Constrained sampling experiments reveal principles of detection in natural scenes

Stephen Sebastian[a,b], Jared Abrams[a,b], and Wilson S. Geisler[a,b,1]

[a]Center for Perceptual Systems, University of Texas at Austin, Austin, TX 78712; and [b]Department of Psychology, University of Texas at Austin, Austin, TX 78712

Edited by Randolph Blake, Vanderbilt University, Nashville, TN, and approved June 2, 2017 (received for review November 27, 2016)

A fundamental everyday visual task is to detect target objects within a background scene. Using relatively simple stimuli, vision science has identified several major factors that affect detection thresholds, including the luminance of the background, the contrast of the background, the spatial similarity of the background to the target, and uncertainty due to random variations in the properties of the background and in the amplitude of the target. Here we use an experimental approach based on constrained sampling from multidimensional histograms of natural stimuli, together with a theoretical analysis based on signal detection theory, to discover how these factors affect detection in natural scenes. We sorted a large collection of natural image backgrounds into multidimensional histograms, where each bin corresponds to a particular luminance, contrast, and similarity. Detection thresholds were measured for a subset of bins spanning the space, where a natural background was randomly sampled from a bin on each trial. In low-uncertainty conditions, both the background bin and the amplitude of the target were fixed, and, in high-uncertainty conditions, they varied randomly on each trial. We found that thresholds increase approximately linearly along all three dimensions and that detection accuracy is unaffected by background bin and target amplitude uncertainty. The results are predicted from first principles by a normalized matched-template detector, where the dynamic normalizing gain factor follows directly from the statistical properties of the natural backgrounds. The results provide an explanation for classic laws of psychophysics and their underlying neural mechanisms.

natural scene statistics | detection | masking | normalization | Weber's law

Visual systems are the result of evolution by natural selection, and, as a consequence, their design is strongly constrained by the properties of natural visual stimuli and by the specific visual tasks performed to survive and reproduce. Thus, to understand the human visual system, it is critical to characterize natural visual stimuli and performance in natural visual tasks (1, 2).

Perhaps the most fundamental visual task is to identify target objects in the natural backgrounds that surround us. It is known that the specific properties of a background can have a strong influence on detectability (Fig. 1). For example, the detectability of a target pattern with given amplitude decreases with increases in background luminance (3, 4), background contrast (5–7), and the similarity of the spatial properties of the background to those of the target (8–12). In addition to the direct effects of such background properties, there are other factors that affect detection performance. Specifically, under natural conditions, the strength (amplitude) and location of the target often randomly vary on every occasion, and the target typically appears against a different background scene on every occasion. The uncertainty created by the random amplitude and location of the target ("target uncertainty") and the random variation in the properties of the background ("background uncertainty") are additional factors that can reduce detection performance (12–15).

What is relatively unknown are (*i*) how these various factors individually affect detection accuracy in natural scenes, (*ii*) how they combine in affecting detection accuracy in natural scenes, and (*iii*) how these factors and the underlying neural mechanisms are related to the statistical properties of natural scenes. Although there have been a number of studies of detection in natural backgrounds (16–23), they have not directly addressed these questions, and have either tested only a small number of natural stimuli (16, 17, 19, 20), tested natural stimuli with altered statistical properties (21, 22), or used experimental paradigms not representative of natural detection tasks (16, 18–20, 23). These latter studies are not as representative of natural tasks, because observers were allowed to directly compare the same image with and without the added target, an advantage that is not normally available under real-world conditions.

Here we address the three questions above using an experimental approach based on sampling from multidimensional histograms of natural stimuli, together with a theoretical analysis based on signal detection theory. This constrained sampling approach is efficient and could be used to address similar questions for other natural tasks. The first step is to obtain a large collection of calibrated natural images. These images then are divided into millions of background patches that are sorted into narrow bins along dimensions of interest. In the present study, each histogram bin corresponds to a narrow range of mean luminance, contrast, and similarity of the target to the background patch. Next, detection performance for simple targets is measured by sampling from a sparse subset of bins spanning the space. In one set of experiments, performance is measured one bin at a time (no background and amplitude uncertainty), and, in the second set of experiments, the bin is randomly selected on each trial (high background and amplitude uncertainty). In both sets of experiments, the location of the target (if present) is fixed; i.e., we did not consider location uncertainty (*Discussion*). The experiments revealed lawful effects of luminance, contrast, and similarity on detection performance, and showed that humans are remarkably

Significance

The visibility of a target object may be affected by the specific properties of the background scene at and near the target's location, and by how uncertain the observer is (from one occasion to the next) about the values of the background and target properties. An experimental technique was used to measure how several background properties, and uncertainty, affect human detection thresholds for target objects in natural scenes. The thresholds varied in a highly lawful fashion—multidimensional Weber's law—that is predicted directly from the statistical structure of natural scenes. The results suggest that the neural gain control mechanisms underlying multidimensional Weber's law evolved because they are optimal for detection in natural scenes under conditions of high uncertainty.

Author contributions: S.S., J.A., and W.S.G. designed research; S.S., J.A., and W.S.G. performed research; S.S. and W.S.G. analyzed data; and S.S. and W.S.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: w.geisler@utexas.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1619487114/-/DCSupplemental.

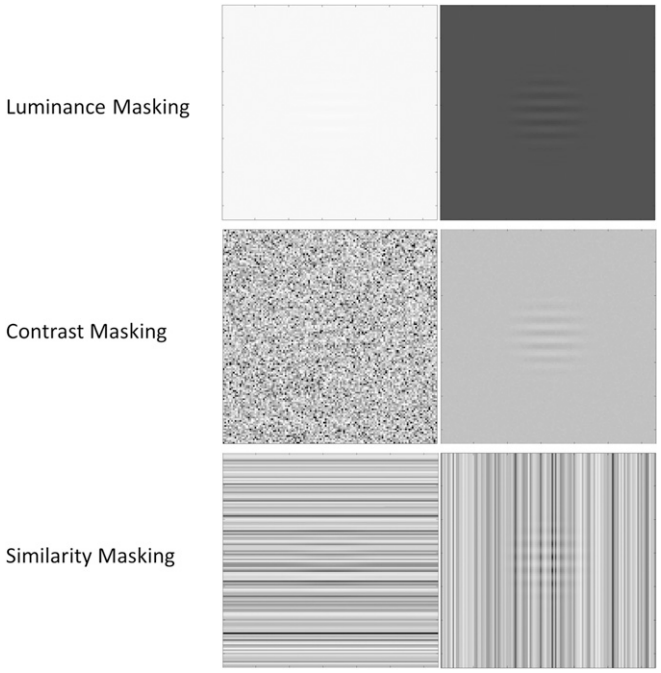

Fig. 1. Demonstration of the major dimensions of masking. In each case, the same target (a windowed sinewave grating) is added to the backgrounds on the left and the right. The target is less visible on the left. (*Top*) backgrounds are uniform with higher luminance on the left. (*Middle*) The backgrounds are 2D Gaussian noise with higher contrast on the left; mean luminance is the same on both sides. (*Bottom*) Backgrounds are 1D Gaussian noise with a horizontal orientation on the left; contrast and mean luminance are the same on both sides.

unaffected by background and amplitude uncertainty. These aspects of human detection performance are predicted quantitatively from first principles by a signal detection analysis of the natural image stimuli. This analysis provides an understanding of the computational principles and evolutionary pressures that underlie the classic laws of visual psychophysics and their associated neural mechanisms.

Results

The two major aims of the present experiments were (*i*) to determine how the local luminance, contrast, and target similarity of natural backgrounds affect the detectability of targets and (*ii*) to determine how uncertainty about the background properties and target amplitude affect detectability.

Natural Background Stimuli. To obtain stimuli for the experiments, we analyzed a large collection of high-resolution gray-scale natural images (Fig. 2*A* and *Methods*). The 1,204 images were divided up into millions of background patches that were the same size (0.8° wide) as the targets used in the detection experiments. For each patch, the mean luminance, root-mean-squared (RMS) contrast, and phase-invariant similarity to the target were measured (see *Methods* for definitions of these measures). These values were used to sort the patches into 3D histograms. Fig. 2*B* shows the histograms for the two targets used in the experiments. (Note that similarity depends on both the target shape and size and hence separate histograms were measured for the two targets.) Most of the 1,000 bins in each of these histograms contained many hundreds of patches, and all of the patches within a bin had nearly the same mean luminance, contrast, and similarity. However, we note that not all image patches fell into a bin, primarily because we restricted the contrast to a maximum of 0.32 RMS and the mean luminance to 0.55 of the maximum luminance in the image—the maxima for which it is practical to measure thresholds on standard displays, like those in the present experiments (*Methods*). Fig. 2*C* shows single background patches (0.8° wide), each randomly sampled from one of 25 bins with the same mean luminance but different RMS contrast and similarity.

Experiment 1: Detection Thresholds with Background Context Present. In the first experiment, detection thresholds were measured in the fovea along each of the three dimensions separately, with the other two held fixed at their median values. Thresholds were measured for two different targets (Fig. 2*B*), one having a single dominant orientation [a windowed 4-cycles-per-degree (cpd) grating] and one with two dominant orientations (a windowed 4-cpd plaid). To obtain the thresholds, psychometric functions were measured in a single-interval identification task with feedback (Fig. 2*D*). On each trial, a background patch was randomly sampled without replacement from the bin being tested. The background that was displayed also included a context region 4.3° wide surrounding the target region. The rest of the display contained a fixed luminance equal to that of the bin. In experiment 1, both the target amplitude and the background bin were "blocked" (i.e., fixed for all trials of a block). Randomly, on half of the trials, the target was added to the background, and the subject reported whether the target was present or absent. Thresholds were defined to be the target amplitude giving 69% correct decisions (*Methods*). (Note that most studies report thresholds in units of contrast.)

The average amplitude thresholds for three subjects are shown in Fig. 3; Fig. 3, *Upper* is for the grating target, and Fig. 3, *Lower* is for the plaid target (results for individual subjects were similar; see Fig. S1). Each plot shows the threshold (solid symbols) as a function of the value along a background dimension. For both targets, the threshold amplitude increased approximately linearly with local mean luminance in natural backgrounds (solid lines), in agreement with the classic finding of Weber's law reported for detection in uniform backgrounds (3, 4). Similarly, the threshold amplitude increased linearly with background RMS contrast (at contrasts above a few percent), in agreement with the classic finding for detection in white noise (7) and more recent findings for targets in 1/f noise (22, 24) and in Gaussianized natural backgrounds (22). Finally, amplitude threshold increases approximately linearly with the similarity of the background to the target, a result not previously reported.

The primary conclusion from this experiment is that thresholds increase approximately linearly along each of the cardinal directions for detection in natural backgrounds, with different slopes and intercepts for each dimension and target. In *Signal Detection Analysis of Detection Under Blocked Conditions*, we show that this entire pattern of results follows directly from a principled signal detection theory analysis of detection in natural images.

Experiment 2: Detection Thresholds with Background Context Removed. The first experiment measured detection thresholds when the backgrounds were substantially larger (4.3° wide) than the target (0.8° wide). It is possible that the surrounding region is helpful because it provides information about the properties of the background in the target region. To examine the effect of the surrounding region on threshold, we carried out a second experiment where the background was windowed to the size of the target. This experiment also provided some of the baseline data for experiment 3 that measured the effects of background and target amplitude uncertainty. For practical reasons (see experiment 3), we fixed the luminance at the median value in both experiments 2 and 3.

Fig. 4 plots the detection thresholds as a function of contrast and similarity (individual subject data are in Fig. S2). For comparison, the contrast and similarity thresholds from experiment 1 also are plotted, in a lighter shade. For both targets, threshold still increased linearly with the background contrast and similarity. However, the overall magnitude of the thresholds was lower. One possibility is that the decrease is due to reduced uncertainty about the location of the target (25, 26). However, between trials, a fixation point was kept at the center of the target location,

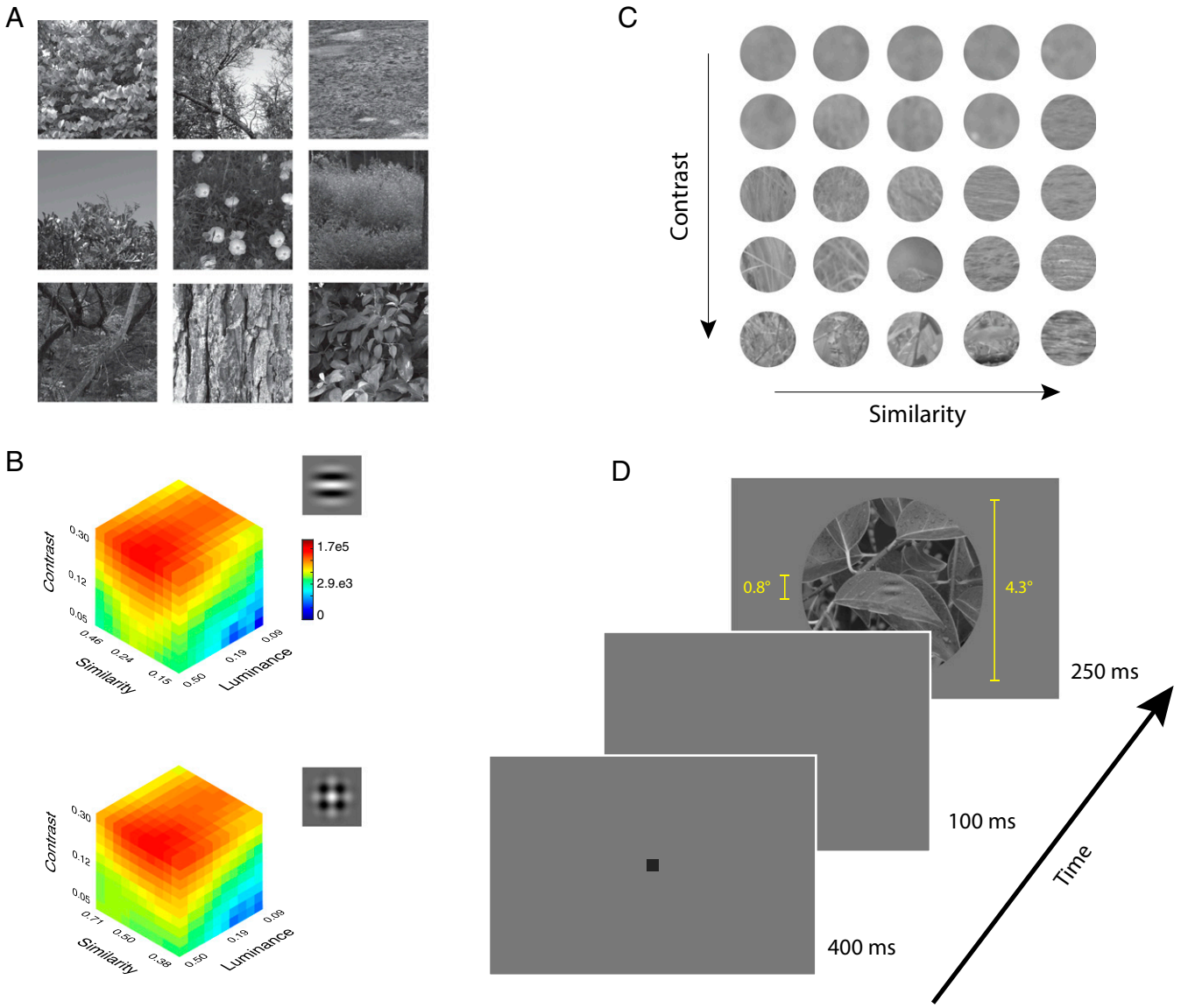

Fig. 2. Stimuli for experiment 1. (*A*) Example 400 × 400 pixel regions from several of the 4,284 × 2,844 full images. (*B*) Three-dimensional histograms along the dimensions of luminance, RMS contrast, and similarity, for the two targets used in the experiments: a windowed 4-cpd grating and a windowed 4-cpd plaid. The similarity depends on the specific target, and, hence, there are two histograms (*Methods*). For each histogram, there are 10 bins along each dimension (bin widths increase geometrically along each dimension), for a total of 1,000 bins. The ranges for each dimension were restricted to those over which it was possible to measure detection thresholds on a standard display screen without clipping (luminance is 0.08 to 0.55 of image maximum, contrast = 0.05 to 0.32 RMS, similarity range for grating target is 0.15 to 0.45, and similarity range for plaid target is 0.24–0.47). The color scale indicates the number of patches falling in the bins. (*C*) Example 101 pixel diameter (0.8°) background patches having the same mean luminance and various contrasts and similarities to the grating target. (*D*) Timeline of a single trial (excluding the response and feedback intervals). The psychometric function for each tested bin was based on at least 350 trials. In the entire experiment, no background patch was presented twice.

and, as will be shown in *Signal Detection Analysis of Detection Under Blocked Conditions*, the differences between experiments 1 and 2 are predicted by a simple model observer with no position uncertainty. Thus, it is likely that the primary effect of removing the surrounding context region was to remove some of the background's power from under the target.

In experiment 3, we consider conditions where the background bin and target amplitude randomly varied from trial to trial. First, however, we consider potential explanations for the results from the blocked conditions (experiments 1 and 2) based on the statistical properties of natural backgrounds.

Signal Detection Analysis of Detection Under Blocked Conditions. An obvious question is why thresholds should vary approximately linearly along all three dimensions. To gain some insight into this fact, we evaluated a simple signal detection model known as a matched-template (MT) observer (Fig. 5*A*). On each trial, the MT observer computes the dot product of a template $f(x,y)$ with the input image $I(x,y)$,

$$R = f \cdot I = \sum_{x,y} f(x,y) I(x,y), \quad [1]$$

where the template is the target (with amplitude equal to 1.0) divided by its energy (the energy of the target is the dot product of the target with itself). Note that the dot product is equivalent to computing the response of a receptive field exactly matching the luminance profile of the target. Also note that this template

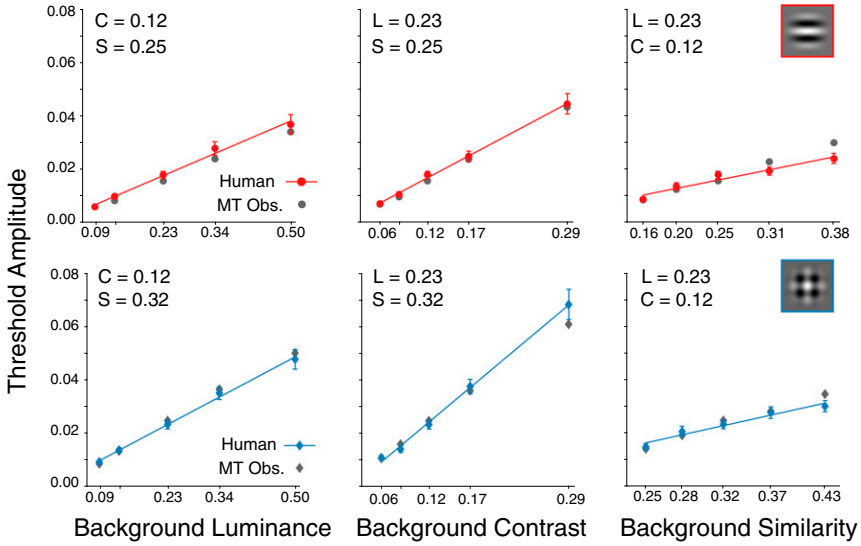

Fig. 3. Experiment 1 results: threshold as a function of background luminance, RMS contrast, and similarity to the target, for (*Upper*) grating and (*Lower*) plaid targets. Thresholds are for the conditions where the background context region is present (Fig. 2*D*). Colored symbols are the mean thresholds for three subjects. Error bars are bootstrapped SDs. The lines are best-fitting linear functions. The black symbols are the predictions of the MT and NMT observer with a single efficiency scale factor (Eq. **4**) applied across all thresholds in experiments 1 and 2 (see also Fig. 4): $\eta = 0.124$.

response is sensitive to the phase of the input, whereas the similarity measure (which is also based on the shape of the target) is insensitive to phase. If the template response R exceeds a decision criterion value γ, then the observer reports that the target is present; otherwise, the observer reports that it is absent. The MT observer is the optimal (ideal) observer when the target is known (as it is here) and the background is Gaussian white noise (7, 27, 28), and for "narrow band" targets, it often is not too far from optimal if the background is Poisson white noise or is correlated Gaussian noise. Although natural image backgrounds have a more complex statistical structure, the MT observer (although not optimal) is a simple, principled signal detection model, and hence a good starting point.

The purple histograms in Fig. 5*B* show the distributions of template responses for the grating target, for all of the background patches in three of the 1,000 bins in Fig. 2*B*. If a target of amplitude a is added to the background, then the distribution remains identical in shape, but is shifted to the right by a (green histograms). The distributions are approximately Gaussian, as indicated by the curves, which are Gaussian distributions having the same means and SDs as the measured distributions. We find that the distributions of template responses are approximately Gaussian for almost all of the bins (Fig. S3; see also ref. 29).

If the goal of the observer is to be as accurate as possible in our experiments, where the target is present on half of the trials, then the decision criterion should be placed halfway between the two distributions ($\gamma = a/2$). In Fig. 5*B*, the target amplitudes have been set so that the accuracy of the MT observer is 85% correct for the three example bins. In our experiments, we define the amplitude threshold a_t to be the amplitude producing 69% correct responses ($d' = a_t/\sigma = 1.0$). In this case, the MT observer's threshold is simply the SD of the template responses for that bin,

$$a_t = \sigma(L, C, S), \qquad [2]$$

where L, C, and S represent the luminance, contrast, and similarity, respectively.

The symbols in Fig. 6 show the measured SDs of the template responses, and hence the MT observer's thresholds, for both targets, for all 2,000 bins. To summarize these image statistics, we fit the following formula, which is separable in luminance, contrast, and similarity:

$$\sigma(L, C, S) = k_0(L + k_L)(C + k_C)(S + k_S), \qquad [3]$$

where k_0, k_L, k_C, and k_S are free parameters (values in Fig. 6 legend). The lines in Fig. 6 show that this formula fits very well (it accounts for 99.6% of the variance in the SDs), implying that

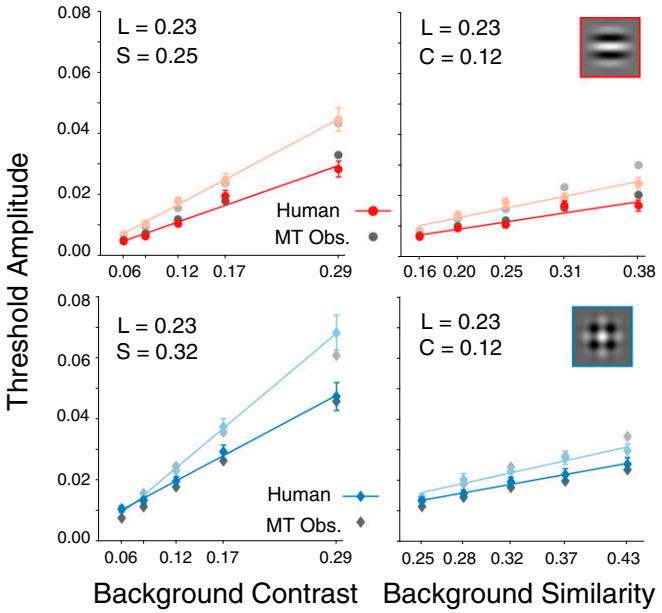

Fig. 4. Experiment 2 results: threshold as a function of RMS contrast and similarity to the target, for grating and plaid targets, with the background context region removed. The colored symbols are the mean thresholds for three subjects. Error bars are bootstrapped SDs. The lines are best-fitting linear functions. The black symbols are the predictions of the MT and NMT observer with a single efficiency scale factor (Eq. **4**) applied across all thresholds in experiments 1 and 2: $\eta = 0.124$. For comparison, the lighter shaded symbols and lines are replotted from Fig. 3, where the background context region is present.

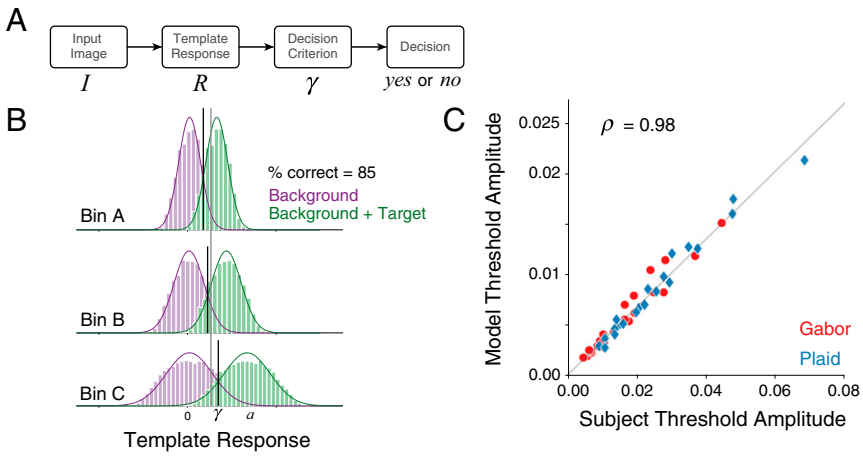

Fig. 5. MT observer. (*A*) The MT observer computes the dot product of the image with a template (receptive field) that matches the luminance profile of the target. If the value of the dot product exceeds a decision criterion the observer responds "yes," that the target is present, and, otherwise, "no." (*B*) Purple histograms show the distribution of template responses for three example bins from the 1000 in Fig. 2*B*. Green histograms show the distributions when the target amplitude is set to produce 85% correct responses. Black vertical lines show optimal criterion placement for each bin. Gray vertical line shows the best single criterion if the stimuli are randomly picked from the three bins on each trial. (*C*) Correlation between the MT observer's and human observers' thresholds for all thresholds in experiments 1 and 2. The symbols show the MT observer's and human observers' threshold for each condition in Figs. 3 and 4. The thin line is the best-fitting line through the origin.

the thresholds of the MT observer are essentially linear in all three dimensions—multidimensional Weber's law.

The prediction of linear threshold functions for each dimension is an important result, but do the MT observer's thresholds actually predict the variation in slopes and intercepts across the different background dimensions and targets in experiments 1 and 2 (shown in Figs. 3 and 4)? To address this question, we first note that, for detection in noise, humans never reach the absolute levels of performance of the MT (ideal) observer, and thus the relative performance of human and ideal observers is often compared by introducing an overall efficiency parameter η, which effectively scales up the variance of the MT responses or, equivalently, scales up all of the MT observer's thresholds by a constant (7, 12, 28, 30),

$$a_t = \sigma(L, C, S)/\sqrt{\eta}. \quad [4]$$

The black symbols in Figs. 3 and 4 show the predictions of the MT observer for a single fixed value of the efficiency parameter

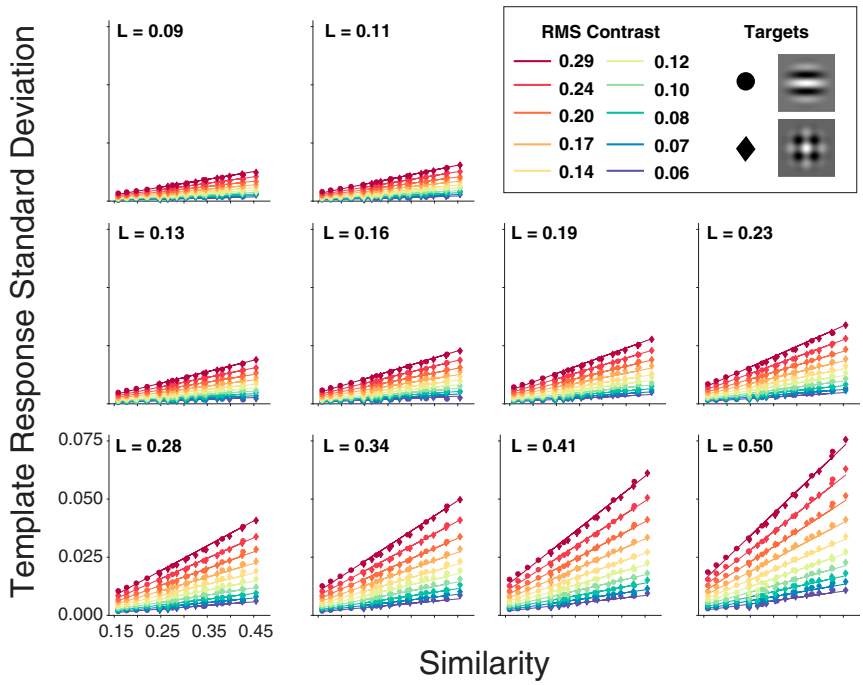

Fig. 6. Template response variability in natural images for the grating and plaid targets. Each symbol shows the SD of the template response (Eq. **1**) for one of the 2,000 bins (1,000 for each target) tiling the space of natural background image patches (Fig. 2*B*). The position of a symbol on the horizontal axis gives the mean similarity of the backgrounds in the bin to the target (units are proportion of the maximum possible similarity). The color of a symbol gives the mean RMS contrast of the backgrounds in the bin. The panel gives the mean luminance of the backgrounds in the bin (units are proportion of maximum luminance in the entire natural image). The solid curves show the fit of Eq. **3**, with the following parameter values: $k_0 = 1.38$, $k_L = 0$, $k_C = -0.0154$, and $k_S = -0.0712$.

($\eta = 0.124$). (To generate the predictions for experiment 2, we measured and used the SDs of the template response for the windowed backgrounds; see Fig. S4.) As can be seen, the values of the slopes and intercepts across the three dimensions, for both targets, and in both experiments, are predicted quite well from the statistics of the template responses to natural backgrounds, with only a single free scaling (efficiency) parameter. Fig. 5*C* shows that the correlation between the predicted (with no efficiency parameter) and observed thresholds for both targets in both experiments was 0.98. These results show, at least for our targets, that the detection mechanisms in the human visual system are tightly matched to the statistical properties of natural scenes.

Experiment 3: Detection Thresholds with Background and Target Amplitude Uncertainty. The third experiment measured the effect on detection performance of randomly varying the background bin and the target amplitude on every trial, because this sort of uncertainty exists under natural conditions. As in experiment 2, we fixed the luminance at the median value (only the contrast and similarity were randomly varying). If we did not fix the luminance bin, then the surrounding uniform region of the display would either provide a nonnatural cue to the luminance (if it varied with the natural background) or it would produce brightness contrast artifacts (if it was kept at a fixed luminance). Using the psychometric functions from experiments 1 and 2, we determined, for each background bin, the target amplitudes corresponding to four specific accuracy levels: 65%, 75%, 85%, and 95%. These target amplitudes were determined separately for each subject. Performance was measured for each of these accuracy levels, with the background bin randomly selected on each trial but the accuracy level blocked. Under these circumstances, both the target amplitude and the background bin varied on each trial, unlike in experiments 1 and 2, where both amplitude and background bin were blocked. Fig. 7*A*, *Left* shows the accuracy in this random condition plotted as a function of the accuracy in the blocked conditions of experiment 1. Fig. 7B, *Left* shows a similar plot for the windowed background conditions of experiment 2. If performance were unaffected by background and target amplitude uncertainty, then the data points should fall along the diagonal (solid black) line. As can be seen, there is little if any effect of uncertainty, even though subjects reported that the background appeared dramatically different from trial to trial.

This is a rather surprising result given the expected effects of uncertainty on target detection. To understand why this is surprising, consider the MT observer for target detection described earlier. On each trial, the observer computes the dot product of the background and template, and, if the dot product exceeds a criterion, then the observer reports that the target is present. Fig. 5*B* shows the distributions of template responses in three specific bins, for a target amplitude producing 85% correct responses. If the background bin and the amplitude of the target are fixed, as in experiments 1 and 2, then the MT observer can maximize accuracy by setting the criterion at the cross point of the two distributions in that bin ($\gamma = a/2$). For bin A, the SD is relatively low, and hence the target amplitude and the criterion that produces 85% correct responses are relatively small (black line). For bins B and C, the SDs are higher, and the target amplitudes and the criterion that produces 85% correct responses are larger. However, when the background bin randomly varies on every trial, as in experiment 3, then no single criterion (gray line) can be in the correct location for all bins. Thus, the maximum accuracy of the MT observer must be lower in the uncertainty conditions. The dashed curves in Fig. 7 *A*, *Left* and *B*, *Left* show the performance of the MT observer, when its decision criterion is set so the MT observer's bias exactly equals the bias estimated from the subjects' hits and false alarms at each accuracy level (see *SI Text* and Fig. S5 for details). The upper limit of MT observer performance in natural images is reported in Fig. S6.

Another prediction of the MT observer is that there must be a strong trade-off in the proportions of hits and correct rejections as a function of the bin SD. The proportion of hits is the area under the green distribution to the right of the criterion; the proportion of correct rejections is the area under the purple distribution to the left of the criterion. As can be seen in Fig. 5*B*, the proportion of hits must increase as the bin SD increases, and the proportion of correct rejections must decrease. The dashed curves in Fig. 7 *A*, *Right* and *B*, *Right* show the dramatic trade-off in the proportion of hits and correct rejections predicted by the MT observer, as a function of the bin SD. The subjects' proportions (symbols) show that the measured trade-off is much smaller, especially when the surround context is present (Fig. 7*A*), which is the case under real-world conditions.

How are the human visual and decision-making systems able to maintain sensitivity, and relatively constant hit and correct rejection rates, under conditions of background and target-amplitude uncertainty? One possibility is that they effectively normalize the template response by subtracting the mean and dividing by the SD implied by the estimated luminance, contrast, and similarity (Fig. 7*C*). The effect of properly normalizing the template responses is illustrated in Fig. 7*D*, which shows the distributions in Fig. 5*B* after normalization. In the normalized space, the SDs all become 1.0, and separation between the distributions becomes the detectability d'. In this case, the same accuracy level in the blocked and random conditions can be obtained with a single fixed criterion for each accuracy level (which was blocked). Furthermore, with this fixed criterion, the proportion of hits and correct rejections will not trade off as a function of bin SD.

Recall that the SD of the template response is a separable product of the luminance, contrast, and similarity (Eq. **3**). Also, the grating and plaid targets integrate to zero, and hence the target-absent distributions have a mean of zero. Thus, the responses of the normalized MT (NMT) observer are given by

$$Z = \frac{R}{k_0\left(\hat{L}+k_L\right)\left(\hat{C}+k_C\right)\left(\hat{S}+k_S\right)}, \qquad [5]$$

where $\hat{L}$, $\hat{C}$, and $\hat{S}$ are the estimated luminance, contrast, and similarity in the target region. These three properties of the background might be estimated from the background region surrounding the target region; this is plausible because the statistical properties of natural images are spatially correlated (e.g., nearby locations have similar contrasts). It is also possible that these properties could be estimated in the target region; however, this might be more difficult because the estimates would be corrupted by the properties of the target, on the target-present trials.

To evaluate this hypothesis, we determined, separately, how well luminance, contrast, and similarity could be estimated by a simple linear model that combines measurements of the property in both the target and surrounding regions (see *SI Text* and Fig. S7). Such a linear weighting of local measurements is a plausible neural computation. We find that the linear model estimates are sufficiently accurate that Eq. **5**, with a fixed criterion for each accuracy level, can account for human performance quite well. Specifically, the solid black symbols (and black solid line) in Fig. 7 *A*, *Left* and *B*, *Left* show the predictions of the NMT observer. Also, for the surround context conditions (Fig. 7*A*, *Right*), the NMT observer does a much better job than the simple MT observer in accounting for the hit and correct rejection rates (Fig. 7*A*, *Right*, *Inset* and solid curves). Interestingly, for the windowed conditions (Fig. 7*B*, *Right*), which are less natural, there is a greater trade-off in hits and correct rejections. In this case, the predictions of the two model observers are equally good (Fig. 7B, *Right, Inset*), which would be predicted by partial or incomplete normalization. This incomplete normalization is consistent with evidence that the contrast normalization component of receptive

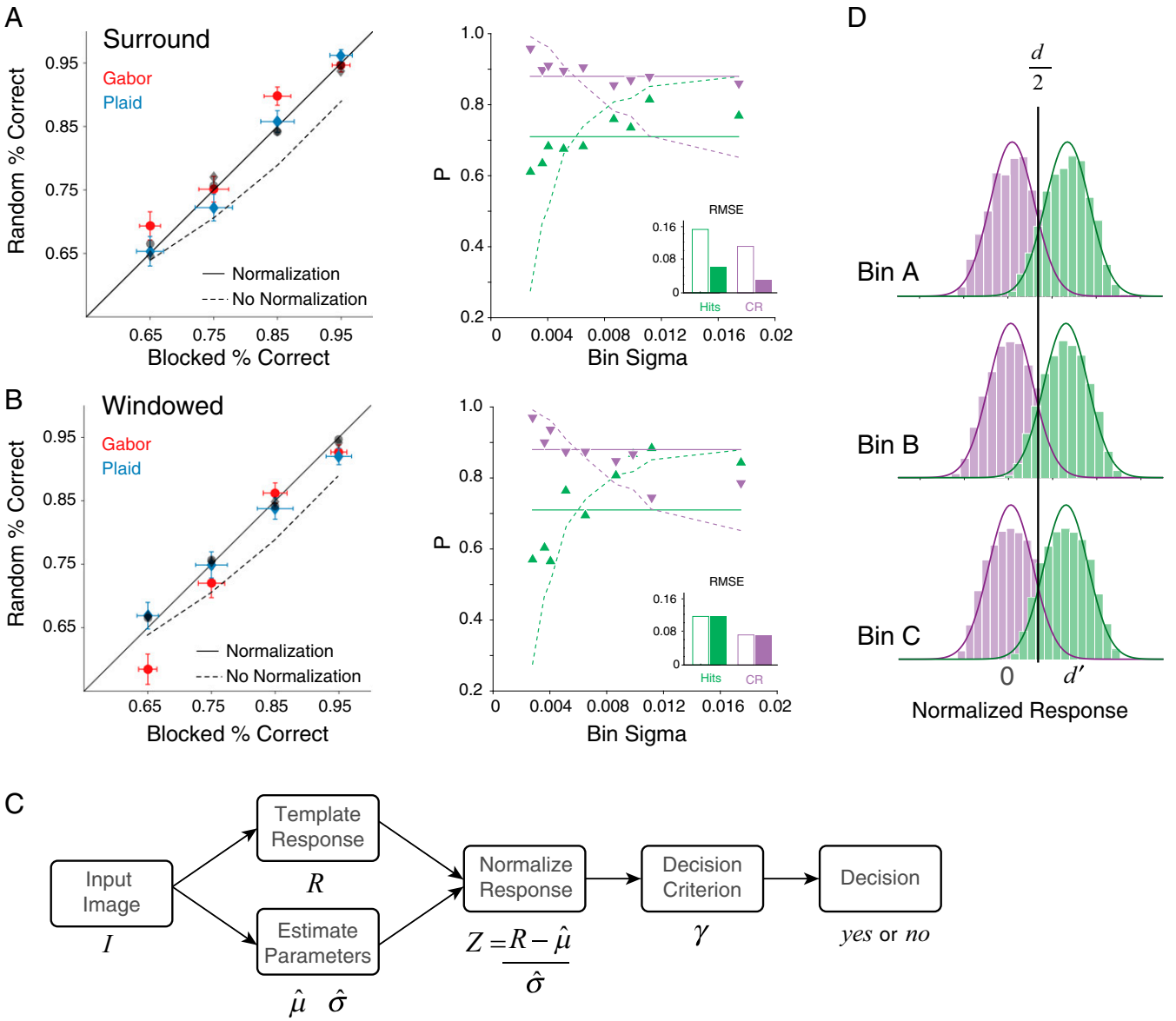

Fig. 7. Comparison of detection performance for blocked and random conditions. (*A*) (*Left*) Percent correct detection in the random conditions as a function of the percent correct in the blocked conditions, when the surrounding background context is present. (*Right*) Proportion of hits and correct rejections as function of the template response SD from the nine randomly selected background bins. (*Inset*) The root mean squared error for the MT observer (open bars) and the NMT observer (solid bars). The dashed curves show the performance of the MT observer. (*B*) Similar plot for conditions where the background has been windowed to the size of the target region. (*C*) NMT observer. The decision variable is the template response normalized by subtracting the estimated mean template response and divided by the estimated SD of the template response. (*D*) The effect of normalization on the example distributions in Fig. 5*B*. With normalization, a single fixed criterion can achieve optimal performance for any fixed accuracy level in experiment 3. The solid black points and solid lines in *A* and *B* show the performance of the NMT observer.

fields in primary visual cortex extends beyond the linear summation component (31, 32). In the *Supplementary Information*, we show that normalization by all three dimensions is necessary to achieve the same detectability in random and blocked conditions, although normalizing by contrast is the most important (Fig. S6).

Discussion

We used a constrained sampling approach to examine the factors affecting detection of known targets in natural backgrounds. Background patches from a database of calibrated natural images were sorted into a 3D histogram having the dimensions of mean luminance, RMS contrast, and (phase-invariant) similarity to the target. We then measured psychometric functions in single-interval, blocked identification experiments in a subset of bins along the three cardinal dimensions of the space, for a grating and a plaid target. We found that amplitude thresholds increased approximately linearly along all three dimensions, both when the background region extended well beyond the target region (experiment 1) and when the background was restricted to the target region (experiment 2). We then showed that a simple MT observer predicted the entire set of thresholds from both experiments with a single efficiency parameter, whose effect is to scale all of the MT thresholds by a single factor. In experiment 3, we examined the effects of background and amplitude uncertainty by randomly sampling a background on every trial from a randomly selected contrast and similarity bin, where the amplitude of the target also depended on the randomly selected bin. We found that, under these conditions, there was essentially no effect of the uncertainty on accuracy, and

a modest effect of background bin SD on the proportion of hits and correct rejections. These results cannot be explained by the MT observer, but can be explained by an NMT observer that estimates the background luminance, contrast, and similarity in the target region via linear weighted summation of the local measurements of each property in the target region. This NMT observer quantitatively accounts for almost all of the results from all three experiments and shows (for our stimuli) that human thresholds in natural backgrounds are accurately predicted from first principles.

The tight relationship between the NMT observer and human observers must break down at very low contrasts, because the NMT observer will perform perfectly (except for the effect of photon noise) on backgrounds of zero contrast. Thus, viable models of human observers based on the NMT observer would need to include another factor such as an additive constant representing a fixed neural noise. Also, we have not yet measured thresholds over the entire space in Fig. 2*B*; however, in pilot data (not presented here), we have found that human thresholds are roughly separable, as implied by the natural scene statistics in Fig. 6.

Interestingly, it is known that MT observers fail to account for significant aspects of human performance. For example, results from classification image experiments strongly suggest that the human visual system uses image features that deviate from those of the MT observer (33–35). Thus, it is unlikely that the subjects in our experiments are carrying out a computation directly equivalent to applying a matched template. However, for detection in white noise as a function of noise amplitude, it is known that the effect of removing some fraction of pixels in a matched template is to reduce efficiency by a fixed scale factor without affecting the pattern of thresholds. We checked whether this invariance holds for natural images, and found that it holds quite well for the whole space of conditions in Fig. 6 (see *SI Text*). We conclude that the pattern of thresholds observed in Fig. 6 and in experiments 1 and 2 (Figs. 3 and 4) are largely due the statistical structure of natural backgrounds, and would be consistent with a range of biologically plausible models. The results of experiment 3 (Fig. 7) would seem to require that plausible models include normalization mechanisms that are separable in luminance, contrast, and similarity.

Perhaps surprisingly, we found that the histograms of template responses are approximately Gaussian for all bins and both targets. It is well known that, in natural images, the response distributions of oriented Gabor filters (like our grating target template) are highly non-Gaussian, with sharp peaks at zero and heavy tails (36, 37). One hypothesis is that this is due to the higher-order structure (contours, edges, lines, etc.) in natural images. However, the patches in any one of our bins contain such structure. Thus, our results suggest, instead, that the heavy-tailed distributions result from the mixture of SDs from the different bins (a mixture-of-Gaussians model that does not depend on the local phase structure of natural images). Also, if the template responses are normalized by the patch luminance, contrast, and phase-invariant similarity (Fig. 7), then they become Gaussian with the same variance, for all bins. To the extent that cortical neuron responses are consistent with such normalization, they will not provide a sparse code in the sense of producing a heavy-tailed distribution of responses to natural images (38).

It is also worth noting that any "neuron" having a linear receptive field can be regarded as a matched template. Hence the SDs of the responses to natural stimuli of neurons having a narrow-band linear receptive field will be the same or very similar to those in Fig. 6. If one of the goals of single neurons in the visual cortex is to identify the presence of features that match their receptive fields (under the real-world conditions of high uncertainty), then there would be a benefit from normalization by background luminance, background contrast, and phase-invariant similarity of the background to the shape of the receptive field (Eq. **5**).

Detection in the Real World. In experiments 1 and 2, stimulus uncertainty was minimized by blocking both the target amplitude and the bin from which the background was sampled. In experiment 3, uncertainty was increased, but was still constrained by blocking trials to a fixed level of accuracy. Under these circumstances, target amplitude and background properties covary in a way that allows the NMT observer to adopt a single optimal fixed criterion for the block. However, in the real world, there is generally no reason to expect the amplitude and the background properties to covary. Nonetheless, the NMT observer supports a simple optimal decision strategy. Under conditions where amplitude is unconstrained, a rational strategy (cost function) is to maximize hit rate for a given desired false-alarm rate—the same strategy used in standard one-tailed statistical tests. For the NMT observer, this cost function corresponds to placing the criterion at a fixed value. For example, a criterion of 1.65 gives a false-alarm rate of 5% and the optimal hit rate, independent of target amplitude and background properties. There is no such fixed criterion for the MT observer. The NMT observer performs much better than the MT observer when the desired false-alarm rate is low (Fig. S8).

The targets in the present experiments were added to the background, and hence the background is at least partially visible through the target. Such transparency occurs in natural scenes, but more common are target objects that occlude the background under them. There are important differences between detection with additive and occluding targets, but the basic principles are the same. For the target-absent trials, the NMT responses will still be approximately Gaussian, with an SD of 1.0.

Here we only considered detection with background uncertainty and target amplitude uncertainty; however, the NMT observer is also appropriate for other forms of target uncertainty, such as location (25, 26), orientation, and spatial frequency (39) uncertainty. In these cases, the normalized matched template would be applied over the region of uncertainty, and the decision criterion is applied to the maximum of the normalized template responses. These kinds of uncertainty are different from amplitude uncertainty because the template would need to be applied to different locations or varied in orientation or shape. Also, unlike amplitude uncertainty, these forms of target uncertainty usually cause a substantial unavoidable decrease in accuracy (12, 25, 26, 28, 30).

The Optimality of Weber's Law for Luminance, Contrast, and Similarity. The classic effects of masking—increases in threshold with background luminance, contrast, and similarity to the target—were primarily discovered and then explored using simple backgrounds that did not randomly vary from trial to trial (4, 5, 8). Furthermore, the effects observed with these nonrandom backgrounds are similar to those we report here. On the surface, this fact seems puzzling. An MT observer, for backgrounds that do not vary from trial to trial, will always perform perfectly, independent of background luminance, contrast, or similarity, because the template response has no variability except that due to the target. So, why should there be a close relationship between the thresholds obtained with random backgrounds and those obtained with fixed backgrounds?

The explanation most likely lies in the fact that the visual system evolved to operate under conditions of high stimulus uncertainty. Under natural conditions, both the background and the amplitude of the target (if present) are generally different on every occasion. What the present scene statistics measurements and modeling show is that the detrimental effects of this uncertainty can be optimally reduced by dividing the template response by the product of background luminance, contrast, and similarity (Eq. **5** and Figs. S6 and S8); this is just the sort of normalization (gain control) observed early in the visual system for the dimensions of luminance and contrast (40–44). Because the visual system is almost always performing detection under uncertainty, it is reasonable to expect evolution to place the

adjustments for this uncertainty into the early, automatic levels of the visual system. However, the side effect of this is that, under laboratory conditions, where we can fix the background, these gain-control mechanisms lead to highly suboptimal performance—the gain control reduces the signal level relative to subsequent neural processing and decision noise. Undoubtedly, if our ancestors had existed in a simple environment with just a few specific backgrounds, then the visual system would have evolved a very different solution (e.g., estimating which of the few possible backgrounds is present and then subtracting it from the input). We argue that the rapid and local neural gain-control mechanisms, and the psychophysical laws of masking, are most likely the result of evolving a near-optimal solution to detection in natural backgrounds under conditions of high uncertainty.

A standard explanation for early gain-control mechanisms is that they keep the responses of the neurons encoding the stimulus within the neurons' dynamic range. This explanation must be true for the slow changes in gain that occur with changes in ambient light level, for the same reason that cameras adjust their gain based on ambient light level, namely, the ambient light level typically varies slowly over 10 orders of magnitude. These mechanisms are not included in the modeling and analysis presented here, because the local luminance changes within a given natural image (and in our experiments) are modest. The gain-control mechanisms that operate under these conditions adjust rapidly to the local luminance and contrast (40–44), and perhaps similarity (32, 45). Indeed, to be useful, they must adjust nearly instantly (within a few tens of milliseconds), because the eyes are in constant motion and local image statistics are largely uncorrelated across fixations (42, 46). Our argument is that it is these rapid gain-control mechanisms that are optimal for detection when fixating around a given natural scene under conditions of high uncertainty. This is not to say that there is not also a synergistic benefit of rapid gain control for keeping signals within the dynamic range of neurons; for example, Fig. 6 shows that the dynamic range of template responses within a typical image is nearly three orders of magnitude.

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

Methods

The scene statistics were computed, and stimuli obtained, from a large collection of calibrated natural images (4,284 × 2,844 pixels) that are 14 bits per color and linear in luminance (the images and camera calibration procedure are available at natural-scenes.cps.utexas.edu). The RGB images were converted to gray scale by converting to XYZ space and then taking the Y (luminance) values. They were then clipped to the top 1% and normalized by the maximum luminance.

To measure the scene statistics, the images were divided into 101 × 101 pixel patches, which was the size of the targets in the experiments, and then sorted into 3D histograms, with 10 bins along each dimension. Briefly, the three stimulus dimensions were defined as follows (for more details, see *SI Text*). The two target stimuli were a 4-cpd cosine grating and 4-cpd plaid windowed with a radial raised-cosine function having a width of 101 pixels. A mean luminance image was obtained by convolving the image with the raised-cosine window. The mean luminance L of a patch was defined as the value of the mean luminance image at the center of the patch. A contrast image was obtained by subtracting the mean luminance image from the image, and then dividing the result by the mean luminance image. The RMS contrast C of a patch was defined as the square root of the dot product of the square of the contrast image with the raised-cosine window centered on the patch. The phase-invariant similarity S was defined as the cosine of the angle between the Fourier amplitude spectrum of the patch (minus its mean) and the Fourier amplitude spectrum of the target, where the two spectra are regarded as vectors.

To generate the stimuli, each 14-bit natural gray-scale image was normalized to a maximum of 255. On each trial, a background patch was randomly sampled from the bin for that trial. On trials where the surrounding context region was presented, the context region was included. On target-present trials, the target was added. The resulting image was then gamma-compressed, based on the calibration of the display device (GDM-FW900; Sony), quantized to 256 gray levels from a 10-bit pallet (maximum gray level = 97 cd/m^2), and displayed at a resolution of 120 pixels per degree.

Stimulus presentation and response collection were programmed in MatLab, using PsychToolbox (47, 48). In experiments 1 and 2, psychometric functions were measured for several bins in each experimental session. Each psychometric function was measured twice on each of the three subjects; the second measurement was taken after all of the psychometric functions had been measured once. Each psychometric function was measured in a single-interval, blocked identification task with feedback. There were five blocks, where each block consisted of 36 trials, with the target amplitude fixed at a particular value. To help the subject adopt the appropriate decision criterion, the first trial in a block always contained the target (the first trial was not included in the data analysis). For each subject, all of the psychometric data for each bin (350 trials) were fitted with a generalized cumulative Gaussian function using a maximum-likelihood procedure (see *SI Text*). Threshold was defined to be the target amplitude corresponding to 69% correct responses ($d' = 1.0$).

For plotting and modeling, the 14-bit gray-scale images were normalized to a maximum of 1.0, and the target at maximum amplitude was normalized to a peak of 1.0. Thus, when the target is present, the stimulus image is given by $I(x,y) = B(x,y) + aT(x,y)$, with amplitude $a < 1$, and the template response is given by $R = B \cdot f + a$ (Eq. **1**).

On each trial in experiment 3, one of the nine contrast and similarity bins was randomly selected, and then a background patch was randomly sampled from those background patches that were sampled from that bin in experiments 1 and 2. In each 50-trial block, the amplitude of the target was set to give a particular accuracy, based on the specific subject's psychometric functions measured in experiments 1 and 2. There were four different blocks (65%, 75%, 85%, and 95%). Each block was repeated three times for a total of 150 trials per accuracy level for each subject.

The experimental protocols for this study were approved by the University of Texas Institutional Review Board, and informed consent forms were obtained from all participants.

ACKNOWLEDGMENTS. We thank Dennis McFadden and the reviewers for helpful comments. This work was supported by National Institutes of Health Grants EY024662 and EY11747.

1. Geisler WS, Diehl RL (2003) A Bayesian approach to the evolution of perceptual and cognitive systems. *Cogn Sci* 27:379–402.
2. Geisler WS (2008) Visual perception and the statistical properties of natural scenes. *Annu Rev Psychol* 59:167–192.
3. König A, Brodhun E (1889) *Experimentelle Untersuchungen über die psycho-physische Fundamentalformel in Beug auf den Gesichtssinn* (Preuss Akad Wiss, Berlin).
4. Mueller CG (1951) Frequency of seeing functions for intensity discrimination of various levels of adapting intensity. *J Gen Physiol* 34:463–474.
5. Nachmias J, Sansbury RV (1974) Letter: Grating contrast: Discrimination may be better than detection. *Vision Res* 14:1039–1042.
6. Legge GE, Foley JM (1980) Contrast masking in human vision. *J Opt Soc Am* 70:1458–1471.
7. Burgess AE, Wagner RF, Jennings RJ, Barlow HB (1981) Efficiency of human visual signal discrimination. *Science* 214:93–94.
8. Campbell FW, Kulikowski JJ (1966) Orientational selectivity of the human visual system. *J Physiol* 187:437–445.
9. Stromeyer CF, 3rd, Julesz B (1972) Spatial-frequency masking in vision: Critical bands and spread of masking. *J Opt Soc Am* 62:1221–1232.
10. Wilson HR, McFarlane DK, Phillips GC (1983) Spatial frequency tuning of orientation selective units estimated by oblique masking. *Vision Res* 23:873–882.
11. Watson AB, Solomon JA (1997) Model of visual contrast gain control and pattern masking. *J Opt Soc Am A Opt Image Sci Vis* 14:2379–2391.
12. Burgess AE (2011) Visual perception studies and observer models in medical imaging. *Semin Nucl Med* 41:419–436.

13. Tanner WP, Jr (1961) Physiological implications of psychophysical data. *Ann N Y Acad Sci* 89:752–765.
14. Pelli DG (1985) Uncertainty explains many aspects of visual contrast detection and discrimination. *J Opt Soc Am A* 2:1508–1532.
15. Eckstein MP, Ahumada AJ, Jr, Watson AB (1997) Visual signal detection in structured backgrounds. II. Effects of contrast gain control, background variations, and white noise. *J Opt Soc Am A Opt Image Sci Vis* 14:2406–2419.
16. Caelli T, Moraglia G (1986) On the detection of signals embedded in natural scenes. *Percept Psychophys* 39:87–95.
17. Rohaly AM, Ahumada AJ, Jr, Watson AB (1997) Object detection in natural backgrounds predicted by discrimination performance and models. *Vision Res* 37:3225–3235.
18. Nadenau MJ, Reichel J, Kunt M (2002) Performance comparison of masking models based on a new psychovisual test method with natural scenery stimuli. *Signal Process Image Commun* 17:807–823.
19. Winkler S, Susstrünk S (2004) Visibility of noise in natural images. *Proc SPIE* 5292: 121–129.
20. Chandler DM, Gaubatz MD, Hemami SS (2009) A patch-based structural masking model with an application to compression. *J Image Video Process* 5:649316.
21. Wallis TSA, Bex PJ (2012) Image correlates of crowding in natural scenes. *J Vis* 12(7):6.
22. Bradley C, Abrams J, Geisler WS (2014) Retina-V1 model of detectability across the visual field. *J Vis* 14(12):22.
23. Alam MM, Vilankar KP, Field DJ, Chandler DM (2014) Local masking in natural images: A database and analysis. *J Vis* 14(8):22.
24. Najemnik J, Geisler WS (2005) Optimal eye movement strategies in visual search. *Nature* 434:387–391.
25. Burgess AE, Ghandeharian H (1984) Visual signal detection. II. Signal-location identification. *J Opt Soc Am A* 1:906–910.
26. Swensson RG, Judy PF (1981) Detection of noisy visual targets: Models for the effects of spatial uncertainty and signal-to-noise ratio. *Percept Psychophys* 29:521–534.
27. Peterson WW, Birdsall TG, Fox WC (1954) The theory of signal detectability. *Trans IRE Prof Group Info Theory* 4:171–212.
28. Green DM, Swets JA (1966) *Signal Detection Theory and Psychophysics* (Wiley, New York).
29. Burge J, Geisler WS (2015) Optimal speed estimation in natural image movies predicts human performance. *Nat Commun* 6:7900.
30. Geisler WS (2011) Contributions of ideal observer theory to vision research. *Vision Res* 51:771–781.
31. Cavanaugh JR, Bair W, Movshon JA (2002) Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *J Neurophysiol* 88:2530–2546.
32. Cavanaugh JR, Bair W, Movshon JA (2002) Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons. *J Neurophysiol* 88: 2547–2556.
33. Murray RF, Bennett PJ, Sekuler AB (2005) Classification images predict absolute efficiency. *J Vis* 5:139–149.
34. Eckstein MP, Beutter BR, Pham BT, Shimozaki SS, Stone LS (2007) Similar neural representations of the target for saccades and perception during search. *J Neurosci* 27:1266–1270.
35. Zhang S, Abbey CK, Eckstein MP (2009) Virtual evolution for visual search in natural images results in behavioral receptive fields with inhibitory surrounds. *Vis Neurosci* 26:93–108.
36. Field DJ (1987) Relations between the statistics of natural images and the response properties of cortical cells. *J Opt Soc Am A* 4:2379–2394.
37. Daugman JG (1989) Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. *IEEE Trans Biomed Eng* 36:107–114.
38. Olshausen BA, Field DJ (1997) Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Res* 37:3311–3325.
39. Davis ET, Graham N (1981) Spatial frequency uncertainty effects in the detection of sinusoidal gratings. *Vision Res* 21:705–712.
40. Albrecht DG, Geisler WS (1991) Motion selectivity and the contrast-response function of simple cells in the visual cortex. *Vis Neurosci* 7:531–546.
41. Heeger DJ (1991) Nonlinear model of neural responses in cat visual cortex. *Computational Models of Visual Perception*, eds Landy MS, Movshon JA (MIT Press, Cambridge, MA), pp 119–133.
42. Mante V, Frazor RA, Bonin V, Geisler WS, Carandini M (2005) Independence of luminance and contrast in natural scenes and in the early visual system. *Nat Neurosci* 8: 1690–1697.
43. Carandini M, Heeger DJ (2011) Normalization as a canonical neural computation. *Nat Rev Neurosci* 13:51–62.
44. Hood DC (1998) Lower-level visual processing and models of light adaptation. *Annu Rev Psychol* 49:503–535.
45. Coen-Cagli R, Kohn A, Schwartz O (2015) Flexible gating of contextual influences in natural vision. *Nat Neurosci* 18:1648–1655.
46. Frazor RA, Geisler WS (2006) Local luminance and contrast in natural images. *Vision Res* 46:1585–1598.
47. Brainard DH (1997) The Psychophysics Toolbox. *Spat Vis* 10:433–436.
48. Pelli DG (1997) The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spat Vis* 10:437–442.
49. Kersten D, Mamassian P, Yuille A (2004) Object perception as Bayesian inference. *Annu Rev Psychol* 55:271–304.

Supporting Information

Sebastian et al. 10.1073/pnas.1619487114

SI Text

Fitting Psychometric Functions. The psychometric functions were fitted with a generalized cumulative normal distribution function. The equations for hit and false-alarm rates were given by the following two equations:

$$P_h(a) = \Phi\left[\frac{1}{2}\left(\frac{a}{a_t}\right)^{\beta} - \gamma_0\right] \quad [S1]$$

$$P_{fa}(a) = \Phi\left[-\frac{1}{2}\left(\frac{a}{a_t}\right)^{\beta} - \gamma_0\right], \quad [S2]$$

where a is the amplitude of the target, a_t is the threshold amplitude, β is the steepness parameter, γ_0 is the bias parameter, and $\Phi(\cdot)$ is the standard normal integral function. To estimate the parameters, we maximized the likelihood function,

$$\begin{aligned}&\ln L(a_t, \beta, \gamma_0)\\ &= \sum_{i=1}^{n} N_h(a_i)\ln P_h(a_i) + N_m(a_i)\ln[1 - P_h(a_i)] + N_{fa}(a_i)\ln P_{fa}(a_i)\\ &\quad + N_{cr}(a_i)\ln\left[1 - P_{fa}(a_i)\right],\end{aligned} \quad [S3]$$

where $N_h(a_i)$, $N_m(a_i)$, $N_{fa}(a_i)$, and $N_{cr}(a_i)$ are number of hits, misses, false alarms, and correct rejections, respectively, for target amplitude a_i. We found that the bias parameter was nearly zero in all cases, so it was set to zero in the final estimates of the thresholds. We note that, although the bias parameter was zero when estimated from the whole hit and false-alarm psychometric functions, it did vary somewhat with the accuracy, which was taken into account in analyzing experiment 3 (see *Predictions for Experiment 3*). We also note that the value of the threshold is independent of β, but that β is higher than that of the matched template observer, 1.0.

Definitions of Dimensions. A local mean luminance image was obtained for each calibrated natural image by convolving the image with a 2D raised-cosine function (Hanning window) normalized to a volume of 1.0,

$$\bar{I}(x,y) = w(x,y) * I(x,y), \quad [S4]$$

where

$$w(x,y) = \frac{W(x,y)}{\sum_{x,y} W(x,y)} \quad [S5]$$

$$W(x,y) = \begin{cases} 0.5 + 0.5\cos\left(\pi\sqrt{x^2+y^2}\,/\,\rho\right) & \sqrt{x^2+y^2} < \rho \\ 0 & \sqrt{x^2+y^2} \geq \rho \end{cases}. \quad [S6]$$

The radius of the raised cosine ρ was equal to the radius of the patch (50 pixels). The mean luminance L of a patch was defined as the value of the local mean luminance image at the center of the patch.

A contrast image was obtained by subtracting the local mean luminance image from the image and then dividing by the local mean luminance image,

$$c(x,y) = \frac{I(x,y) - \bar{I}(x,y)}{\bar{I}(x,y)}. \quad [S7]$$

The contrast of a patch was defined to be square root of the dot product of the square of the contrast image and the 2D raised cosine centered on the patch (see Eq. **1** for the definition of the dot product),

$$C = \sqrt{w \cdot c^2}. \quad [S8]$$

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

$$S = \frac{A_T \cdot A_I}{\|A_T\|\|A_I\|}. \quad [S9]$$

The amplitude spectrum of the target was obtained by taking the complex absolute value of the fast Fourier transform (FFT) of the target. The amplitude spectrum of the patch was obtained by taking the complex absolute value of the FFT of the image patch, after subtracting the mean of the patch and then windowing the patch at its boundary by a raised cosine ramp having a width of 10 pixels.

Individual Subject Data. Fig. S1 shows the individual subject data for experiment 1, and Fig. S2 shows the individual subject data for experiment 2.

Kurtosis of Template Response Distributions. Histograms of the excess kurtosis of the matched template responses for the grating and plaid target are given in Fig. S3. The excess kurtosis of a Gaussian distribution is 0.0.

Template Response Variability for Windowed Backgrounds. Fig. S4 shows the MT response SDs for the windowed backgrounds.

Predictions for Experiment 3. In experiment 3, the background bin and target amplitude randomly varied on each trial, where the amplitudes were constrained to correspond to a fixed level of accuracy in experiments 1 and 2. Four fixed levels of accuracy were tested (65%, 75%, 85%, and 95%) for both target types (grating and plaid) and for both surround conditions (with and without). The amplitudes needed for each accuracy level were estimated separately for each subject.

To analyze the data, we first computed, for each accuracy level, the average decision bias of the subjects in the blocked conditions of experiments 1 and 2 and in the random conditions of experiment 3. These bias values were calculated directly from the proportion of hits and proportion of false alarms using the standard signal detection formula,

$$\gamma_0 = -\frac{\Phi^{-1}(p_{hits}) + \Phi^{-1}\left(p_{fa}\right)}{2}, \quad [S10]$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal integral function (note that unbiased corresponds a bias value of zero). These bias values are plotted in Fig. S5. Because of the bias in the blocked conditions (open squares in Fig. S5), the actual d' values corresponding to the fixed accuracy levels were slightly higher than expected given zero bias. For example, the d' value for the 75% correct condition was 1.39 rather than the nominal

1.35. In the signal detection theory framework, $d' = a_i/\sqrt{\eta}\sigma_i$, where a_i is the amplitude of the target in bin i, σ_i is the SD of the template response in bin i, and η is the subjects' efficiency. Thus, the larger d' value effectively scales all of the target amplitudes up by a small factor (this scaling has only a small effect).

We then computed the performance of the MT observer and the NMT observer, where each was constrained to produce the exactly the same bias values as the human subjects in the random conditions (black squares in Fig. S5). The proportion of hits and false alarms of the MT observer are given by the following equations:

$$p_h = 1 - \frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{\gamma - a_i}{\sigma_i}\right) \quad \textbf{[S11]}$$

$$p_{fa} = \frac{1}{n}\sum_{i=1}^{n}\Phi\left(-\frac{\gamma}{\sigma_i}\right), \quad \textbf{[S12]}$$

where n is the number of background bins (nine, in the present case). For each accuracy condition, we varied the criterion γ in Eq. **S11** and **S12** until the bias γ_0 computed with Eq. **S10** matched the human subjects. These criterion values determined the predictions of the MT observer shown in Fig. 7 *A* and *B*. Similarly, for each accuracy level, the criterion value of the NMT observer was varied to match the bias value of the human subjects.

Estimation of Local Background Properties. The performance of the NMT observer depends on how accurately the properties of the background in the target region can be estimated; this is a potentially tricky problem because, on each trial, the observer does not know whether the target is present or absent. If the target is present, it could bias the estimate of the background properties, thereby leading to a reduction in performance. Interestingly, we discovered that a simple linear model is able to estimate the natural background properties with sufficient accuracy that performance is essentially unaffected by the random presence of the targets with different amplitudes. We considered two linear models. The first model takes into account the surrounding background context region and is only appropriate for experiment 1. The second model only considers the background in the target region and can be applied to either experiment 1 or experiment 2. In both cases, we learn a separate linear model for each background property. We trained the model by randomly sampling a large number of backgrounds from the entire space, and, for half the samples, we added a target with contrast randomly sampled from a uniform probability distribution over a large range (0.01 to 0.35).

In the first model, we measured, for each training stimulus, the value of the stimulus property at the target location and at the eight surrounding locations. We also measured the template response in the target region. This gave a vector of 10 numbers for each training stimulus. We then applied linear regression to learn the 10 weights that best predict the ground truth background property value at the target location. Fig. S7 shows the learned weights for each of the three dimensions. As can be seen, the most weight is put on the center (target) location, and the next most is put on the template response. The negative weight on the template response partially discounts stimulus energy that is aligned in phase with the target, and hence is likely to come from the target.

In the second model, we measured the value of the stimulus property at the target location, and we measured the template response. Thus, there were only two weights to learn. As might be expected given the weights in Fig. S7, the estimates of the background properties were of similar accuracy in the two models. Thus, in practice, all information away from the target location can be ignored. For each trial in experiment 3, we used these fixed linear weights to estimate the background luminance, contrast, and similarity in the target region, and then substituted those estimated values into Eq. **5** to obtain the normalized response for that trial. Finally, we applied a single, fixed decision criterion (for each percent-correct condition) to obtain the predicted black points and solid curves in Fig. 7 *A* and *B*.

Robustness of NMT Observer. The NMT observer is able to account for almost all of the data reported here with a single efficiency parameter whose effect is to scale all of the NMT observer's thresholds up by a fixed factor. An important question is, how sensitive are the predictions to the specific assumptions of the NMT observer?

As mentioned in *Discussion*, there is evidence from classification image experiments that humans use image features that deviate from those used by the MT (and NMT) observer (33–35). For detection in white-noise backgrounds, removing features from the optimal matched template simply reduces the overall efficiency without changing the shape of the predicted threshold functions. We ran a few checks of this principle for our natural image backgrounds and found that it appears to hold quite well: The correlation between the predicted thresholds for the MT observer and one with 70% of the template pixels randomly removed was 0.97, and the correlation with a template that was windowed to about half the area was 0.94. Thus, it seems likely that the predictions of the MT and NMT observer are fairly robust to deviations from the matched template. In other words, there are likely to be a number of models that predict the pattern of results in experiments 1 and 2. This finding strongly suggests that this pattern of results is largely due to the statistical properties of natural backgrounds and not the detailed properties of the detection mechanisms.

Another property of the NMT observer is that the normalization involves all three stimulus dimensions: luminance, contrast, and similarity. Fig. S6 shows, for all background bins in Fig. 2*B*, the effect of normalizing separately by all three dimensions, by only luminance and contrast, by each dimension separately, and by no dimensions (the MT observer). For these calculations, we assumed a flat prior (all bins equally likely, as in experiment 3) and that the criterion was placed at the optimal location. As can be seen, all three dimensions provide a benefit, although contrast normalization is the most important.

The no-normalization predictions in Fig. S6 are those of the MT observer with an optimally placed criterion. A more sophisticated model observer that does not use normalization (i.e., does not use estimates of L, C, and S) is a Bayesian observer with knowledge of the SDs for each bin and of the prior over bins. In this case, the observer computes the probability of the observed template response given each possible SD and then integrates (marginalizes) across SD and amplitude (49). Given the flat prior, the decision variable reduces to

$$X = \frac{\sum_{i=1}^{n} p_{T+B}(R|\sigma_i, a_i)}{\sum_{i=1}^{n} p_B(R|\sigma_i, a_i)}, \quad \textbf{[S13]}$$

where $p_{T+B}(R|\sigma_i, a_i)$ is the probability of the template response given a particular bin SD and target amplitude when the target is present, and $p_B(R|\sigma_i, a_i)$ is the probability with background alone. This observer responds that the target is present if this decision variable is greater than 1.0. We simulated this observer and found its performance to be indistinguishable from that of the MT observer (gray curve) shown in Fig. S6. Thus, a standard Bayesian observer without normalization is also inconsistent with the results of experiment 3.

Under natural conditions, both the properties of the background and the amplitude of the target (if present) would be unknown and largely independent from one occasion to the next. Further, the prior probability of a target being present would generally be low. Under such circumstances, a simple and sensible decision rule is to

pick a criterion γ that produces a small desired false-alarm rate (like a one-tailed statistical test). For the MT observer, the false-alarm and hit probabilities are given by

$$p_{fa} = 1 - \sum_{i,j,k} \Phi\left[\frac{\gamma}{k_0(L_i + k_l)(C_j + k_c)(S_k + k_s)}\right] p(L_i, C_j, S_k) \quad \textbf{[S14]}$$

$$p_h = 1 - \sum_{i,j,k} \Phi\left[\frac{\gamma - a}{k_0(L_i + k_l)(C_j + k_c)(S_k + k_s)}\right] p(L_i, C_j, S_k), \quad \textbf{[S15]}$$

where Φ is the standard normal integral function. For the NMT observer, the false-alarm and hit probabilities are given by

$$p_{fa} = \Phi(\gamma) \quad \textbf{[S16]}$$

$$p_h = 1 - \Phi\left[\gamma - a \sum_{i,j,k} \frac{1}{k_0(L_i + k_l)(C_j + k_c)(S_k + k_s)} p(L_i, C_j, S_k)\right]. \quad \textbf{[S17]}$$

Fig. S8 shows proportion of hits as a function of target amplitude for several false-alarm rates, again assuming a flat prior over bins. The blue curves show the proportion of hits for the MT observer, and the orange curves show the proportion of hits for the NMT observer. As can be seen, the NMT observer has a much greater hit rate (i.e., much greater power) than the MT observer, especially when the desired false-alarm rate is low (which is appropriate under real-world conditions where the prior probability of target present is low). These calculations further demonstrate the potential value of normalization by local luminance, contrast, and similarity.

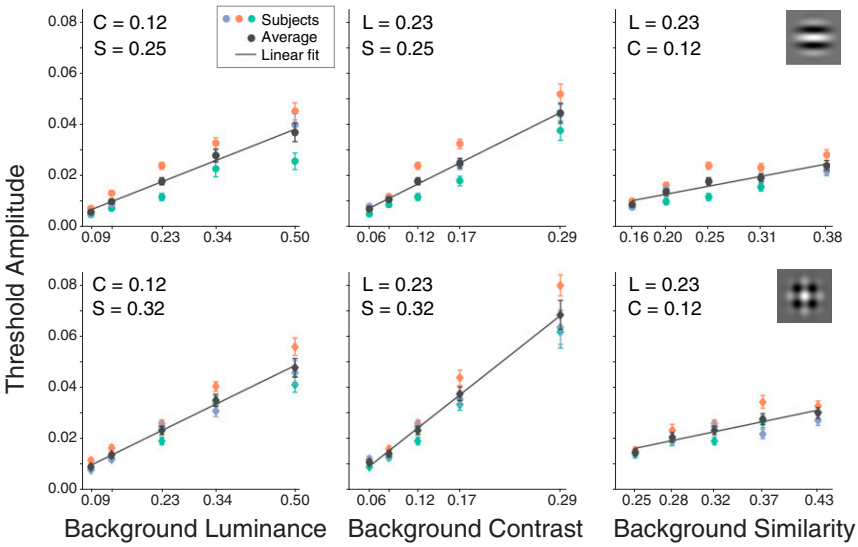

Fig. S1. Individual threshold functions from experiment 1. Colored symbols are thresholds for the different subjects; black symbols are the average. The lines are best-fitting linear functions to the average threshold curves.

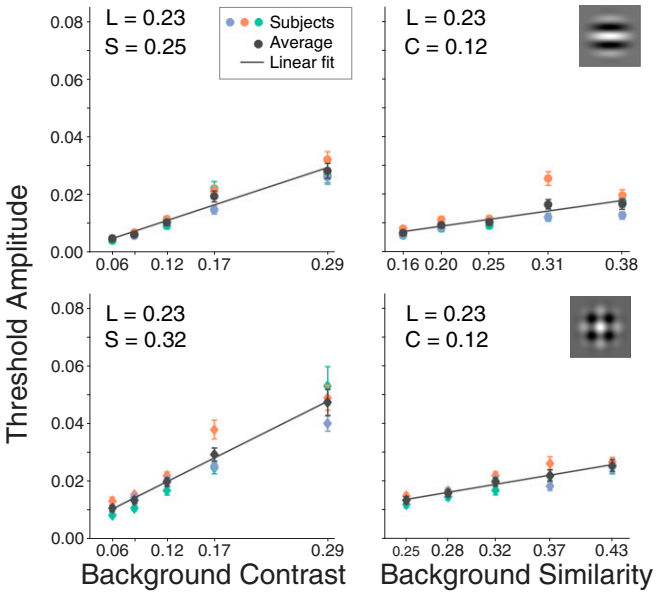

Fig. S2. Individual threshold functions from experiment 2. Colored symbols are thresholds for the different subjects; black symbols are the average. The lines are best-fitting linear functions to the average threshold curves.

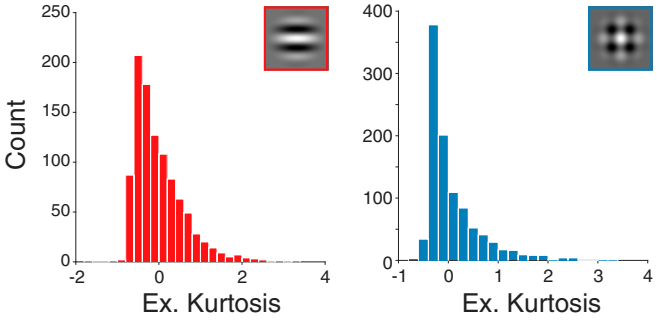

Fig. S3. Histograms of the excess kurtosis of the template response distributions for the grating and plaid targets, for all bins in Fig. 2*B*. The excess kurtosis is 0.0 for a Gaussian distribution. The mean excess kurtosis for the grating target is −0.0095 and, for the plaid target, is 0.0724.

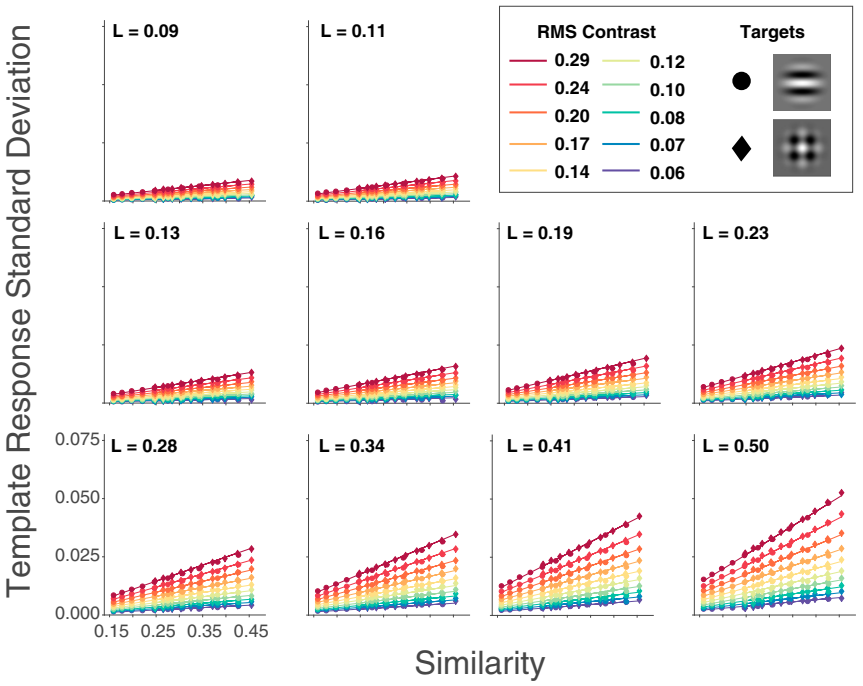

Fig. S4. Template response variability in windowed natural images for the grating and plaid targets. Each symbol shows the SD of the template response (Eq. **1**) for one of the 1,000 bins tiling the space of natural background image patches (Fig. 2*C*). The position of a symbol on the horizontal axis gives the mean similarity of the backgrounds in the bin to the target (units are proportion of the maximum possible similarity). The color of a symbol gives the mean RMS contrast of the backgrounds in the bin. The panel gives the mean luminance of the backgrounds in the bin (units are proportion of maximum luminance in the entire natural image). The solid curves show the predictions of Eq. **3**, with the following parameter values: $k_0 = 0.885$, $k_L = 0$, $k_C = -0.0139$, and $k_S = -0.0397$.

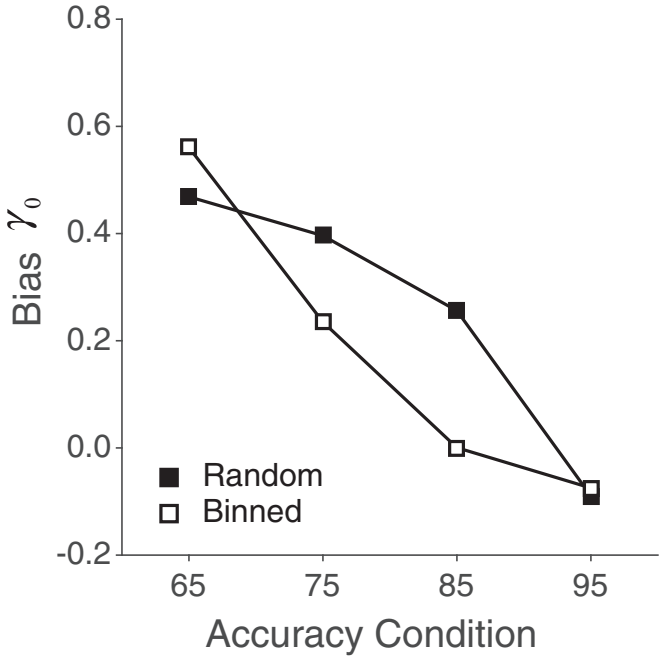

Fig. S5. Average bias value in blocked conditions of experiments 1 and 2, and random conditions of experiment 3 as a function of the measured detection accuracy.

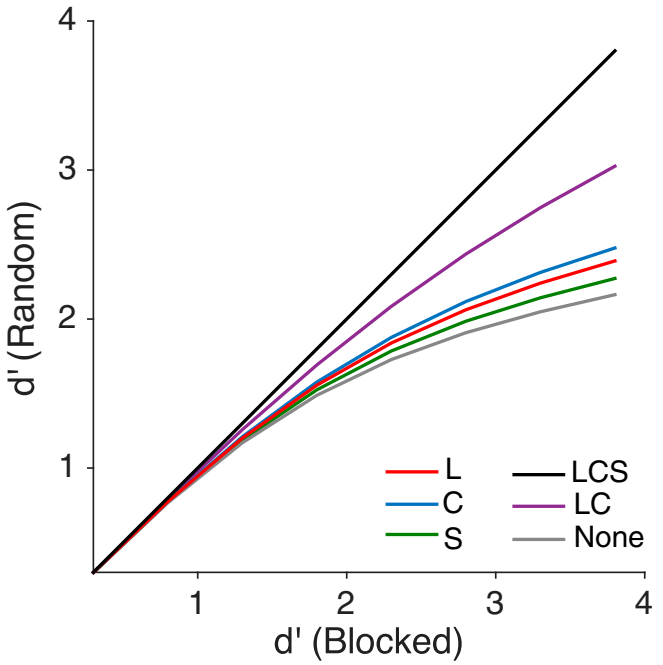

Fig. S6. Contributions of luminance, contrast, and similarity normalization under random background and target amplitude conditions. The gray curve shows the performance of the MT observer, for the space of natural background bins in Fig. 2*B*, with a decision criterion placed optimally for each level of accuracy (here converted to *d'*). The black curve shows the performance of the NMT observer for normalization by all three dimensions (Eq. **5**). The red, blue, and green curves show the performance of the NMT observer for normalization by luminance, contrast, and similarity alone. The purple curve shows the performance for normalization by luminance and contrast (no similarity normalization).

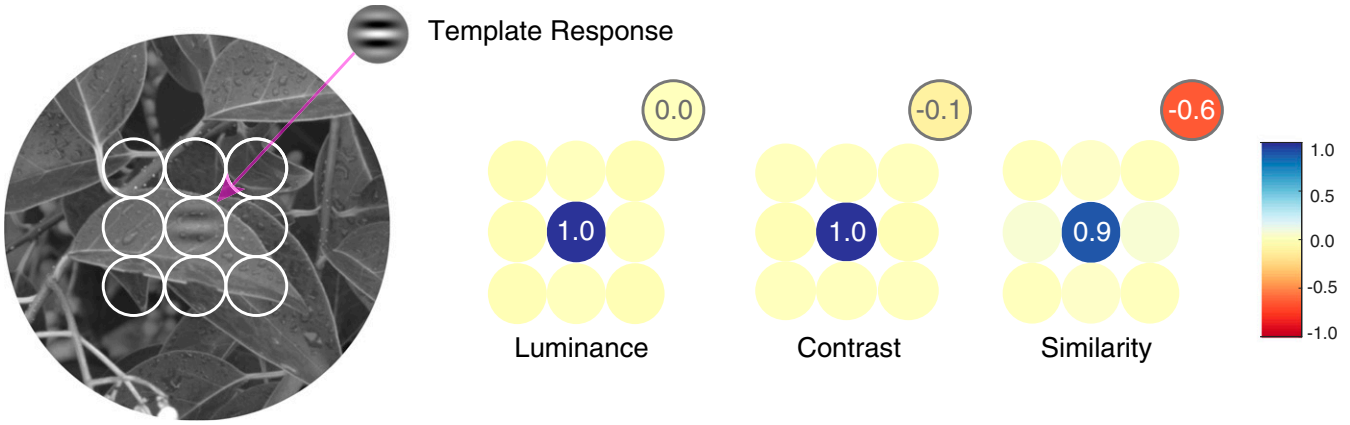

Fig. S7. Linear weights for estimating background luminance, contrast, and similarity in the target region, for the grating target. Linear regression was applied separately for each dimension to learn 10 weights. Nine of the weights were for the dimension value (e.g., similarity) at the target location and eight surrounding locations. The tenth weight was for the template response in the target region. Almost all of the weight is put on information at the location of the target (central location). The same weights were obtained for the plaid target. Using just the two weights for the target location is sufficient to achieve essentially optimal performance under conditions of simultaneous background uncertainty and target amplitude uncertainty.

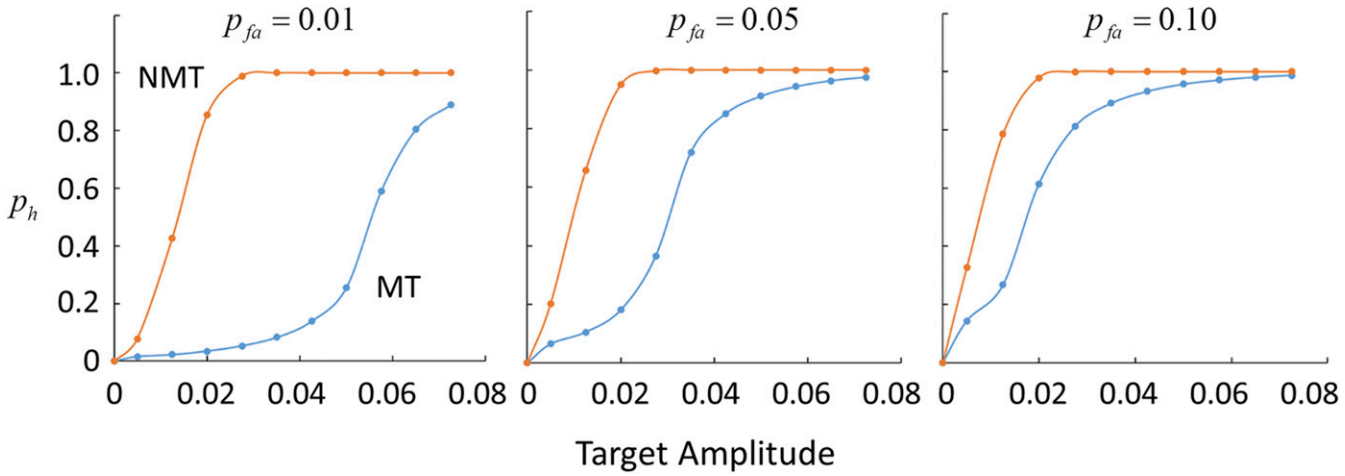

Fig. S8. Comparison of proportion of hits (statistical power) of the MT and NMT observer for a fixed proportion of false alarms (type I error rate), as a function of target amplitude, for several false-alarm rates. The statistical power of the NMT observer is much greater than that of the MT observer, especially when the desired proportion of false alarms is low, which is appropriate under natural conditions where the probability of target present is low.