made to the background. The images were then gamma-compressed based on the measured gamma of the monitor (Sony GDM-FW900) and quantized to 8-bit precision (maximum gray level = 97 cd/m^2). The images were presented at a display resolution of 60 pixels per degree (full display size 1920 x 1200 pixels).

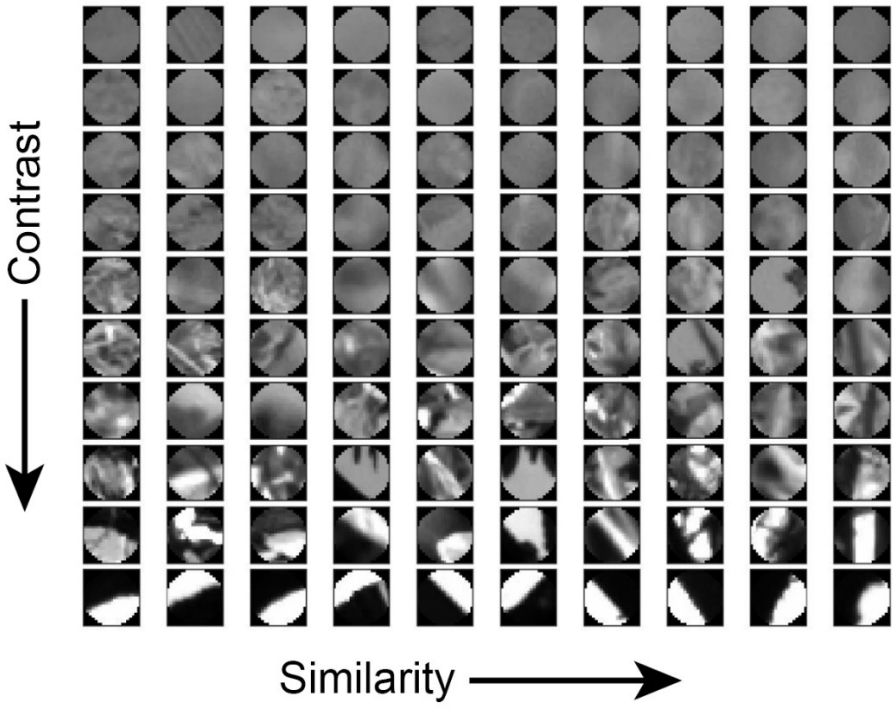

Figure 2. Example patches for the vertical target when the luminance is at the median value.

Procedure

Stimulus presentation and response collection were programmed in MATLAB using Psych Toolbox (Pelli 1997; Brainard 1997). Occluding target detection performance was measured in a yes-no task, with bias correction. We used a yes-no task because it is more similar to detection under natural conditions than is a spatial or temporal two-interval task—under natural conditions one does not get to see the same scene location with and without the target. The measurements were blocked by target type, bin, and eccentricity. For each target type and bin tested, detectability was measured at 5 eccentricities in the right visual field. The specific eccentricities were picked to span the accuracy range and varied somewhat across the observers. For each eccentricity, 60 unique stimuli were presented with 50% containing the target, for a total of 300 trials, measured in two sessions (total trials per observer = 4 targets x 15 bins x 5 eccentricities x 60 trials = 18,000) . The first session for every bin and target was completed before running the second sessions. In each session, the subject would complete 30

trials at the eccentricity closest to the fovea and then proceed to conduct blocks at farther eccentricities until all trials in that session were completed. Stimulus eccentricity was modified by moving the fixation cross to a new location, while keeping the stimulus position fixed (this method minimizes uncertainty about target location). Eye position was monitored using an EyeLink 1000 (SR Research). We found that our practiced observers made only about 1% fixation errors, and hence all trials were included in the data analysis.

A practice trial, with target present, was run at the beginning of each block and was not included in the analysis. On each trial, after a delay period of 500 ms, the stimulus appeared and remained on the screen for 250 ms. The stimulus was then removed and the subject was given 1000 ms to respond whether the target was present or absent. Feedback was then provided and another trial was initiated until all trials in the block had been completed. Eccentricity thresholds were defined to be the eccentricity at which detection performance fell to a bias-corrected accuracy of 69% correct (see below).

All subjects completed more than 1000 trials of practice in the task prior to participating in the experiment.

Analysis methods

Eccentricity psychometric functions were fit to the responses in each bin. Commonly in a detection task, a psychometric function relates target strength to a measure of performance such as percent correct. In the present case, the target strength was held at a constant value and eccentricity of the stimulus from the fovea was varied. To summarize the data and estimate eccentricity thresholds, the hit and false-alarm rates were fit simultaneously by the following descriptive equations:

$$p_{\text{hit}}(e) = \Phi\left(\frac{1}{2}d'_f \frac{e_2^\beta}{e^\beta + e_2^\beta} - \gamma\right) \quad (1)$$

$$p_{\text{fa}}(e) = \Phi\left(-\frac{1}{2}d'_f \frac{e_2^\beta}{e^\beta + e_2^\beta} - \gamma\right) \quad (2)$$

where e is the eccentricity, d'_f is the detectability in the center of the fovea, β is a steepness parameter, γ is a bias parameter, e_2 is the eccentricity at which detectability reaches half max ($d' = d'_f/2$), and $\Phi(\cdot)$ is the standard normal integral function. An eccentricity threshold was defined for each bin as the eccentricity corresponding to a detectability of 1.0 (bias corrected accuracy of 69%). The parameters of each fitted psychometric function were obtained by maximizing likelihood (see Appendix). Standard errors of the thresholds were computed by bootstrap resampling.

Results

Human Detection of Occluding Targets

Eccentricity detection thresholds were measured for four different occluding targets in natural backgrounds sampled from 15 bins in LCS space. The symbols in Figure 3A show example eccentricity psychometric functions (bias corrected) measured for the three observers for the

four targets. The curves show the fits of equations (1) and (2). As can be seen, the psychometric functions are similar across observers. The black symbols in Figures 3B-D show the average eccentricity thresholds for the three observers along the dimensions of luminance, contrast, and similarity, respectively. The blue symbols (with different saturation) show the thresholds for the individual observers. Note that low thresholds correspond to conditions where the target is less visible in the periphery (low sensitivity).

The subpanels plot the eccentricity thresholds for the different targets separately. For all four targets, the luminance dimension had a non-monotonic effect on thresholds. Eccentricity thresholds increased away from a minimum located in either the 3rd (11 cd/m^2) or 4th (21.5 cd/m^2) luminance bins. Both bins have a background luminance that is near the mean luminance of the target (17.8 cd/m^2). The background RMS contrast had a monotonically decreasing effect on eccentricity thresholds. Along the contrast dimension, eccentricity thresholds were highest at low contrast and saturated to a minimum by the highest contrast bin tested (0.8059 RMS). The dominant effect of similarity was a decrease in eccentricity threshold with increasing similarity.

Model Observer for Detection of Occluding Targets

The model observer incorporates three image cues (pattern, silhouette, edge) and four well-known properties of the human visual system (optical blur, blurring and downsampling by the ganglion cells, divisive normalization, intrinsic position uncertainty). We also evaluated model observers that excluded some of these well-known properties.

In simulating observer models, 2300 patches including the surrounding context (241 x 241 pixels) were randomly selected from each LCS bin tested in the behavioral experiment, and for each of 5 retinal eccentricities (see below). The image size was the same as the image in the behavioral experiment (4^o). A duplicate of the image was created and a target was placed in the center of the image. The cropping procedure was identical to the creation of the experimental stimulus.

In the experiment, the display resolution was 60 pixels/deg. Therefore, to match the image resolution to the sampling rate of the photoreceptors in the human fovea (120 receptors/deg), the image patches were up-sampled in size to 482 x 482 pixels by the nearest-neighbor interpolation. (Note that nearest-neighbor up-sampling simply replaces each pixel by four pixels of the same gray level.) Finally, they were padded to 512 x 512 pixels (pad gray level = mean gray level of the patch).

In the human observers, and in the model observer, target detectability (d') varies continuously as a function of retinal location. However, it is not practical to compute the model observer's detectability for every retinal eccentricity; therefore, we computed detectability at the five retinal eccentricities where ganglion-cell sampling decreased by successive factors of 2. We then fit the model observer's detectabilities with the same smooth function used to fit the human observers' detectabilities (see Eqs. 1 and 2) : $d'(e) = d'_f \, e_2^\beta \Big/ \left(e^\beta + e_2^\beta\right)$. From these smooth functions (which fit very well) we were then able to compute the predicted detectabilities at all eccentricities, as well as the predicted eccentricity thresholds.

Retinal image and ganglion-cell encoding

To model the effect of human optics, images were first passed through an optical filter that approximates the average foveal optical point-spread function, $h_o(\mathbf{x})$, of the human eye when the pupil has a diameter of 4 mm (Watson, 2013; Watson & Yellott 2012). We call this blurred image the level-0 image:

$$I_0(\mathbf{x}) = (h_0 * I)(\mathbf{x}) \tag{3}$$

where $\mathbf{x} = (x, y)$ is a pixel location, and $*$ represents the operation of convolution. (Note that here bold letters represent vectors.)

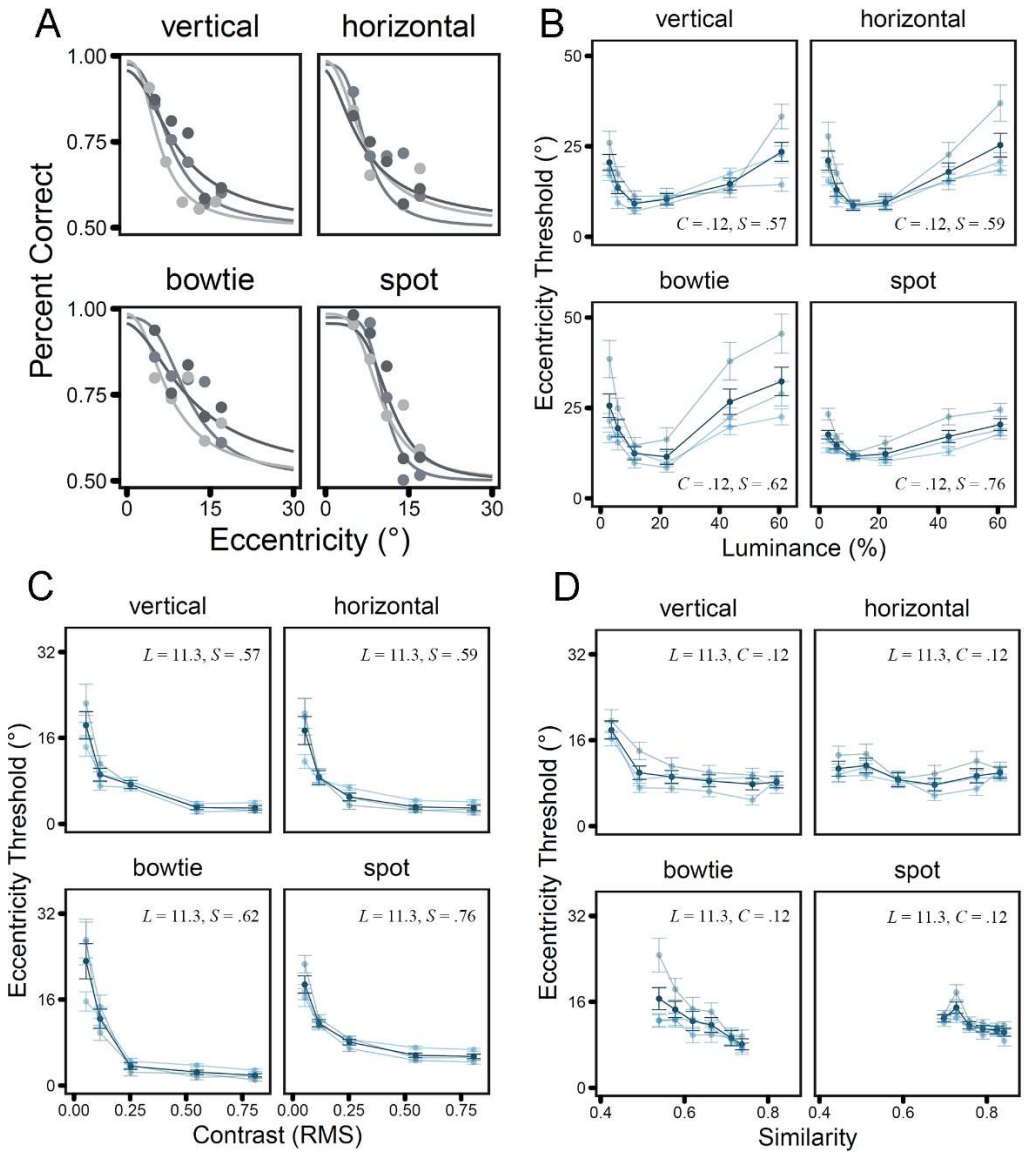

Figure 3. Psychometric functions and eccentricity thresholds. A. Example eccentricity psychometric functions for the four targets, for the central bin (5,5,5) in the histograms (see Fig. 1) Different gray-level symbols and curves are for the three observers. B-C. Average eccentricity thresholds (black symbols) and for individual observers (blue symbols) as a function of background luminance L, contrast C and similarity S, respectively. The subpanels show

the thresholds for the four different targets. In each subpanel, the two other background dimensions were held fixed at the values indicated. Error bars are boot-strapped 95% confidence intervals.

To obtain the ganglion-cell sampled image in the fovea we blurred the level-0 image by the ganglion cell center size in the fovea to obtain what we call the level-1 image:

$$I_1(\mathbf{x}) = (h * I_0)(\mathbf{x}) \tag{4}$$

To obtain the sampled image at other retinal locations we computed a multiresolution Gaussian pyramid obtained by successively blurring and down-sampling the level-1 image by factors of 2 to obtain a total of five images:

$$I_r(\mathbf{x}) = (h * I_{r-1})_{\downarrow 2}(\mathbf{x}) \quad r = 2, \cdots, 5 \tag{5}$$

The standard deviation of the Gaussian kernel used in computing the pyramid was based on the finding, in macaque, that the RF centers (half-max regions) form a complete tiling of the retinal image (Field & Chichilnisky, 2007). We approximated this rule by setting the standard deviation of the Gaussian kernel to correspond to 1 pixel in level r-1; thus,

$$h(\mathbf{x}) = \exp\left(-0.5\|\mathbf{x}\|^2\right) / \sqrt{2\pi} \tag{6}$$

The five images in the pyramid represent ganglion-cell responses at five discrete retinal distances from the center of the fovea. The retinal locations corresponding to these images were determined from anatomical data reported by Drasdo et al. (2007) from six human retinas. Specifically, we assumed 120 pairs of on and off midget ganglion cells per degree (30 arc sec spacing) in the center of the fovea (one pair for each cone), and then determined the retinal location of each pyramid level from the spacing reported in Drasdo et al. (2007). The retinal locations corresponding to the five levels are 0°, 1.3°, 4.2°, 12.3°, and 54.5° of temporal eccentricity (the direction with highest ganglion-cell density). Watson (2014) also developed a descriptive formula for the same human anatomical data. His formula is similar, but diverges substantially beyond 30° eccentricity. We also generated predictions using Watson's formula which pins the temporal retinal locations for the five levels at 0.36°, 1.8°, 5.4°, 14.9°, and 34.4°.

The 2300 unique image patches for each stimulus condition were filtered and down-sampled by the procedure described above to create a database of retina-processed images. To generate predictions, we also applied the same filtering and down-sampling to the pattern template $t(\mathbf{x})$ and the silhouette template $s(\mathbf{x})$ for each target (see next two subsections).

It is important to keep in mind that the multiresolution pyramid representation is not part of the model observer, but is just a computational trick for quickly computing model observer-performance at arbitrary eccentricities.

Pattern template response

Template matching is known to be the optimal computation for detection of targets that are added to backgrounds of Gaussian white noise (Peterson et al. 1954; Green & Swets 1966; Burgess et al. 1981). On each trial the template-matching observer multiplies the image by a template (receptive field) having the shape of the target and compares the sum of the values to a criterion (i.e., the observer compares the dot product of the image and template to a

criterion). If the response (dot product) exceeds the criterion the observer reports the target is present, otherwise that it is absent. Template matching is also known to perform well for additive targets in structured noise backgrounds (e.g., 1/f noise). In the present task, the backgrounds are natural scenes and the signals occlude rather than add to the background, therefore template matching alone is not sufficient to approach optimal performance, in large part because of the sharp bound created between the occluding target and the background. Here we assume that the pattern template response, when combined optimally with cues representing the target's silhouette and edge strength at the boundary, is a good approximation to the optimal computation.

Consider first the pattern template response. The target can be described as having two components, the mean luminance of the target l and the pattern of luminance modulation about that mean $t(\mathbf{x})$:

$$T(\mathbf{x}) = t(\mathbf{x}) + l \tag{7}$$

where $t(\mathbf{x})$ sums to zero (see Appendix).

The pattern template response to the stimulus at pyramid level r is obtained by taking the dot product of the blurred and down-sampled pattern template with the blurred and down-sampled image on that trial:

$$\mathrm{R}_P = \mathbf{t}_r \cdot \mathbf{I}_r \tag{8}$$

This pattern template response contains much of the pattern information available for detecting whether the target is present or absent. Without loss of generality the pattern template can be scaled to have an energy of 1.0, $\|\mathbf{t}_r\| = 1$. Figure 4 (left) illustrates the unscaled pattern template for the vertical-edge target, before blurring and down-sampling.

Silhouette template response

The pattern template response captures the information created by the spatial pattern internal to the target boundary, but it does not capture the information created by the difference in luminance between the target region and the surrounding background region. Much of this information can be captured with a silhouette template:

$$s(\mathbf{x}) = 2\mathrm{1}_T(\mathbf{x}) - \mathrm{1}_{T'}(\mathbf{x}) \tag{9}$$

where $\mathrm{1}_T(\mathbf{x})$ (an indicator function) is 1.0 in the target region and is 0.0 elsewhere, and $\mathrm{1}_{T'}(\mathbf{x})$ is 1.0 over an expanded target region that preserves the shape of the target as closely as possible, but contains exactly twice as many pixels, and thus $s(\mathbf{x})$ sums to 0.0. In the present case, the silhouette template has a circular center-minus-surround structure where the center, $\mathrm{1}_T(\mathbf{x})$, has the diameter of the target and the surround, $\mathrm{1}_{T'}(\mathbf{x})$, a diameter that is $\sqrt{2}$ larger. The silhouette template response to the stimulus at pyramid level r is obtained by taking the dot product of the blurred and down-sampled silhouette template with the blurred and down-sampled image on that trial:

$$R_S = \mathbf{s}_r \cdot \mathbf{I}_r \tag{10}$$

Without loss of generality the silhouette template can be scaled to have an energy of 1.0, $\|\mathbf{s}_r\| = 1$. Figure 4 (middle) illustrated the unscaled silhouette template for all four targets, before blurring and down-sampling.

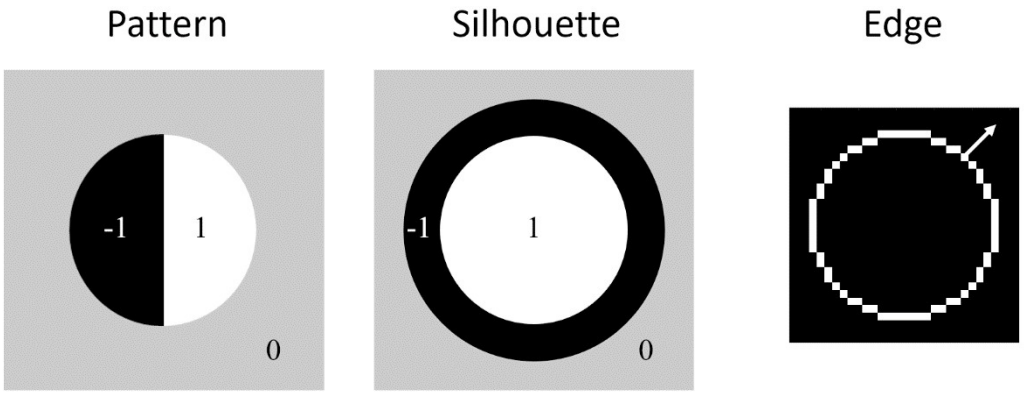

Figure 4. Pattern, silhouette, and edge cues. The pattern template (left) is for the vertical edge target. The silhouette template (middle) is the same for all four targets. Shown are the templates before optical blurring, and blurring and down-sampling by the ganglion cell array. The edge cue is the sum of the squared luminance gradient in the direction normal to the boundary (arrow) at each boundary location (white pixel), normalized by the local luminance and contrast at that boundary location. The sizes of the pattern and silhouette templates decrease with retinal eccentricity, as does number of boundary locations (ganglion cells on the boundary).

Edge-energy response

The silhouette template response captures information about the mean luminance of the target region relative to the surrounding background region, and hence represents both the mean-luminance information and some of the luminance boundary information. However, background luminance may modulate arbitrarily over space and the target luminance may also modulate over space. These luminance modulations tend to create local luminance gradients that are perpendicular to the local orientation of the target boundary and that often vary randomly in sign and amplitude. Thus, even when the mean luminance of the target and surrounding region are identical there generally remains substantial boundary information. This boundary information can be captured with an edge-energy measure, which we define to be the sum, over the boundary pixels, of the square the luminance-gradient amplitude perpendicular to the target boundary normalized by the local luminance and contrast at the boundary pixel location:

$$R_E = \sum_{\mathbf{x} \in boundary} \left[\frac{\nabla \mathbf{I}_r(\mathbf{x}) \cdot \mathbf{n}_r^{\perp}(\mathbf{x})}{L_r(\mathbf{x}) C_r(\mathbf{x})} \right]^2 \tag{11}$$

where $\nabla \mathbf{I}_r(\mathbf{x})$ is the luminance gradient, and $\mathbf{n}_r^{\perp}(\mathbf{x})$ is the unit vector perpendicular to the boundary. The boundary pixel locations are defined in the Appendix. The gradients were computed using a pair of orthogonally oriented derivative-of-Gaussian filters with a standard

deviation σ_r matched to the center size of ganglion cell RFs at that the given level of the pyramid. Derivative-of-Gaussian filters are steerable (Freeman & Adelson 1991), whereby the same gradient computation in any direction can be determined from the output of the pair of orthogonal filters. Thus, the gradient for each boundary pixel $\mathbf{x}$ was determined by computing:

$$\nabla \mathbf{I}_r(\mathbf{x}) = \left[\left(\frac{\partial g_r}{\partial x} * I_r \right)(\mathbf{x}), \left(\frac{\partial g_r}{\partial y} * I_r \right)(\mathbf{x}) \right] \quad (12)$$

with

$$\frac{\partial g_r}{\partial x}(\mathbf{x}) = -\frac{x}{\sigma_r^2} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma_r^2} \right)$$

and

$$\frac{\partial g_r}{\partial y} g_r(\mathbf{x}) = -\frac{y}{\sigma_r^2} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma_r^2} \right)$$

where $\sigma_r = \sqrt{\left(\sigma_o / 2^{r-1}\right)^2 + 1}$ and σ_o is the standard deviation of the Gaussian approximation to the optical point-spread function (recall that the standard deviation of the blur kernel at each pyramid level is 1). We defined the unit vector perpendicular to the boundary to be the unit gradient vector calculated for a uniform target on a uniform background. The local luminance and RMS contrast was computed under a Gaussian envelope having the same standard deviation (σ_r) used to compute the gradients. When the target is present, the gradient will tend to be normal to the boundary at location $\mathbf{x}$ and hence the magnitude of the dot products in equation (11) will tend to be larger when the target is present. Figure 4 (right) shows an example of the boundary pixels and a unit normal vector.

Divisive normalization

There is a long history of evidence showing that neural responses measured along the visual pathway are consistent with local divisive normalization by luminance (for review see Hood 1998) and by contrast (Heeger 1991;1992, Albrecht & Geisler 1991, for review see Carandini & Heeger 2012). Indeed, responses measured in many brain areas and in many tasks are consistent with divisive normalization (Carandini & Heeger 2012). Recently, Sebastian et al. (2017) showed that human detection thresholds for additive sinewave and plaid targets in natural backgrounds are proportional to the product of the background luminance, contrast, and similarity (as defined in the current Appendix), and that this separable Weber's law is consistent with pattern template matching (the relevant cue for their additive targets) and with divisive normalization by luminance, contrast, and similarity. They show that divisive normalization is particularly valuable under real-world conditions, where there is almost always high uncertainty about background properties and target amplitude, and where there is low prior probability of the target being present at any given location in the background. The benefit of multidimensional normalization is that it makes it possible to obtain near optimal

performance with a simple decision rule (a single fixed decision criterion). We incorporate divisive normalization into the model by computing on each trial, for every level of the pyramid, the luminance, contrast, and similarity within the pattern template region and the silhouette template region, using the definitions in the Appendix. The responses given by equations (8) and (10) are then divided by the product of the estimated luminance, contrast, and similarity (see Figure 5). Divisive normalization by local luminance and contrast is also incorporated into the edge energy measure (equation 11); however, in this case the normalization is at a smaller scale and does not include normalization by similarity (which is not practical to compute at this smaller scale).

Intrinsic position uncertainty

The final known property of the visual system we incorporate into the model observer is intrinsic position uncertainty (Swenssen & Judy 1981; Pelli 1985). Intrinsic position uncertainty (internal uncertainty about target location even when the actual location is the same on every trial) has been found to increase approximately linearly with retinal eccentricity (Michel & Geisler 2011), which is consistent with the hypothesis that the intrinsic uncertainty radius is proportional to the spacing between midget ganglion cells at any given retinal location. Given that the midget-ganglion-cell magnification factor is similar to (a little less than) the cortical magnification factor (Wässle et al. 1989), our hypothesis is that there is a roughly fixed topographic uncertainty when reading out V1. To model the effects of intrinsic position uncertainty, we use a recent analysis of position uncertainty in template-matching observers (Geisler, 2018) that found that detectability under position uncertainty is given by

$$d' = \ln\left(\frac{\exp\left[d'_0\right] + u}{1 + u}\right) \tag{13}$$

where d'_0 is the detectability without uncertainty, and u is the uncertainty constant. This formula is an approximation that holds quite accurately for the all the cases we have simulated so far. Assuming that the uncertainty standard deviation is proportional to ganglion cell spacing we used the measurements of errors in target localization as function of retinal location from Michel & Geisler (2011) to estimate that the uncertainty standard deviation in the center of the fovea is approximately 5 arc min (10 times the foveal ganglion-cell spacing). We then assumed that the standard deviation was the same number of ganglion cells at each level of the pyramid (i.e., 10 times the ganglion-cell spacing corresponding to that level of the pyramid). Finally, we then measured simulated psychometric functions (for detection of additive targets in 1/f noise) for each pyramid level, and fit the psychometric function with equation (13) to estimate the value of the uncertainty constant. The estimated values of u for levels 1 through 5 are: 3.64, 6.96, 12.86, 25.11, 39.63.

Predictions for individual cue responses

The model observer combines the pattern-template, silhouette-template and edge-energy responses when deciding whether the target is present or absent, but it is informative to first consider the three cues individually. Figure 5 shows the processing for a single cue at the eccentricity corresponding to level r in the pyramid representation.

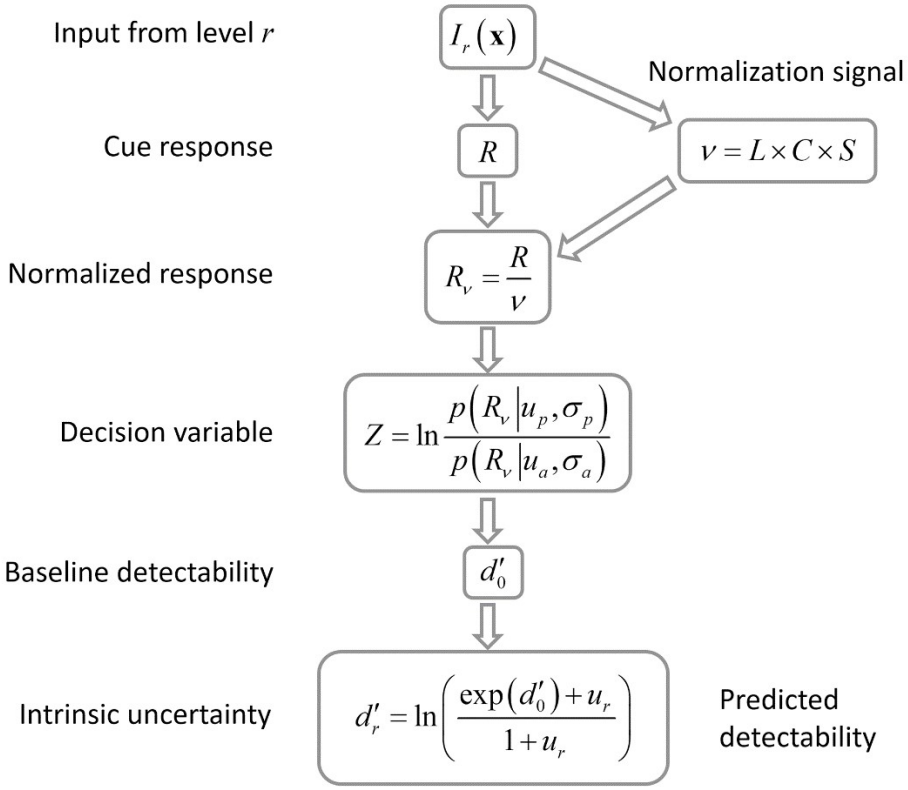

Figure 5. Processing steps for a single cue at the retinal eccentricity corresponding a single level of the multiresolution-pyramid representation. This figure is accurate for the pattern and silhouette cues (eqs. 8 and 10). For the edge energy cue (eq. 11) the normalization is in luminance and contrast at the scale of the receptive fields that measure the local luminance gradients.

The input is the blurred and down-sampled image on a trial. The cue response and the normalization signals are computed from the input image, and then combined to obtain the normalized cue response. The cue response R represents either the pattern-template response R_P (equation 8), the silhouette template response R_S (equation 10) or the edge energy response R_E (equation 11). However, we note that for the edge energy response the normalization is only by local luminance and contrast. The optimal decision variable is the log likelihood ratio of the normalized response given the response means and standard deviations estimated from thousands of target present and absent trials, in the given background bin. The distributions are approximately Gaussian distributed, but with different means and standard deviations for target present and target absent. Thus, on both target-present and target-absent trials the decision variable has (approximately) a generalized chi-squared distribution. To determine the detectability without intrinsic position uncertainty, d'_0, we integrate the generalized chi-squared distributions on those sides of the optimal decision criterion (bound) corresponding to errors to obtain the error rate p_e, which is then converted to detectability using the standard formula: $d'_0 = 2\Phi^{-1}(1 - p_e)$, where Φ^{-1} is the inverse of the standard normal integral function. (Code for integrating generalized chi-squared distributions is available at https://github.com/abhranildas/classify.) Finally, we include the effect of intrinsic position

uncertainty at the eccentricity corresponding the given level of the pyramid to obtain the predicted detectability d'_r.

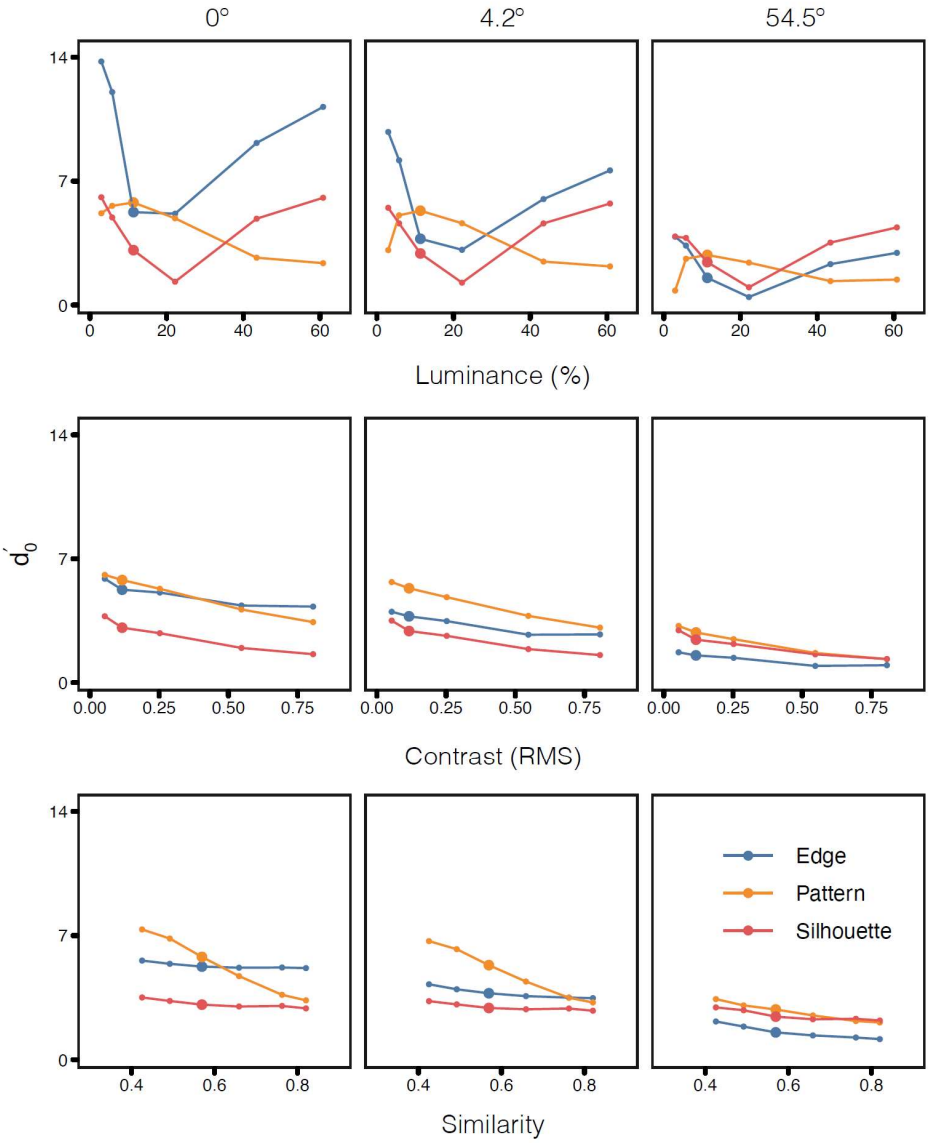

Figure 6 Single cue predictions. Predicted detectability without intrinsic position uncertainty (d'_0) a function of the three background dimensions (rows), at three retinal eccentricities (columns), for the three different cues (colors). The larger symbols indicate the values along the background dimensions that were fixed in the experimental conditions, when the value along one of the other background dimensions was varied (e.g., the big symbols in the contrast and luminance plots indicate the fixed values of luminance and contrast when similarity was varied).

Figure 6 shows the detectability of the vertical edge target, without uncertainty d'_0, as a function of the three background dimensions (rows), at three different eccentricities (columns), for each cue (colors). For all three cues and eccentricities, detectability decreases monotonical with background contrast and similarity. However, as a function of background luminance, detectability values for the edge and silhouette cue decrease to a minimum at approximately the luminance of the target and then increase. When the background luminance is near the

target luminance the pattern cue tends to provide the most useful information. When the background luminance is very different from the target luminance, then the edge and silhouette cues tend to provide the most information.

Also, for all three cues and background dimensions, detectability decreases monotonically retinal eccentricity. The effect of eccentricity is different for the three cues. At near eccentricities, detectability is generally higher for the pattern and edge cues, but at farther eccentricities, relative detectability increases for the silhouette cue, in agreement with intuition.

Predictions for joint cue responses

To generate predictions for the joint cue responses we first compute the detectability for the joint responses without intrinsic position uncertainty. We have computed the joint-response detectability in two different ways. The most general is to use a decision variable based on the multivariate normal distributions of the cue responses for target present and target absent

$$Z = \ln \frac{p\left(\mathbf{R}_v \left| \mathbf{u}_p, \boldsymbol{\Sigma}_p \right.\right)}{p\left(\mathbf{R}_v \left| \mathbf{u}_a, \boldsymbol{\Sigma}_a \right.\right)} \tag{14}$$

where $\mathbf{u}_p$, $\mathbf{u}_a$, $\boldsymbol{\Sigma}_p$, and $\boldsymbol{\Sigma}_a$ are the mean vectors and covariance matrices of the target-present and target-absent distributions. This decision variable, for both target present and target absent, is also a generalized chi-squared distribution and hence detectabilities can be computed in the same way as for the single cues. The drawback of this approach is that the covariance matrices must be estimated for all conditions. Also, one might wonder whether such a representation and computation is biologically plausible. However, we note that it is not implausible that the visual system has at least some implicit knowledge of the approximate correlations of the three cues at different retinal locations.

The other way we have computed the joint-response detectability is to assume statistical independence of the cue responses. Specifically, we compute the detectabilities for the three cues separately using the procedure in Figure 5, and then compute the combined detectability using the standard formula for independent cues from signal detection theory (Green & Swets 1966):

$$d'_0 = \sqrt{d'^2_P + d'^2_S + d'^2_E} \tag{15}$$

We note that this formula gives the detectability for reliability-weighted cue combination when the cues are statistically independent.

We have computed detectabilities both ways, but interestingly, the predictions are quite similar even though there are some substantial correlations when the target is present. Presumably this occurs because cue variances are much larger in target absent trials and hence detectability is largely determined by the covariance matrix for the target-absent trials, which is nearly a diagonal matrix (consistent with statistical independence).

Once the joint-detectabilities are computed we include the effect of intrinsic position uncertainty (final step in Figure 5). Finally, we estimate the model observer's psychometric

functions using the same procedure used to estimate the human observers' psychometric functions.

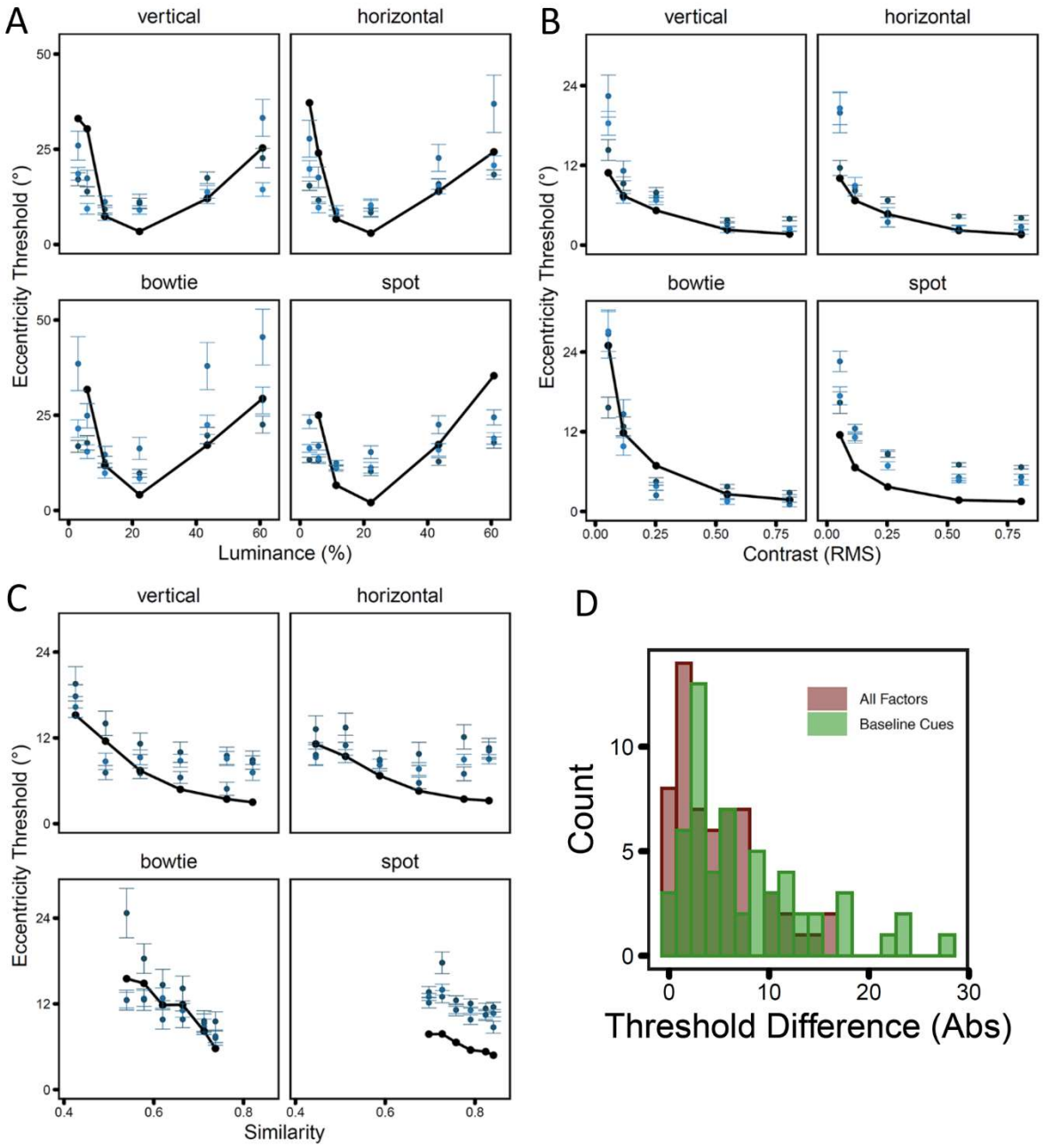

Figure 7. Comparison of human and model performance. A-C. Threshold as a function of each background dimension, for each of the four targets. Thresholds for the model observer (black symbols and solid curves). Thresholds for three human observers (blue symbols); error bars are boot strapped confidence intervals. D. Histogram of the differences in thresholds between model and human observers. The green bars are for the model observer with only the baseline cues (no normalization and no intrinsic position uncertainty) and a single overall efficiency parameter. The brown bars are for the full model observer.

To compare the predicted thresholds of the model observer with those of the human observers shown in Figure 3, we introduce a single overall efficiency parameter η that is less than 1.0 and

that multiplies all of the model observer's values of d_0'. The black symbols and curves in Figure 7A-C show the predicted thresholds for independent cue combination (equation 15, $\eta = 0.5$), and the blue symbols (and error bars) show the thresholds (and confidence intervals) of each subject in each condition (from Figure 3B-D).

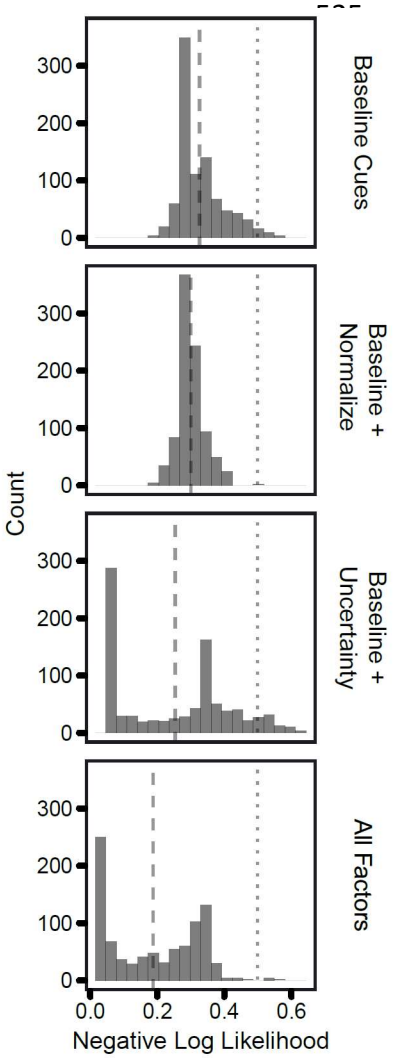

Figure 8. Negative log likelihood histograms for four model observers.

The predicted thresholds capture most the variance in the data, although there are some systematic differences. The most obvious difference is that the model observer is relatively less sensitive (thresholds are lower) for the spot target than the other targets. Another difference is the model observer is relatively more sensitive than the human observers at lower background luminance values. Nonetheless, the results suggest that many aspects of human ability to detect known occluding targets in natural backgrounds, at known locations across the visual field, are predicted from first principles by a near-optimal observer with only a single overall efficiency parameter.

We also generated predictions using Watson's (2014) formula for midget ganglion cell spacing. The predictions are almost the same as in Figure 7, but with a fit that is slightly better, especially for the spot target.

The model observer contains four known biological factors, blurring by the optics of the eye, blurring and down-sampling by midget ganglion cells, divisive normalization, and intrinsic position uncertainty. The model observer predictions are substantially worse if any of these factors are left out. Figure 7D summarizes the effect leaving out divisive normalization and intrinsic uncertainty. Specifically, the brown bars are the histogram of magnitude of the differences in model and human thresholds for all conditions and subjects (proportion of variance explained $R^2 = 0.62$). The green bars are the histogram when normalization and intrinsic uncertainty are left out ($R^2 = 0.18$). The efficiency parameter was estimated separately to obtain the best fit for each version of the model.

We also compared the goodness of fit of different versions of the model to the hits and false alarms for each condition, eccentricity, and subject. Again, in each case, a separate overall efficiency parameter η was estimated. Figure 8 shows histograms of the negative log likelihood of the observed responses according to four versions of the model: the baseline model with no divisive normalization or intrinsic position uncertainty, the baseline model with normalization, the baseline model with uncertainty, and the full model. The dashed vertical line shows the mean negative log likelihood (the smaller the negative log likelihood the more accurate prediction). For reference, the

dotted line shows the mean negative log likelihood for the degenerate model that assumes a fixed value of detectability for all conditions and eccentricities.

Discussion

Most research in the areas of visual masking, visual search, and visual memory involves identification of target objects that can be described mathematically as added to the background image, and thus some information about the background is available at each pixel location. However, under natural conditions target objects are often opaque, and hence they occlude or partially occlude the features or objects in the background. The result is that target objects have a particular internal spatial pattern that contains no background information, they create a sharp boundary in luminance and texture, and they create T-junctions from any occluded contours (e.g., Marr 1982). Relatively little is known about human detection of occluding targets.

In this study, we measured psychometric functions for detection of occluding targets in natural backgrounds that were varied along the dimensions of luminance, contrast, and similarity to the target. The psychometric functions for four occluding targets (vertical edge, horizontal edge, bowtie and spot) were measured by varying the targets' retinal eccentricity. The eccentricity psychometric functions and thresholds were similar across the three subjects tested. For all four targets, threshold decreased monotonically with background contrast and similarity and was u-shaped with background luminance (recall that lower thresholds are lower sensitivity).

To interpret the results, we developed a model of occluding-target detection that was based on three image cues: the internal spatial pattern of the target, the silhouette of target, and the edge energy created at the target boundary. The model also includes four major factors known to affect detectability: the optics of the eye, the blurring and down-sampling of the midget ganglion cell array, divisive normalization, and intrinsic position uncertainty (which increases with retinal eccentricity). These four parameter-free factors were taken directly from previous anatomy, physiology, psychophysical measurements reported in the literature. The only free parameter was a single overall efficiency parameter. We found the three cues and these well-known factors account for most of the variance in the measured psychometric functions and thresholds, although there were some systematic deviations from the predictions.

We emphasize that our aim here was to implement a minimal model that represents the main principles and factors underlying detection of known occluding targets in natural backgrounds. There should exist more biologically-plausible versions of the model that incorporate essentially the same principles and factors, and hence make similar predictions.

Occluding versus additive targets

An important difference between occluding and additive (or distortion) targets is that when the target is present there is no background signal in the ganglion-cell responses from the target region except what encroaches into the target region due to optical blur and due to blurring and down-sampling by the ganglion cell array. This causes the background properties, especially background luminance, to have different effects on detectability for occluding targets than they

do for additive targets. For additive targets detectability declines monotonically as background luminance increases (i.e., Weber's law); whereas, for occluding targets detectability is u-shaped, reaching a minimum when the luminance of the background matches that of the target (Figures 6 and 7).

Given that there is relatively little background signal from the ganglion cells in the target region, it may at first seem surprising that background luminance, contrast, and similarity have substantial effects on the detectabilities and eccentricity thresholds of occluding targets. There are several reasons for these effects. One is the encroachment of the background signals into the target region due to blurring and downsampling. The luminance, contrast, and similarity of the background in the target region it strongly correlated with luminance, contrast, and similarity in the immediately surrounding region that is encroaching into the target region.

Another reason is that detectability is determined by the variability of the cue responses in both target present and target absent trials. On target absent trials the unnormalized response standard deviations for the pattern and silhouette templates increase in proportion to the product of the background luminance, contrast, and similarity (Sebastian et al. 2017): $\sigma \propto L \times C \times S$. Detectability goes down as the standard deviation of the cue responses on target-absent trials increases.

Another reason is the divisive normalization. Sebastian et al. (2017) note that normalization has benefits for decision making under conditions of high background and target uncertainty, and because real-world detection is almost always under high degrees of uncertainty the normalization mechanisms evolved to operate automatically in the early levels of the visual system. Normalizing by the estimated values of the background properties results in template responses that, in target absent trials, have a fixed standard deviation (like a z score), independent of the values of the background properties. This makes it possible to reach near optimal performance with a single fixed decision criterion on the log-likelihood-ratio decision variables. This is particularly useful under conditions where the prior probability of a target being present at a given location is low, as it generally is when looking for a target in a natural scene (the target is absent in most image locations). Without normalization, obtaining good performance requires setting a different criterion value for every background location, which is a computational disaster for efficient parallel processing. The downside of normalization is that it involves effectively dividing by estimated quantities, which injects some variability into the decision variables. However, we note that if the decision criterion had to be estimated for every background location those estimates would also effectively inject variability into the decision variables. It should also be mentioned that the edge-energy cue already includes local normalization and we find that leaving out the local normalization reduces the usefulness (reliability) of the cue.

A somewhat surprising conclusion from the present study is that the mechanisms that predict detectability of localized occluding targets in natural backgrounds also predict the detectability of additive localized targets in natural backgrounds. These mechanisms undoubtedly evolved primarily to support detection of occluding targets, but they are also appropriate for additive targets. Sebastian et al. (2017) found that foveal thresholds for additive sinewave and plaid targets follow the separable multidimensional Weber's law, $a_t \propto L \times C \times S$, and that this

behavior was accurately predicted with only the pattern-template cue together with divisive normalization by the product of background luminance, contrast, and similarity. This result is consistent with the present study because in the Sebastian et al. experiment the silhouette and edge cues provide no information (and hence could be down-weighted in the model observer), and because intrinsic position uncertainty has a minimal effect in the fovea. Also, it is easy to show, using equation (13), that position uncertainty only affects overall efficiency, not the prediction of separable multidimensional Weber's law (see Appendix). In short, the Sebastian model for additive targets is a special case of the current model observer for occluding targets.

Crowding effects in natural backgrounds

There is much evidence that in certain kinds of cluttered backgrounds the falloff in identification performance with eccentricity is more rapid than the falloff in visual acuity measured on uniform backgrounds (Bouma 1970; for reviews see Levi 2008; Pelli & Tillman 2008; Whitney & Levi 2011). These "crowding" effects are strongest when there is high similarity between target and background features (Whitney & Levi 2011). The observer model described here includes some factors that may contribute to the crowding effects reported in the literature, including blurring and down-sampling by the midget ganglion cells array and intrinsic position uncertainty. Crowding effects are known to be more consistent with the increase in cortical-cell spacing (cortical magnification) than with the increase in cone-photoreceptor spacing. The midget-ganglion-cell RF spacing increases more rapidly than that of cone photoreceptors, and only slightly less rapidly than cortical-cell RF spacing in V1 (Wässle et al. 1989). Thus, intrinsic position uncertainty also increases approximately in proportion to cortical-cell RF spacing. Nonetheless, it is unlikely that these factors alone can account for the magnitude of the crowding effects often reported.

This raises the question of why the simple principled model described here is able to predict, fairly well, human identification performance in natural backgrounds, which are highly complex and cluttered. A plausible hypothesis is that in our experiments the similarity between the target and background features (luminance, contrast, orientation, shape) is on average relatively low, which is just the situation where crowding effects are known to be weaker. In all the conditions of our experiment there was a least some difference in luminance and contrast with the background. This is undoubtedly representative of most real-world conditions, where arbitrary target objects are viewed from an arbitrary direction.

There is evidence of crowding effects in natural backgrounds when the features of the target and background are similar. For example, Freeman & Simoncelli (2011) find that human ability to discriminate distorted from undistorted natural image patches decreases with eccentricity in a fashion consistent with classic crowding effects and with the sizes of receptive fields in macaque V2. Similarly, Wallis & Bex (2012) find that human ability to detect occluding "dead-leaves" textures (having the same luminance and contrast as the natural background against which they appear) also decreases with eccentricity at a rate consistent with crowding effects and with the findings of Freeman & Simoncelli (2011).

Edge-energy cue and detection of camouflaged objects

An important special case, where the similarity of target and background features is high, is when the target objects are organisms that have evolved camouflage that mimics the backgrounds against which they normally appear (Stevens & Merilaita 2011). Recently, Das & Geisler (2018) considered the limiting (maximally-camouflaged) case where the target and background texture are random samples of the same texture. In this case, both the pattern and the silhouette cues provide little or no information. Das & Geisler find that human detection accuracy as a function of edge energy is the same for both 1/f noise and synthetic bark textures, supporting the appropriateness of the edge-energy cue in modeling occluding target detection. A future direction will be to measure how detection of maximally camouflaged targets varies with retinal eccentricity. This is an important case where crowding mechanisms may play a particularly strong role.

Silhouette cue

Most target objects in natural scenes differ in mean luminance from the background that surrounds them, and thus the silhouette cue provides useful identification information. For our small targets, the silhouette information tends to be most useful at larger eccentricities (Figure 6). Although as yet unexplored, it is intuitive that the relative value of the silhouette cue will increase when the target is larger and the target's surface texture is finer or of lower contrast. The silhouette template needs a center-surround structure in order to be useful for identification in scenes where the target location is not known. For example, a template that is constant in the target region and zero elsewhere will produce the same response at every location in a large uniform field that it does to a uniform patch that just fills the template. The results from classification-image experiments (Eckstein et al., 2007) and from visual search optimization (Zhang et al., 2009) are consistent with center-surround templates. Finally, we note that the luminance normalization in the model observer, makes all three cue responses invariant to scaling the overall illumination of a natural scene.

Conclusion

Eccentricity thresholds were measured in three subjects for four different occluding targets as function the luminance, contrast, and phase-invariant similarity of the background to the target. The complex pattern of results was consistent across subjects and was largely explained by a principled model observer (with only a single efficiency parameter) that combines three image cues and four well-known properties of the human visual system. The model observer is a generalization of an earlier model for foveal detection of additive targets in natural backgrounds. The results and model observer should help to lay a foundation for a more general theory of visual search and object identification in natural and other complex backgrounds.

Acknowledgments

Supported by NIH grant EY011747.

References

Alam, M. M., Vilankar, K. P., Field, D. J., & Chandler, D. M. (2014). Local masking in natural images: A database and analysis. *Journal of Vision*, 14(8):22, 1–38.

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.

Bex, P. J. (2010). (In) sensitivity to spatial distortion in natural scenes. *Journal of Vision*, 10(2):23, 1‑15

Bex PJ, Solomon SG, & Dakin SC (2009).Contrast sensitivity in natural scenes depends on edge as well as spatial frequency structure, *Journal of Vision*, 9(10):1, 1‑19

Bouma, H. (1970) Interaction effects in parafoveal letter recognition. *Nature* 226, 177–178

Bradley C, Abrams J & Geisler WS (2014) Retina-V1 model of detectability across the visual field. *Journal of Vision,* 14(12):22, 1-22.

Brainard, D. H. (1997) The Psychophysics Toolbox, *Spatial Vision 10*:433-436.

Burgess AE (2018) Signal detection: a brief history, in Samei E & E. Krupinski E, Eds. *The Handbook of Medical Image Perception and Techniques, Second Edition*. Cambridge: Cambridge University Press.

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

Caelli T. & Moraglia G. (1986) On the detection of signals embedded in natural scenes. *Perception and Psychophysics*, 39, 87-95.

Campbell F. W. & Kulikowski J. J. (1966) Orientation selectivity of the human visual system. *Journal of Physiology*, 187, 437-445.

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

Das A & Geisler WS (2018) Understanding camouflage detection, *Journal of Vision,* 18(10):549-549.

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 MP, Beutter BR, Pham BT, Shimozaki SS & Stone LS (2007) Similar neural representations of the target for saccades and perception during search. *Journal of Neuroscience*, 27(6), 1266 –1270.

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

761 Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*,
762 14(9), 1195‑1201.

763 Freeman, W. T. & Adelson, E. H. (1991). The design and use of steerable filters. *IEEE*
764 *Transactions on Pattern Analysis and Machine Intelligence*, **13** (9): 891–906.

765 Geisler, WS (2008) Visual perception and the statistical properties of natural scenes. *Annual*
766 *Review of Psychology*, 59, 167-192.

767 Geisler W. S. (2011) Contributions of ideal observer theory to vision research. *Vision Research*,
768 51, 771-781.

769 Geisler WS (2018) Psychometric functions of uncertain template matching observers, *Journal of*
770 *Vision* 18(2):1, 1–10.

771 Green DM, & Swets JA (1966) *Signal Detection Theory and Psychophysics*. New York: Wiley.

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

775 Heeger DJ (1992) Normalization of cell responses in cat striate cortex. *Visual Neuroscience, 9*,
776 191-197.

777 Hood DC (1998) Lower-level visual processing and models of light adaptation. *Annual Review of*
778 *Psychology*, 49, 503-535.

779 König A & Brodhun E (1889) Experimentelle Untersuchungen über die psycho-physische
780 Fundamentalformel in Beug auf den Gesichtssinn. *Zweite Mittlg. S. B. Preuss. Akad. Wiss*., p.
781 641.

782 Legge GE & Foley JM (1980). Contrast masking in human vision. Journal of the Optical Society of
783 America, 70, 1458–1471.

784 Marr D, (1982) *Vision. A Computational Investigation into the Human Representation and*
785 *Processing of Visual Information.* W.H. Freeman and Company.

786 Michel MM & Geisler WS (2011) Intrinsic position uncertainty explains detection and
787 localization performance in peripheral vision*. Journal of Vision*, 11(1):18, 1-18.

788 Mueller CG (1951) Frequency of seeing functions for intensity discrimination at various levels of
789 adapting intensity. *Journal of General Physiology*, 34, 463-474.

Nadenau MJ, Reichel J, & Kunt M (2002) Performance comparison of masking models based on a new psychovisual test method with natural scenery stimuli. *Signal Processing: Image Communication*, 17(10), 807–823.

Pelli DG (1985) Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America A*, 2, 1508-1532.

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

Pelli, D.G. and Tillman, K.A. (2008) The uncrowded window of object recognition. Nat. Neurosci. 11, 1129–1135

Peterson WW, Birdsall TG, and Fox WC (1954) The theory of signal detectability. *Trans. IRE PGIT-4*, 171-212.

Rohaly AM, Ahumada AJ, & Watson AB (1997*)* Object Detection in natural backgrounds predicted by discrimination performance and models*, Vision Research,* 37, 3225-3235*.*

Sebastian S., Abrams J. & Geisler W.S. (2017) Constrained-sampling experiments reveal principles of detection in natural scenes. *Proceedings of the National Academy of Sciences*, 14:28, E5731–E5740.

Stevens M & Merilaita S (2011) Animal camouflage: Function and mechanisms. In Stevens M & Merilaita S (Eds) *Animal Camouflage: Mechanisms and Function*. New York: Cambridge University Press.

Stromeyer CF & Julesz B (1972) Spatial frequency masking in vision: Critical bands and the spread of masking. *Journal of the Optical Society of America*, 62, 1221-1232.

Swensson RG and Judy PF (1981) Detection of noisy visual targets: models for the effects of spatial uncertainty and signal-to-noise ratio, *Perception & Psychophysics,* 29, 521-534.

Wallis TSA, & Bex PJ (2012) Image correlates of crowding in natural scenes. *Journal of Vision*, 12(7):6, 1–19.

Watson AB (2013) A formula for the mean human optical modulation transfer function as a function of pupil size, *Journal of Vision* 13(6):18, 1–11.

Watson AB (2014) A formula for human retinal ganglion cell receptive field density as a function of visual field location, *Journal of Vision* 14(7):14, 1–17

Watson AB & Yellot JI (2012) A unified formula for light-adapted pupil size. *Journal of Vision* 12(10):12, 1–16

Wässle H, Grunert U & Rohrenbeck J (1989) Cortical magnification factor and the ganglion cell density of the primate retina. *Nature*, 341, 643-646.

Whitney D & Levi DM (2011) Visual crowding: a fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, 15(4), 160-168.

Wilson, HR, McFarlane, DK, & Phillips, GC (1983) Spatial-frequency tuning of orientation selective units estimated by oblique masking. *Vision Research*, 23(9), 873–882.

Zhang S, Abbey CK, & Eckstein MP (2009) Virtual evolution for visual search in natural images results in behavioral receptive fields with inhibitory surrounds. *Visual Neuroscience*, 26, 93–108.

Appendix

Target definitions

The vertical edge target pattern was defined by

$$t(\mathbf{x}) = \begin{cases} 1 & \text{if } x < 0 \\ -1 & \text{if } x > 0 \\ 0 & \text{if } (x = 0) \text{ or } \sqrt{x^2 + y^2} > \rho \end{cases} \quad \text{(A1)}$$

The horizontal edge target pattern was defined by

$$t(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathrm{y} > 0 \\ -1 & \text{if } \mathrm{y} < 0 \\ 0 & \text{if } (\mathrm{y} = 0) \text{ or } \sqrt{x^2 + y^2} > \rho \end{cases} \quad \text{(A2)}$$

The bowtie target pattern was defined by

$$t(\mathbf{x}) = \begin{cases} 1 & \text{if } (\mathrm{y} > 0 \ \wedge \ \mathrm{x} > 0) \ \vee \ (\mathrm{y} < 0 \ \wedge \ \mathrm{x} < 0) \\ -1 & \text{if } (\mathrm{y} < 0 \ \wedge \ \mathrm{x} > 0) \ \vee \ (\mathrm{y} > 0 \ \wedge \ \mathrm{x} < 0) \\ 0 & \text{if } (x = 0) \ \vee \ (\mathrm{y} = 0) \ \vee \ \sqrt{x^2 + y^2} > \rho \end{cases} \quad \text{(A3)}$$

The spot target pattern was defined by

$$\mathrm{t}(\mathbf{x}) = \begin{cases} 1 & \text{if } \delta < \sqrt{x^2 + y^2} \le \rho \\ -1 & \text{if } \sqrt{x^2 + y^2} \le \delta \\ 0 & \text{if } \sqrt{x^2 + y^2} > \rho \end{cases} \quad \text{(A4)}$$

In these definitions $\mathbf{x} = (x, y)$, ρ defines the radius of the target region and was set as 10 pixels (at 60 pixels/deg), and δ defines the radius of the interior circular region of the target. The radius for the interior region of the spot was 7 pixels. All the target patterns satisfied the property that $\sum_{\mathbf{x}} t(\mathbf{x}) = 0$.

The displayed target was obtained by adding a mean luminance l: $T(\mathbf{x}) = t(\mathbf{x}) + l$

Definitions of background properties

The local mean luminance of each patch was defined as:

$$L = \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{x}_i) \quad \text{(A5)}$$

where n specifies the number of pixels in the patch, $\mathbf{x}_i$ are the coordinates of pixel i, and $I(\mathbf{x}_i)$ is the luminance of pixel i.

The root-mean-squared (RMS) contrast was defined as:

$$C = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(I(\mathbf{x}_i) - L)^2}{L^2}} \quad \text{(A6)}$$

The phase-invariant similarity was defined as the cosine of the angle between the Fourier amplitude spectrum of the target $A_T(u,v)$ and the Fourier amplitude spectrum of the patch $A_I(u,v)$,

$$S = \frac{\mathbf{A}_T \cdot \mathbf{A}_I}{\|\mathbf{A}_T\|\|\mathbf{A}_I\|} \quad \text{(A7)}$$

where u and v are the horizontal and vertical spatial frequency. In other words, S is the dot product of the amplitude spectra, represented as vectors normalized to a length of 1.0. To prevent artifacts in the Fourier transform, the amplitude spectrum of the target was obtained by first windowing the target with a circular aperture having a raised cosine ramp width at the edge of 2 pixels:

$$w(\mathbf{x}) = \begin{cases} 1.0 & \text{if } \|\mathbf{x} - \mathbf{x}_c\| < r_1 \\ 0.5 + 0.5\cos\left(\pi\left(\|\mathbf{x} - \mathbf{x}_c\| - r_1\right)/\left(r_0 - r_1\right)\right) & \text{if } r_1 < \|\mathbf{x} - \mathbf{x}_c\| < r_0 \\ 0.0 & \text{if } \|\mathbf{x} - \mathbf{x}_c\| > r_0 \end{cases} \quad \text{(A8)}$$

Then a fast Fourier transform (FFT) was applied and the complex absolute value of the Fourier spectrum was taken. To obtain the amplitude spectrum of the patch, the patch was first windowed in the same way and then the complex absolute value of the FFT was taken, after subtracting the mean.

Definition of boundary pixels

Here we define the boundary pixels for each pyramid level. Let the center of the target region at level 1 of the pyramid be (x_1, y_1) and the radius (in pixels) of the target region be ρ_1, then the center and the radius at level r are given by $(x_r, y_r) = (x_1/2^{r-1}, y_1/2^{r-1})$, and $\rho_r = \rho_1/2^{r-1}$.

For each image pixel location (x_i, y_i) the direction of the pixel from the center is $\theta_i = \operatorname{atan2}(y_i - y_r, x_i - x_r)$ and the real-valued location on the boundary (x, y) is given by

$x = \rho_r \cos\theta_i + x_r$, and $y = \rho_r \sin\theta_i + y_r$. An image pixel location is defined to be a boundary location if $\sqrt{(y - y_i)^2 + (x - x_i)^2} < 0.5$.

Luminance, contrast and similarity Bins

The widths of the bins were determined by a geometric spacing rule:

$$a = \left(x_{\max} / x_{min}\right)^{1/n}$$

where n defines the number of bins and $x_{\min}$ and $x_{\max}$ define the minimum and maximum of the lower and upper bins. The bin boundaries are determined from the spacing rule

$$x_i = x_{\min} a^i$$

where and i is the index for the i^{th} bin. The values for $x_{\min}$ and $x_{\max}$ were determined as the 5th and 95th percentile of the scene statistics distributions for each dimension. Table 1 shows the center values of the bins.

Table 1 - Bin Centers

Bin	Luminance (% Max)	Contrast (RMS)	Similarity			
			Vertical	Horizontal	Bowtie	Spot
1	2.96	0.0247	0.4255	0.4454	0.5394	0.6973
2	4.1447	0.0364	0.4577	0.4774	0.5584	0.712
3	5.7984	0.0536	0.4923	0.5117	0.5781	0.727
4	8.1117	0.079	0.5295	0.5484	0.5986	0.7423
5	11.348	0.1163	0.5696	0.5878	0.6197	0.7579
6	15.8755	0.1713	0.6127	0.6299	0.6416	0.7739
7	22.2093	0.2523	0.659	0.6751	0.6643	0.7902
8	31.07	0.3716	0.7088	0.7236	0.6877	0.8069
9	43.4659	0.5472	0.7625	0.7755	0.712	0.8239
10	60.8073	0.8059	0.8201	0.8312	0.7372	0.8412

Maximum likelihood estimation of psychometric function parameters

The parameters of the psychometric functions were estimated via maximum likelihood estimation. The log-likelihood function was defined as,

888 $$\ln L(\boldsymbol{\theta} = (e_2, \beta, \gamma) \mid e) = \sum_{i=1}^{n} N_h(e) \ln P_h(e \mid \boldsymbol{\theta}) + \sum_{i=1}^{n} N_{fa}(e) \ln P_{fa}(e \mid \boldsymbol{\theta}) + \sum_{i=1}^{n} N_m(e) \ln[1 - P_h(e \mid \boldsymbol{\theta})] + \sum_{i=1}^{n} N_{cr}(e) \ln[1 - P_{fa}(e \mid \boldsymbol{\theta})]$$

889 where $N_{\cdot}(e)$ are the number of hits, false alarms, misses and correct rejections and $\boldsymbol{\theta}$ is the
890 vector of parameters with the maximum log likelihood

891 $$\hat{\boldsymbol{\theta}}_{\max} = \arg\max_{\boldsymbol{\theta} \in \Omega} \left[\ln L(\boldsymbol{\theta} \mid e) \right]$$

892 **Effect of intrinsic position uncertainty on separable Weber's law**

893 Sebastian et al. (2017) found that the detectability of the pattern template observer without
894 intrinsic position uncertainty, for additive targets in natural backgrounds, is given
895 approximately by

896 $$d_0' \propto \frac{a}{L \times C \times S}$$

897 where a is the target amplitude, and L, C, and S are the background luminance, contrast, and
898 similarity. Setting the detectability to 1.0 shows that the model observer's threshold satisfies
899 separable Weber's law: $a_t \propto L \times C \times S$. Text equation (13) shows that the effect of intrinsic
900 position uncertainty on detectability is given by

901 $$d' = \ln\left(\frac{\exp\left[d_0' \right] + u}{1 + u} \right)$$

902 Substituting into this equation and setting to 1.0 shows that

903 $$\frac{a_t}{L \times C \times S} = \ln\left[e(1 + u) - u \right]$$

904 Thus, given that e and u are constants, we still have separable Weber's law: $a_t \propto L \times C \times S$.

905