An image-computable psychophysical spatial vision model

Heiko H. Schütt

Neural Information Processing Group,
University of Tübingen, Tübingen, Germany
Department of Experimental and Biological Psychology,
University of Potsdam, Germany

Felix A. Wichmann

Neural Information Processing Group,
University of Tübingen, Tübingen, Germany
Bernstein Center for Computational Neuroscience,
Tübingen, Germany
Max Planck Institute for Intelligent Systems,
Tübingen, Germany

A large part of classical visual psychophysics was concerned with the fundamental question of how pattern information is initially encoded in the human visual system. From these studies a relatively standard model of early spatial vision emerged, based on spatial frequency and orientation-specific channels followed by an accelerating nonlinearity and divisive normalization: contrast gain-control. Here we implement such a model in an image-computable way, allowing it to take arbitrary luminance images as input. Testing our implementation on classical psychophysical data, we find that it explains contrast detection data including the ModelFest data, contrast discrimination data, and oblique masking data, using a single set of parameters. Leveraging the advantage of an image-computable model, we test our model against a recent dataset using natural images as masks. We find that the model explains these data reasonably well, too. To explain data obtained at different presentation durations, our model requires different parameters to achieve an acceptable fit. In addition, we show that contrast gain-control with the fitted parameters results in a very sparse encoding of luminance information, in line with notions from efficient coding. Translating the standard early spatial vision model to be image-computable resulted in two further insights: First, the nonlinear processing requires a denser sampling of spatial frequency and orientation than optimal coding suggests. Second, the normalization needs to be fairly local in space to fit the data obtained with natural image masks. Finally, our image-computable model can serve as tool in future quantitative analyses: It allows optimized stimuli to be used to test the model and variants of it, with potential applications as an image-quality metric. In addition, it may serve as a building block for models of higher level processing.

Introduction

The initial encoding of visual information by the human visual system has been studied extensively since the seminal studies of the late 1960s and early 1970s (e.g., Blakemore & Campbell, 1969; Campbell & Robson, 1968; Carter & Henning, 1971; Graham & Nachmias, 1971; Nachmias & Sansbury, 1974). Their insights have shaped how we now think about the first computations of the visual system: spatial frequency and orientation specific channels followed by a static nonlinearity. This conceptual model is both broadly consistent with physiology up to primary visual cortex, as well as with normative theories on how the available information should be processed.

As a conceptual framework, the standard model of spatial visual processing is useful and successful. Computational models of it, however, are usually only implemented to work with an abstract representation of visual stimuli, not with "real" images. Typically, the models start with activity in the frequency channels, calculated—or taken—from the parameters of the simple one-dimensional stimuli (e.g., Goris, Putzeys, Wagemans, & Wichmann, 2013; Foley, 1994; Itti, Koch, & Braun, 2000; Legge, Kersten, & Burgess, 1987). This simple implementation of early spatial vision models is highly efficient because first, it bypasses the computational intensive multiscale image

Citation: Schütt, H. H., & Wichmann, F. A. (2017). An image-computable psychophysical spatial vision model. *Journal of Vision, 17*(12):12, 1–35, doi:10.1167/17.12.12.

doi: 10.1167/17.12.12 Received April 13, 2017; published October 20, 2017 ISSN 1534-7362

decomposition and second, it requires few computational units because it is only one-dimensional (1D)—the models are only applicable to (simple) one-dimensional stimuli. Historically, it was the lack of computational power which precluded image-computable models.

Implementing a model to be image-computable, i.e., to work on any image as input, helps to generalize its application to a wide range of tasks and datasets—only image-computable models allow quantitative predictions for any input image (c.f. the discussion of the importance of image-computability by Yamins & DiCarlo, 2016, in the context of convolutional deep neural networks (DNNs) as models of object recognition). Furthermore, image-computable models may reveal—and make it easier to explore—potentially counter-intuitive effects of nonlinearities in one's model. Another benefit is that image-computable models of early spatial vision may be useful beyond spatial vision, because they can be used as psychophysically plausible preprocessors in investigations of higher level processing and for more natural tasks. Finally, image-computable models allow the investigation of statistics of the model *output*, comparing it to normative theories from, e.g., the efficient coding hypothesis (Attneave, 1954; Barlow, 1959; Olshausen & Field, 1996; Schwartz & Simoncelli, 2001; Simoncelli & Olshausen, 2001).

But even for spatial vision, an image-computable model may aid further development: An image based implementation necessarily requires that the model is implemented in full 2D, including orientations and the spatial sizes of filters and normalization pools; they necessitate thinking about spatial vision jointly in the space as well as the spatial-frequency domain. This aspect is likely important for the understanding of visual processing (Daugman, 1980), but is typically not implemented in the abstract, 1D models (Goris et al., 2013).

In this paper we present a psychophysical, image-computable model for early spatial visual processing; we aim to explain human performance in behavioral tasks and thus evaluate our model only on behavioral data from human observers.

History and classical experiments in spatial vision

Psychophysics has a long tradition of quantifying behavior, summarizing it using equations—often called "laws" to mimic physics (Fechner, 1860; Stevens, 1957; Weber, 1834). We have a good quantitative understanding of sensitivity to luminance differences, the dependence of luminance discrimination on wavelength, and the size of test patches (reviewed by Hood, 1998; Hood & Finkelstein, 1986). These early results allow us to convert physical light patterns first into luminance patterns and subsequently into contrast images. The contrast images largely determine detection and discrimination performance (once the display is sufficiently bright).

Arguably, the advent of modern spatial vision came with the discovery of spatial frequency and orientation tuned "channels" (Campbell & Kulikowski, 1966; Campbell & Robson, 1968). Later, the existence of these channels was confirmed by numerous studies, including signal mixture and adaptation experiments (e.g., Blakemore & Campbell, 1969; Graham & Nachmias, 1971). The postulate of independent spatial frequency and orientation channels allows predicting detection thresholds for any signal pattern from the knowledge of the Fourier spectrum of the stimulus and the sensitivity to single sinusoidal gratings of different frequencies, i.e., the contrast sensitivity function.

Because of its pivotal role in the early linear channel model, the contrast sensitivity function was measured under many different conditions, including peripheral presentation (Baldwin, Meese, & Baker, 2012; Rovamo & Virsu, 1979a, 1979b; Virsu & Rovamo, 1979), different luminances (Hahn & Geisler, 1995; Kortum & Geisler, 1995; Rovamo, Luntinen, & Näsänen, 1993, 1994), different temporal conditions (Kelly, 1979; Watson, 1986; Watson & Nachmias, 1977) and different spatial envelopes (Robson & Graham, 1981; Rovamo, Mustonen, & Näsänen, 1994).

Another line of research investigated how the (putative) spatial frequency channel responses are further processed and combined to produce visual behavior. This line of research started with contrast discrimination experiments, measuring the contrast increment needed in addition to a pedestal contrast to produce a detectable difference (Foley & Legge, 1981; Nachmias & Sansbury, 1974). Typically the so-called "dipper function" is found: Low pedestal contrasts facilitate detection; i.e., discrimination can be better than detection, while discrimination requires progressively larger contrast increments for growing pedestal contrast (as to be expected from Weber's law). To explain the shape of the dipper function, Legge and Foley (1980) proposed a Naka-Rushton nonlinearity (Naka & Rushton, 1966) on the spatial frequency channel outputs. Later Foley (1994) revised this model to replace the single-channel nonlinearity with a normalization by the other channel responses to explain oblique masking data, i.e., experiments in which the mask grating had a different orientation than the signal to be detected. This across-channel-normalization is in spirit very close to the typical divisive contrast-gain control introduced to explain the behavior of simple cells in V1 (Cavanaugh, Bair, & Movshon, 2002a; Geisler & Albrecht, 1995; Heeger, 1992).

Finally, the last processing step of (most) models in vision is one of decoding: deriving the open behavioral response from the activity in the model. In older spatial vision models, simple task-independent Minkowski norms were used (the popular "max-rule" or "winner-takes-all-rule"; i.e., the decision is based on the maximally active unit or channel only, corresponds to a Minkowski norm with large—in the limit infinite—exponent). Decoding as an important part of spatial vision models was first discussed by Pelli (1985) in the context of uncertainty. In more modern models, channels are explicitly modelled to respond noisily such that the decoding can be understood in its original statistical meaning of deriving the response from the noisy channel responses. Frequently, this decoding is assumed to be optimal (e.g., Goris et al., 2013; May & Solomon, 2015a, 2015b).

Much of the history of the field, its psychophysical experiments and the purely abstract 1D spatial vision models are summarized and discussed in the comprehensive book of Graham (1989).

There have been earlier attempts to make image-computable models of spatial visual processing, for example by Teo and Heeger (1994) and by Watson and Solomon (1997). However, these earlier models were limited by the available computational power at their time, which required them to tailor their models to the processed stimuli or to limit the possible computations, for example, to entirely local normalization. Recently some more models were implemented to work on images (e.g., Alam, Patil, Hagan, & Chandler, 2015; Bradley, Abrams, & Geisler, 2014). These models usually do not cover the whole complexity, but simplify the normalization steps to reach a computationally more efficient model (Bradley et al., 2014) or are based on entirely different approaches like neural networks trained to predict the detectability of specific distortions (Alam et al., 2015).

One major incentive to develop image-computable models of early visual processing is the applications in image processing. The classical aim here is image quality assessment, i.e., to produce a metric which measures how bad a particular distortion of an arbitrary image is as perceived by humans. Consequentially, the classical models were immediately proposed as such image quality metrics (Teo & Heeger, 1994; Watson, Borthwick, & Taylor, 1997). Such an image quality metric can then be used to optimize various image processing algorithms like compression or tone mapping. This cascade towards application has recently been demonstrated for a different biologically inspired model, the normalized Laplacian pyramid (Laparra, Ballé, Berardino, & Simoncelli, 2016; Laparra, Berardino, Ballé, & Simoncelli, 2017). Our model seems to be a good start for a similar path towards application as it makes valid predictions what distortions are visible to humans and also the optimization of suprathreshold distortions yields reasonable predictions as we shall see below.

Outline

In the following, we first describe how we implemented the spatial vision model to operate on images. We then show that our model reproduces classical psychophysical spatial vision findings, namely those which gave rise to the now accepted model structure in terms of linear filters and divisive normalization. Thereafter, we evaluate our model on a dataset measuring masking by natural images. Then we show that the model produces a sparse representation, as predicted—and desired—from normative considerations. As a final step, we create optimized stimuli to maximize or minimize differentiability according to our model.

Model description

Like most image processing spatial vision models, our model contains four major parts: Images are first *preprocessed*. Then they are *decomposed* into spatial frequency and orientation specific channels and pass an accelerating *nonlinearity and normalization*. Finally, for *decoding*, we assume additive noise and optimal decoding to predict how well images can be differentiated.

Preprocessing

In most psychophysical experiments, stimuli are directly defined in contrast units because the pattern and the contrast together explain most variance, once the stimuli are bright enough. Thus, these stimuli could be passed into our model as they are defined, without any preprocessing.

Nonetheless, we implement the conversion from physical light patterns into the contrast-coded input to our main processing explicitly for two reasons: First, we aim for a model, which can process arbitrary images displayed on a screen and images are usually not given in contrast units (as the example image in Figure 1A). Second, optical effects and retinal processing could be modelled in more detail than we do here. Thus, our simple preprocessing steps mark, where in the model more complex precortical processes fit in and which properties of them we model.

First, we convert all images to luminance values at each pixel. The stimuli used in the classical experiments were already given in luminance values. For modelling

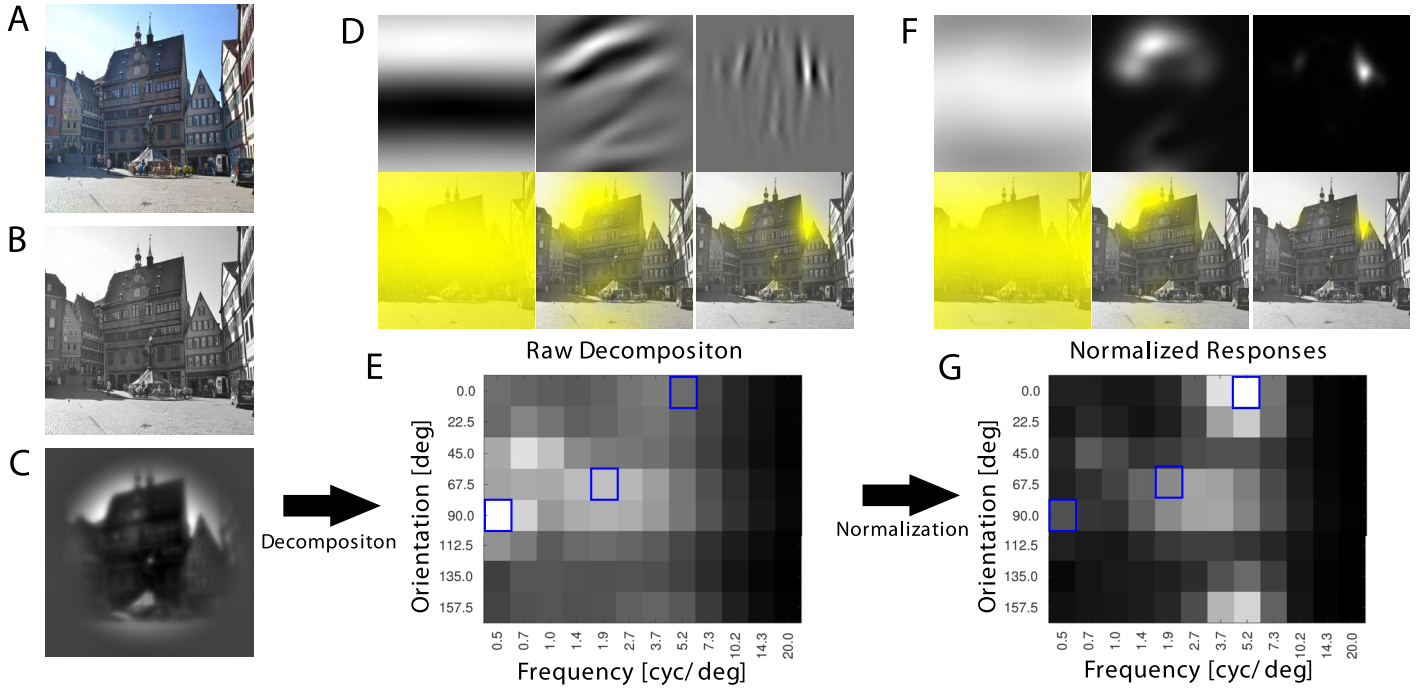

Figure 1. Overview over the model processing. As an example, a photograph of the town hall of Tübingen (A) is passed through the model. (B) Shows the image after conversion to luminance, and (C) shows it after incorporation of eye optics, a hand-tuned contrast sensitivity function and cut-out of the fovea. The image is then decomposed into spatial frequency and orientation channels. The output of these channels for the example image is displayed in (D) and (E). (D) Shows the real part of the output and the absolute value of the output overlaid on the image for three example channels marked in (E). (E) Shows the mean absolute value of each channel overlaid over the original luminance image. Finally, each channel's activity is passed through an accelerating nonlinearity and is normalized by a surrounding normalization pool. The result of this is displayed in (F) and (G). (F) Shows the activity of the same three channels as (D) after normalization, first isolated and then overlaid over the original image. (G) Shows each channel's mean activity over the image after normalization.

the natural image masking database by Alam, Vilankar, Field, and Chandler (2014) as we describe below, we use the pixel value to luminance conversion function as provided with the data. For all other natural images, we used measured spectra from a monitor in our lab (Mitsubishi Diamond Pro 2070) and the V_λ curves as given by Sharpe, Stockman, Jagla, and Jägle (2005) to convert the pixel values to luminance. This monitor (the one we used for the eye movement experiments) we use for the evaluation of our models' responses below. For display in this paper we converted them back to RGB values by calculating the nearest value with equal strength in all three channels (See Figure 1B for an example of this).

Next, we apply optical distortions according to the mean modulation transfer function of a well corrected human eye. To do this, we use a formula by Watson (2013), which was based on optical aberration measurements by Thibos, Hong, Bradley, & Cheng (2002) on 200 eyes of 100 healthy, well-corrected eyes. We fixed the pupil diameter required for these formulas at 4 mm for our simulations. The pupil diameter could be measured, experimentally controlled, or estimated from the luminance over the visual field (Watson & Yellott, 2012). However, in none of the experiments that we fit here pupil diameter or luminance were varied explicitly and conditions were reasonably similar in all experiments, such that we opted for this slight simplification.

Conversion of stimuli to contrast is then performed by dividing by their mean and subtracting 1. Then we cropped the stimulus to an area of $2° \times 2°$ of visual angle around the assumed fixation location (for most classical stimuli the center of the stimuli where they reach maximal nominal contrast). If the stimulus was smaller than $2° \times 2°$, we filled the rest of the area with zeros. Finally, we resized the image to 256×256 pixels using MATLAB's "imresize" function, which performs a bicubic interpolation.

We then implement the higher neuronal sensitivity for medium to high spatial frequencies as an additional linear filter similar to the "high pass filtering of neural origin" of Rovamo et al. (1993), which depends on presentation time. As in earlier approaches, we estimated the neuronal influence on contrast sensitivity simply as the necessary filter to match contrast sensitivity. To implement this filter with as few

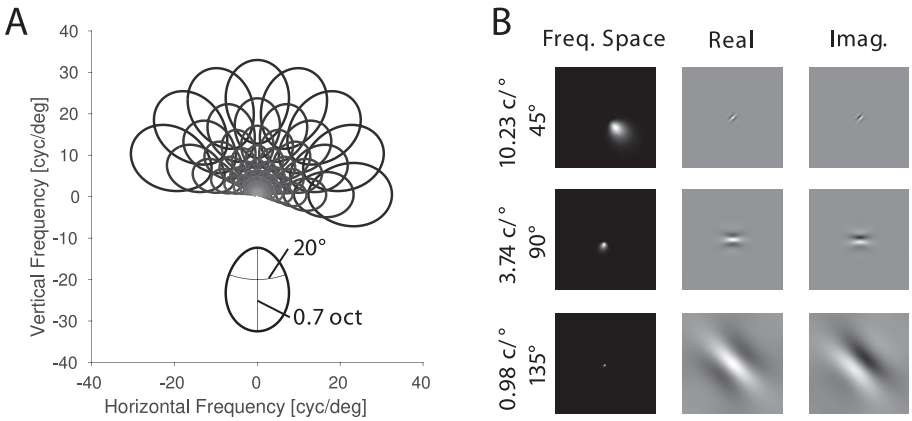

Figure 2. Illustration of the filters used for the decomposition. (A) half response curves in frequency space for all filters. Lighter gray for higher frequency channels. Additionally, one filter is displayed separately to show the half bandwidths at half height of each channel. The distribution of channels may appear tilted in the figure, because we included filters in the cardinal directions; however, by mirror symmetry the filters cover or tile the space equally. (B) Three example channels of different frequency and orientations relative to horizontal. For each channel, a heat map of the weights in frequency space and the real and imaginary part of the filter weights in space are given. The similarity of the filters to receptive fields of V1 neurons is not incidental.

assumptions as possible, we fitted its modulation transfer function (MTF) by hand as a third order spline.

To complete preprocessing, we smoothly cut out the image patch corresponding to the fovea, as we want to restrict ourselves to foveal processing here and to avoid any border effects in later processing. For this purpose, we use a $2° \times 2°$ raised cosine window. This window is above half height over the central disc of 1° diameter, roughly fitting the size of the foveola with maximal resolution and sensitivity.

The final preprocessing result for an example image is displayed in Figure 1C.

Decomposition

Next we aimed to implement the well-established orientation and spatial frequency selective channels (Campbell & Robson, 1968). These were implemented as a dense filter-bank with each individual filter fitting psychophysical and neuronal measurements of channel specificity, as illustrated in Figure 2.

Many functional forms exist that can represent the filter shape of the psychophysical channels closely enough. Here we chose to use a log-Gabor as the basic filter shape, which corresponds to a Gaussian shape in log-frequency and in orientation. A log-Gabor is directly and completely defined by its preferred spatial frequency and orientation and its bandwidth in each dimension, which are all properties estimated from psychophysical and physiological data routinely. Additionally, Gabor-filters are maximally localized jointly in space and frequency, have a monotonically and smoothly decreasing response for frequencies and orientations moving away from the preferred parameters, and no response to uniform fields. These are all desirable properties for a subband decomposition, which gives our filter choice some normative justification. Ultimately, however, any functional form that closely represents the specificities of the psychophysical channels (and thus, V1 neurons) will yield indistinguishable responses in the channels and thus results indistinguishable from our choice.

Additionally to spatial frequency and orientation specificity, linear filters are also tuned to the phase of the stimulus as simple cells in primary visual cortex are (Daugman, 1980). However, psychophysical performance seems not to depend on absolute phase. The most parsimonious model to achieve such phase independent behavior is to use a quadrature pair, i.e., filters which differ only in their phase preference and exactly by 90°. Such a quadrature pair is usually written as a single complex filter with one filter defining the real and one defining the imaginary part of the filter, optimizing the implementation further. From a quadrature pair, the response of a filter preferring any phase can be computed as a linear combination of these two filters. Especially, we can compute the absolute value of the complex response, which represents the response of an optimally phase-tuned channel at each position. For our channels we implemented this scheme and pass only the absolute value of each channel's response on to further processing, as illustrated in Figure 3. As we demonstrate in Figure 3B, this treatment of phase leads to a phase independent response.

Quadrature pairs could be implemented neuronally using four phase preference types of neurons for positive and negative responses of the two filters in the pair as discussed by Watson and Solomon (1997). Indeed, neurons in macaque primary visual cortex

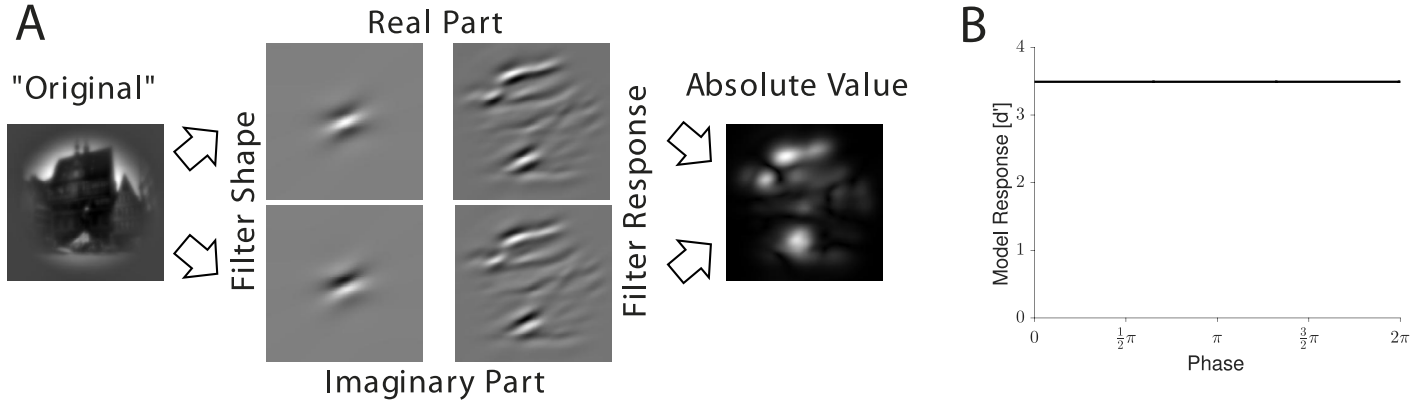

Figure 3. Illustration of the phase handling in the model. (A) The preprocessed example image called "Original" passes through the processing of a single channel. The complex filtering corresponds to filtering with two filter shapes, an even phase filter for the real part and an odd phase filter for the imaginary part, which are illustrated in the second column. In the third column the responses of the filters to the image are shown, which are then combined to the absolute value at each position illustrated in the last panel. (B) The model response (plotted as the signal-to-noise ratio d' for detection of the stimulus) to a $3° \times 3°$ Hanning-windowed horizontal grating of $10 \frac{cyc}{deg}$, changing the phase of the grating. The response of the model is phase independent up to numerical precision.

cluster around even and odd symmetric phases (Ringach, 2002). However, there are neurons at all preferred phases, and strongly orientation tuned neurons tend to prefer odd phase while less tuned neurons tend to prefer even phase. Both of these observations are incompatible with a direct implementation of quadrature pairs in neurons. Consequently, quadrature pairs must be seen as a simplification.

We set the bandwidth of the channels based on the literature, as we do not include data here that could constrain the spatial frequency selectivity of the channels. For spatial frequency, we chose a standard deviation σ_F of 0.5945 octaves corresponding to 0.7 octaves half bandwidth at half height, roughly matching the adaptation data of Blakemore and Campbell (1969) and the neural data of Ringach, Shapley, and Hawken (2002). For orientation, we chose a standard deviation σ_θ of 0.2965, corresponding to 20° half bandwidth at half height based on early psychophysical measurements (Campbell & Kulikowski, 1966; Phillips & Wilson, 1984). These measurements used data similar to our oblique masking data to estimate the bandwidth of the channels. Consequently, any substantial deviation of the estimates would be noticeable when comparing our predictions to these data. Additionally, these estimates are in good agreement with physiological measurements (Campbell et al., 1968), as already noted in the original papers and do fit more modern measurements like Ringach et al. (2002). Nonetheless, our filter collection only roughly approximates the neural population, because there is substantial variability in the specificity of cortical neurons (Goris, Simoncelli, & Movshon, 2015; Ringach et al., 2002) and we ignore known dependencies between preferred spatial frequency and the bandwidths (Phillips & Wilson, 1984), an issue on which we comment in more detail in the Discussion.

Finally, we need to specify how many channels at which spatial frequencies and orientations to use. Normative theory from signal processing tells us that two different orientations and octave-spaced spatial frequency channels suffice to represent the whole information present in an image as it is done for wavelet decompositions (Strang & Nguyen, 1996). Commonly, pyramid schemes are applied to achieve such a decomposition with as few filter responses and as little computation as possible (Simoncelli, Freeman, Adelson, & Heeger, 1992; Watson, 1987). Specific types of filters allow these pyramids to achieve additional advantageous properties like steerability or shiftability (Freeman & Adelson, 1991; Simoncelli et al., 1992).

To achieve this, however, one needs to choose specific filter shapes which need to be broad in frequency and orientation. Using narrower filters, more different filters are required to cover all orientations, and there are only discrete choices which fix both bandwidth and number of channels in each scheme. Even worse, for spatial frequency the whole pyramid scheme breaks down once one wants channels that are not octave-spaced because downsampling by other factors than two is much less efficient. Thus, these pyramid schemes do not allow us to fit the channel bandwidths and the density of channels independently and limit us to octave-spaced channels.

One could glance over this and approximate the filters with the best fitting pyramid as Watson and Solomon (1997) did, for example, were we not using nonlinear processing after the decomposition. A stimulus that matches a filter in the model leads to a single large response in that channel, whereas a stimulus between channels leads to several smaller

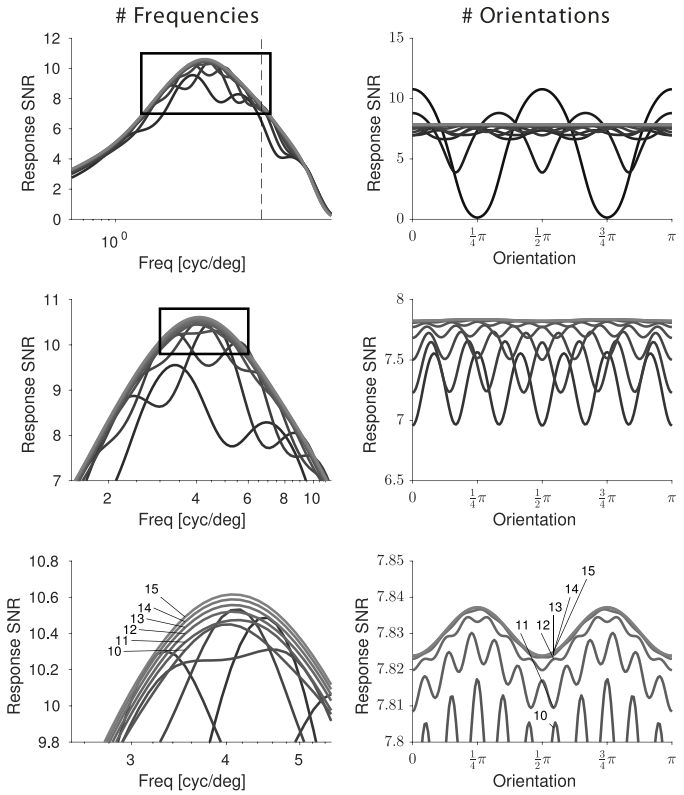

Figure 4. Illustration of the effects of using fewer channels on model performance. Left column: Estimated signal-to-noise ratio for a 3° × 3° Hanning-windowed horizontal grating with 10% contrast against the frequency of the grating. Lighter gray levels correspond to more channels. Each row shows the marked area in the row above, representing different zoom levels. Right column: As the left column, but fixing the frequency of the grating to 10 $\frac{cyc}{deg}$ (marked in the left column by a dashed line) and varying the orientation of the grating instead. The results shown here were obtained for 256 × 256 pixel images, but the effects are largely independent of image size.

responses. Then the accelerating nonlinearity amplifies the larger response more than the several smaller responses leading to a stronger model response to stimuli that match a channel than to stimuli that fall between channels.

In our model this leads to an oscillating response with peaks at the orientations and spatial frequencies of the channels (see Figure 4). Note that such oscillatory behavior must occur for any model that employs nonlinearities after the decomposition in channels for specific frequencies and orientations. A nonlinearity imposes different weights on the channels depending on signal strength. However, the activities of any set of linear channels keep the same relative strength when the absolute signal strength changes. Thus, no linear channel shape can fully compensate for the nonlinearity unless the nonlinearity is computing energy, i.e., squaring and summing over channels.

In contrast, one observes neither oscillating performance nor any clustering in preferred spatial frequency or orientation in either psychophysics or neurophysiology. Neurons seem to cover every frequency and orientation in the range they cover, and human performance on psychophysical tasks seems to change smoothly with scale and orientation.

To mimic the dense neural covering of spatial frequencies and orientations, we chose to simply increase the number of frequency and orientation channels until the oscillations of performance were sufficiently small (see Figure 4). This method allows us to keep the implementation as a convolution, which is still necessary to reach an acceptable computation time. An implementation that includes a realistic sampling of the channels would go far beyond our horizon here as this seems not to be constrained psychophysically, and such decompositions with variable channels were not studied in detail so far.

Following these considerations, we used a complex-valued log-Gabor filterbank with 8 × 12 filters for orientation and spatial frequency for our decomposition. The eight preferred orientations were equally spaced over 180°, covering half the frequency space. The 12 preferred spatial frequencies were placed logarithmically on the spatial frequency axis from 0.5 $\frac{cyc}{deg}$ to 20 $\frac{cyc}{deg}$, which roughly covers the range of frequencies visible to human observers. The kind and range of filters we used are illustrated in Figure 2.

Each of the filters was precomputed in frequency space. We then calculated the filter response by multiplying the Fourier transform of the preprocessed image with the frequency space representations of the filters, which yields a complex-valued images for each channel. This complex-valued image contains responses of an even symmetric filter as its real part and the responses of an odd symmetric filter as its imaginary part. As discussed above, we pass the absolute value of this response on to further processing, dropping phase entirely.

The results of the whole decomposition stage are illustrated for the example natural image in Figure 1D and E. In D we show the results before and after removing phase information for three example channels. In E you find an overview over all channels in which we display only the average absolute response of each channel.

Normalization and nonlinearity

Masking and contrast discrimination experiments show clearly nonlinear relationships between thresholds and mask contrast (Legge & Foley, 1980). To model these psychophysical results and the corresponding interactions observed in primary visual cortex neurons

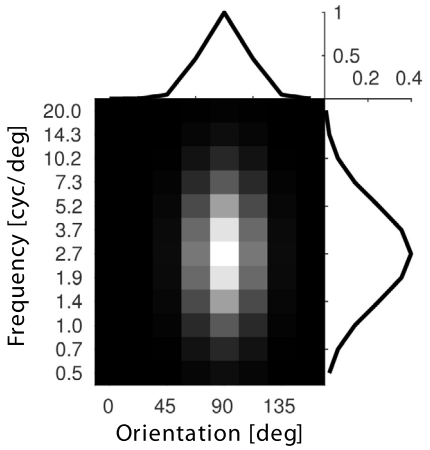

Figure 5. Illustration of the normalization pool G over spatial frequency and orientation. Both are shown for the central pixel in the 2.7 $\frac{cyc}{deg}$, 90° orientation channel.

(Cavanaugh et al., 2002a; Heeger, 1992), the channel activities are passed through a divisive normalization (Carandini & Heeger, 2011; Foley, 1994; Heeger, 1992; Watson & Solomon, 1997). In our model, we restrict the normalization to a pool localized in space, spatial frequency, and orientation. The localization in space and frequency is not controversial, while it is sometimes claimed that the normalization pool is not orientation selective, on which we comment in the Discussion.

In older models, this step was modelled as a Naka-Rushton nonlinearity (Foley & Legge, 1981; Legge & Foley, 1980; Naka & Rushton, 1966), which is equivalent to this normalization with an extremely narrow pool that contains only the channel itself as an input.

In our model the formula for divisive normalization of original channel activities $A = (a_i)_{i\in\mathcal{I}}$ to compute normalized final responses $R = (r_i)_{i\in\mathcal{I}}$ is

$$r_i = \frac{a_i^{p+q}}{C^p + b_i} \quad (1)$$

Using an index set $\mathcal{I}$, which indexes all different channels and all positions, a constant C, exponents p and q and $B = \{b_i\}_{i\in\mathcal{I}}$, an array of normalization coefficients, which are computed from the element wise powers $A^p := (a_i^p)_{i\in\mathcal{I}}$:

$$B = A^p * G \Leftrightarrow b_i = \sum_{j\in I} G(x_i - x_j) a_j^p, \quad (2)$$

by convolution with G, a 4D Gaussian normalization pool with standard deviations $\omega_x = \omega_y$ in space, ω_F in spatial frequency, and ω_θ in orientation.[1]

The weights for the pool in spatial frequency and orientation are displayed in Figure 5. For frequency we set this to a rough estimate of $\omega_F = 1$ octave standard deviation. For orientation we fit the pool bandwidth ω_θ based on oblique masking data (displayed in Figure 11), as explained in more detail below.

For the spatial extent we first implemented the model using a Gaussian profile. However, we lack the data to constrain the size of the normalization pool in space. Instead of arbitrarily setting a pool size, we tested the extreme cases of such a model here. Specifically, we set the normalization pool to be either the exact pixel to be normalized only or all responses over the image weighted equally. These cases correspond to an infinitely small and an infinitely large pool respectively. For the classical grating based data, we find that the normalization over the whole image leads to a better result and more consistent parameter estimates, whereas the natural image data is better explained by the perfectly local normalization.

Nonetheless, we believe neither that the normalization pool is perfectly local nor that is fills the whole space. Both psychophysical (Snowden & Hammett, 1998) and neural data (Cavanaugh et al., 2002a) suggest that the normalization pool has some extent beyond the classical receptive field (roughly 2.5–3 times the radius from the neuronal data). Also our model allows arbitrary intermediate sizes for the normalization pool, and sporadic fits we made with intermediate pool sizes yielded good fits to the classical data as well. Consequently, we do not argue against the normalization pool having a nonzero spatial extent.

We require the additional exponent q, because a single saturating function per channel cannot explain the discrimination thresholds at high contrasts, which grow much less than predicted from a saturating response function (Goris et al., 2013). This approach was used earlier by Foley (1994) and Watson and Solomon (1997) in their models.

The neural mechanism allowing high contrast discrimination with saturating neurons seems to be neurons with higher C, which start to respond only at higher contrasts. Following Watson and Solomon (1997), we interpret the function in (1) as the sum of responses of neurons responsible for different contrast ranges. For such a sum, the formula with $q > 0$ is practically equivalent as Watson and Solomon (1997) discuss in detail (see their figure 16 and discussion point 4.E). As we are not aware of any psychophysical data requiring a separation into contrast channels, we do not include this complication here.

The results after the nonlinearity and normalization are displayed for a natural image in Figure 1F and G. As for the raw decomposition in D and E, the spatially resolved responses for three example channels are displayed in F, and the average response for all channels in G.

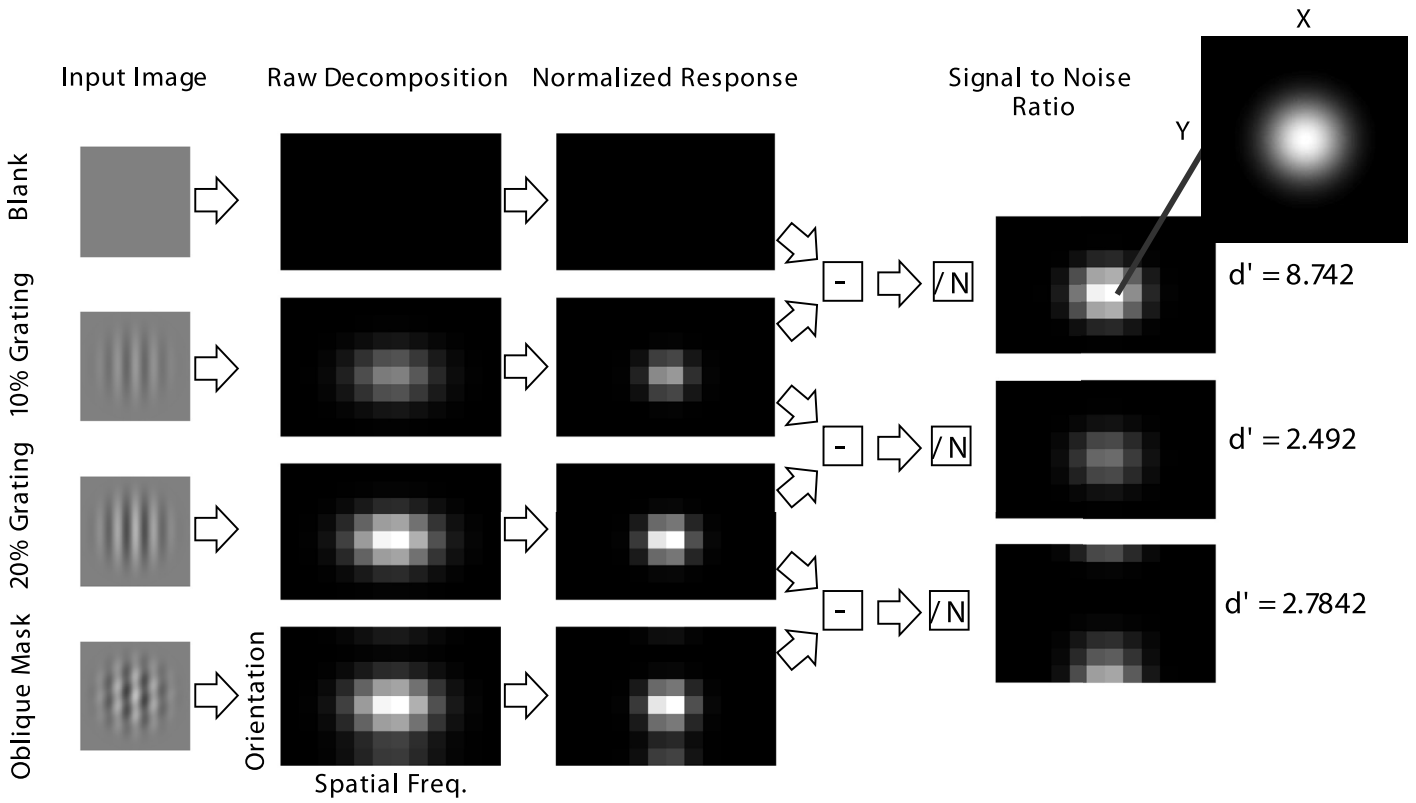

Figure 6. Illustration of the readout mechanism of the model. For four different typical spatial vision stimuli, we show the channel mean of the raw decomposition results and of the final normalized responses. To predict how well two images can be differentiated in psychophysical experiments, the responses for each pixel in each channel are subtracted from each other and divided by the noise standard deviation. This formula results in a signal-to-noise ratio for each position in each channel indicating how well this pixels' activity differentiates the two images. The mean of these signal-to-noise ratios over each channel are shown in the last column. For one channel we also show the spatial distribution of the differentiability and for each we computed the overall discriminability d'. The three pairs of stimuli correspond to contrast detection, contrast discrimination, and oblique masking experiments, respectively.

Noise and decoding

Finally, we need a method to quantify how well stimuli can be discriminated based on their model representations. Here we model noise on the channel outputs and then assume that the rest of the brain optimally decodes from the noisy channel outputs. This allows us to predict whole psychometric functions, i.e., how the proportion of correct responses grows with growing differences. Additionally, it provides a more plausible mechanical interpretation than just computing the difference and pooling with some Minkowski norm as done by earlier models. Our computations for this are illustrated for some typical spatial vision stimuli and tasks in Figure 6.

For our model, we assume independent Gaussian noise for each individual pixel in each channel whose variance scales linearly with the activity in the channel. This model allows us to scale smoothly between pure constant noise and noise that scales completely with the response. Obviously, the independent Gaussian is a specific choice. However, the decision variable will be roughly Gaussian distributed whatever the original distribution was, as our decoding combines many responses for any decision. We also include no noise correlations here, as it would impose a high computational hurdle and is most probably not constrained by the psychophysical data. We discuss our choice of noise in some more detail in the Discussion.

Using this noise model, we can compute a signal-to-noise ratio for each pixel's ability to discriminate a pair of images. Finally we combine the information using optimal linear decoding, which boils down to a weighting by the signal-to-noise ratio, as the pixels are modelled as independent.

First, we calculate the variance of the Gaussian noise n_i for any response r_i of the model:

$$n_i = N_c + N_f r_i \quad (3)$$

using two parameters, the variance of a constant noise source N_c and the factor for the linear noise N_f. When fitting to data, we found that q and N_F can compensate each other, such that we set $N_F = 0$ regressing to constant noise below (see Appendix A for details on this).

For the ith pixel we can then calculate the signal-to-noise ratio for differentiating two images (1) and (2) from the model responses $r_i^{(1)}$ and $r_i^{(2)}$ at this pixel:

$$s_i = \frac{(r_i^{(1)} - r_i^{(2)})}{\sqrt{n_i^{(1)} + n_i^{(2)}}} \quad (4)$$

Using this signal-to-noise ratio we can calculate the mean value d_i and variance η_i for each pixel weighted by its signal-to-noise ratio for discriminating this specific pair of images:

$$d_i = s_i(r_i^{(1)} - r_i^{(2)}) \qquad \eta_i = s_i^2(n_i^{(1)} + n_i^{(2)}) \quad (5)$$

From that we arrive at the summed signal d and its variance η and can calculate the percent correct p_c' for a 2AFC task using the standard cumulative normal distribution Φ:

$$p_c' = \Phi\left(\frac{d}{\sqrt{\eta}}\right) = \Phi\left(\frac{\sum_{i\in\mathcal{I}} d_i}{\sqrt{\sum_{i\in\mathcal{I}} \eta_i}}\right) = \Phi\left(\frac{1}{\sqrt{\sum_{i\in\mathcal{I}} \eta_i}} \sum_{i\in\mathcal{I}} s_i(r_i^{(1)} - r_i^{(2)})\right) \quad (6)$$

Note that this system applies only for exactly two images to be compared. If one wanted to decode information about groups of stimuli, the optimal decoder is almost always more complex.

For the natural images we once chose a simpler decoding principle. The simpler decoder weights all pixels and channels equally, i.e., (5) is replaced by $d_i = |r_i^{(1)} - r_i^{(2)}|$ and $\eta_i = n_i^{(1)} + n_i^{(2)}$. This essentially assumes that the decoder weights all channels in the correct direction, but has no information on how well each channel discriminates.

Finally, to handle rare lapses of subjects, we simulate a lapse rate of 1% by rescaling p_c' into the final p_c

$$p_c = \lambda + (1 - 2\lambda)p_c' \quad (7)$$

with $\lambda = 0.005$. Taking these lapses into account is necessary as a predicted p_c of 1 renders failures impossible. Thus, without a modelled lapse rate, lapses at high stimulus levels can strongly influence parameter estimates (Wichmann & Hill, 2001).

Calculating thresholds

Our model calculates percent correct for differentiating two images. Thus, we require a method to calculate thresholds. We chose to calculate thresholds by a bisection method.

We start by testing whether the model predicts observers to be correct at maximal displayable contrast (one minus the mask contrast) with a probability higher than a threshold (typically 75%). If this is the case, we start the bisection method with 0 contrast and the maximal displayable contrast defining the first interval.

In each step of the bisection method, we calculate the predicted percent correct for the center of interval calculated so far and take this point as the new top or bottom end of the interval depending on whether the predicted percent correct is larger or smaller than the threshold percent correct.

We repeat bisection method steps until the width of the interval divided by the lower end is less than 5% and use the center of the last interval as the threshold estimate.

Parameter fits

We fixed our model up to the decomposition into different spatial frequency channels without free parameters. After this regulating, however, there are some parameters that we need to fit to data. Namely the two exponents p and q, the constant of the normalization C, the bandwidth of the normalization pool $\omega_{\theta,}$ and the noise strengths N_C and N_F.

To fit parameters, we calculated a single maximum likelihood fit to the data obtained from all observers. This adequately weights the different datasets we have for estimating parameters and uses all data well.

In short, we started with a grid search over the unset parameters. As a conclusion from this grid search, we restricted ourselves to a purely constant noise source setting N_F to zero as we found that changing q can fully compensate for different N_F, such that the model can explain the data equally well, largely independent of N_F. Additionally, we fixed the bandwidth of the normalization pool ω_θ based on the oblique masking data starting an optimization of this parameter from the grid search result.

Using the fixed normalization bandwidth and the purely constant noise source, we then fitted the other parameters to the contrast discrimination data for each presentation time and once additionally for the oblique masking data. For this fitting step we used a quasi-Newton optimization.

Additionally, we decided to fit the parameters again for the ModelFest dataset. As we have only threshold data for this dataset, we had to convert these thresholds into contrast, percent correct pairs for fitting. When we use only a data point at threshold, this favors shallow psychometric functions that predict threshold percent correct for any pair of stimuli. To avoid this problem, we added a data point at 1.5 times threshold contrast with 199 of 200 trials correct and a data point at the third of the threshold with 100 or 200 trials, which represents change performance. As threshold detection data usually do not constrain the normalization exponent q, we fixed it to the value from our longest

presentation time of 1497 ms. Fits with this parameter free yield similar prediction quality.

A more detailed description of our method of fitting is given in Appendix A.

Data for model evaluation

The data for contrast detection, contrast discrimination, and oblique and plaid masking were collected during the doctoral studies of Wichmann (1999). Some of the data are published in Bird, Henning, and Wichmann (2002). In these reports all technical details can be found, and we report only an overview here.

The classical psychophysical data were collected as temporal two alternative forced choice (2-AFC) experiments; i.e., two stimuli were presented in succession and the observers' task was to report which time interval contained the signal. Presentation time was marked with tones, and there was immediate auditory feedback indicating which was the correct interval. In total, seven observers participated; they were all experienced psychophysical observers, were aware of the purpose of the experiments, and had normal or corrected-to-normal visual acuity. Stimuli were presented on a calibrated, digitally linearized CRT screen with a mean luminance of $88.5\,\frac{cd}{m^2}$ with a refresh rate of 152.3 Hz. To guarantee independence of signal and mask in the stimuli, they were presented in different refreshes combining three refreshes into one frame (one for the signal and one for each of two possible masks). There were three different temporal presentation modes: (a) Stimuli were presented for a single frame, i.e., three refreshes, nominally for 19.7 ms. (b) Stimuli were presented for 4×3 frames, nominally 79 ms. (c) Stimuli were presented with the contrast of all components following a Hanning window of 1497 ms total duration. All reported contrasts are the peak contrast at the center of the time interval. To extract thresholds from the data we fitted the data using psignifit 4 with the standard prior set based on the tested stimulus range (Schütt, Harmeling, Macke, & Wichmann, 2016). Error-bars represent 95% credible intervals.

We also present data from the Modelfest dataset (Watson & Ahumada, 2005) and a natural image-masking database (Alam et al., 2014) here. The Modelfest dataset consists of contrast detection thresholds for 43 different 256×256 pixel targets presented at 120 pixels/°. Target contrast was temporally modulated by a Gaussian envelope with a standard deviation of 125 ms. The natural image-masking database consists of the detection thresholds for $3.7\,\frac{cyc}{deg}$ log-Gabor-filtered noise targets masked by 1080 natural image patches taken from 30 black and white digital photographs. Thresholds were measured using a spatial three alternative force choice task. Three stimuli were presented simultaneously, and subjects had 5 s to indicate which stimulus contained the noise Gabor target overlaid over the natural image patch. Further technical details for these datasets are provided in the original studies.

Results

Classical psychophysical results

We first test our model on classical psychophysical experiments. These experiments were specifically designed to test hypotheses about early spatial visual processing. To achieve this, the stimuli are composed of sinusoidal gratings intended to activate the spatial frequency and orientation channels as specifically as possible. We shall start with the sensitivity of single channels and continue with masking experiments, which test how well activation of additional channels masks the signals.

Contrast detection

We present detection data for three different temporal presentation modes, roughly 20 ms and 80 ms with hard on and offsets and contrast changing according to a 1.5-s long Hanning window/raised cosine window.

The data are presented in the form of *contrast sensitivity functions* (CSFs) in Figure 7. The contrast sensitivity functions show the typical bandpass shape for long presentation times and the more low-pass shape for the short presentation times.

The model reproduces the contrast sensitivity functions closely. This is not surprising as we fitted a weighting for the spatial frequencies in our preprocessing for each presentation time.

ModelFest

Next we evaluate our model against the ModelFest database, incorporating detection performance for 43 different patterns measured with many observers in different labs.

The results of our model for these data are displayed in Figure 8. First we ran our model with a new contrast sensitivity function and the parameters fitted for the adjacent presentation times. With these parameters we already obtained promising fits displayed as the gray lines in Figure 8, which fitted almost all patterns in the data. The main error seems to be a constant offset, which we could probably correct by adjusting the initial weighting filter. Using parameters fitted to the data, we

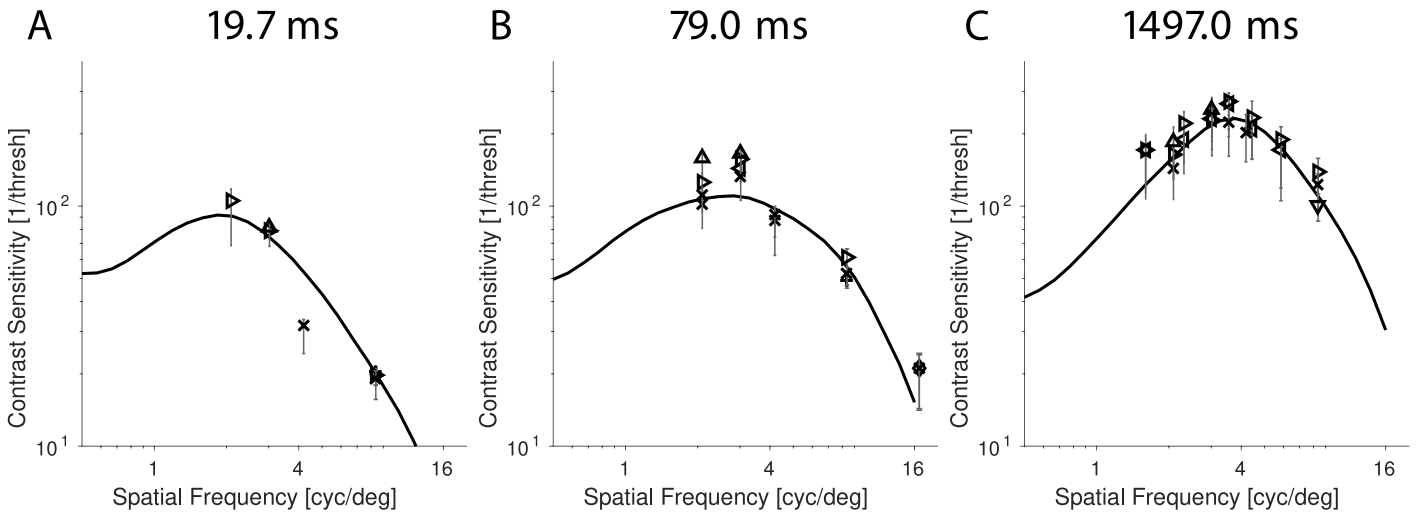

Figure 7. Results for the contrast detection data for different presentation times. Different symbols represent the measured data from different observers. Each observer has their own fixed symbol across all figures. Error bars represent 95% credible intervals from a Bayesian analysis of individual psychometric functions. The continuous line represents the prediction of the model. (A and B) Both 19.7- and 79-ms (three and 12 frames) presentation time with hard on- and offsets. (C) Contrast Hanning-windowed in time with a total presentation time of 1497 ms.

obtain an even slightly better fit to the data plotted as the black line in Figure 8.

The clearly largest deviation from the data for all parameter settings is the Gaussian blob (stimulus #26). This very low spatial frequency target is strongly affected by our initial luminance normalization. Consequently, we believe that this represents a problem of our overly simplistic preprocessing, which ignores stimulation before and after the stimulus, which sets the adaptation level differently from the mean of the image presented.

Contrast discrimination

The next type of data we compare our model to is contrast discrimination data, which originally motivated the nonlinearity (Foley & Legge, 1981; Legge & Foley, 1980). Here the task is to report which of two presented gratings has the higher contrast, i.e., to discriminate gratings, that differ only in contrast.

We start by investigating only the 78.8-ms presentation data presented in Figure 9. At all spatial frequencies the thresholds for discrimination follow the classically observed dipper shape (Foley & Legge, 1981; Legge & Foley, 1980). All curves first decrease

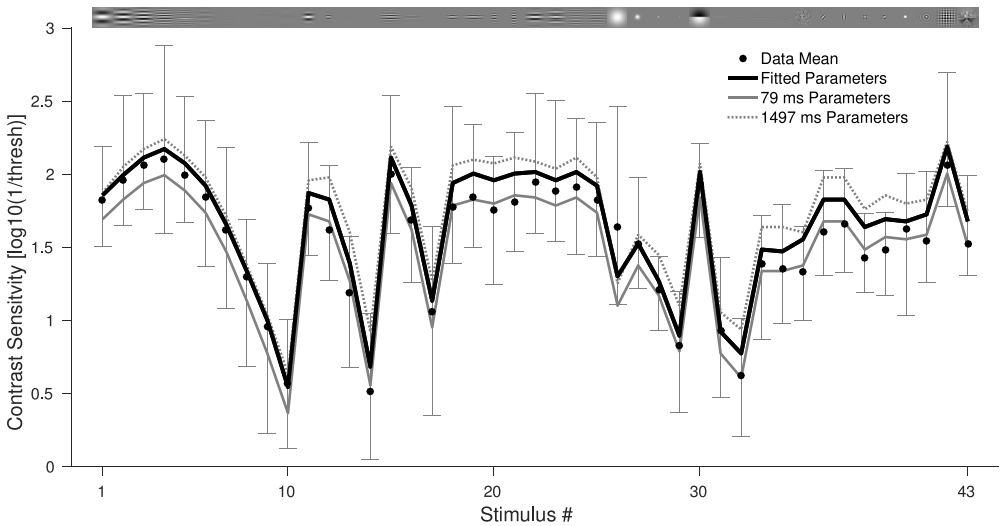

Figure 8. Results for the ModelFest dataset. We here plot (log-) contrast sensitivity for the 43 different stimuli ordered along the *x* axis. The dots represent the average measured threshold, with error bars representing the range of measured thresholds. The lines represent the predictions of our model using different parameters. Above the plot we show tiny full contrast images of the stimuli.

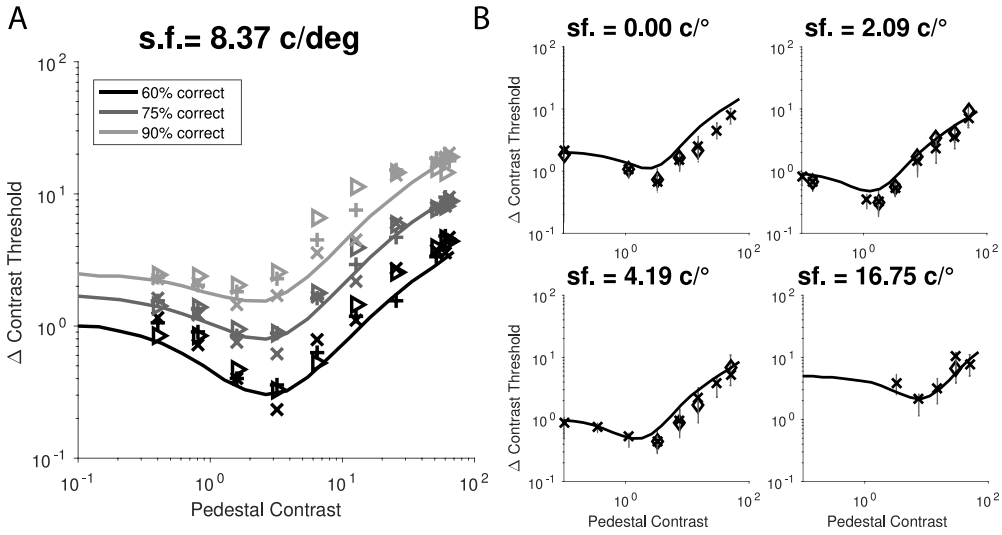

Figure 9. Results for contrast discrimination data. All data were collected with 79-ms presentation time with hard on and offsets. (A) Data for 8.37 $\frac{cyc}{deg}$, the frequency for which we have the most data. The different gray values indicate different percent correct to be reached to define the threshold. The difference between these lines illustrates the change in the slope of the psychometric function over the range of contrasts. Specifically it is shallower in the dip and steepest for detection. (B) Results for different spatial frequencies. Here only the data for the 75% contrast are shown. 0.00 $\frac{cyc}{deg}$ indicates discrimination in the brightness of a blob. All other conventions are as in Figure 7.

such that at low pedestal contrasts, contrast discrimination is easier than detection (Nachmias & Sansbury, 1974). At higher contrasts, discrimination thresholds lie roughly on a straight line in the log-log plot, indicating a power law for the contrast discrimination threshold.

The model reproduces the contrast discrimination curves quite well for all spatial frequencies. Also the slopes of the psychometric functions seem to be captured by the model, since we fit thresholds at different performance levels. Especially the shallower psychometric functions in the dipper reported by Bird et al. (2002) are reproduced.

Next, we can investigate how contrast discrimination performance varies with presentation time. For the 8.37 $\frac{cyc}{deg}$ target we also have contrast discrimination data at the two other presentation times of 19.7 ms and the 1497-ms Hanning window.

These data with model fits are plotted in Figure 10. In each panel we show the data measured with given presentation time together with three different fits. All of these fits use the contrast sensitivity filter fitted for the correct presentation time, but normalization

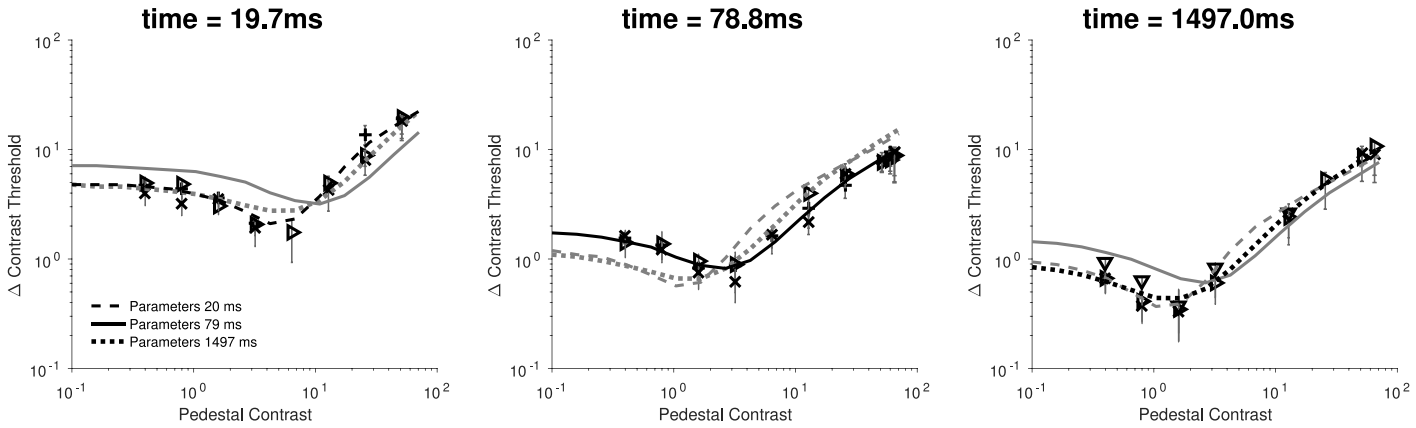

Figure 10. Results for contrast discrimination data for different presentation times. Each panel shows the contrast discrimination data for the 8.37 $\frac{cyc}{deg}$ for one presentation time. Again different symbols show the 75% threshold from different observers with 95% credible intervals. The lines represent the predictions from three different sets of parameters. In each panel the prediction with parameters fit to the displayed data is highlighted in black. All other conventions are as in Figure 9.

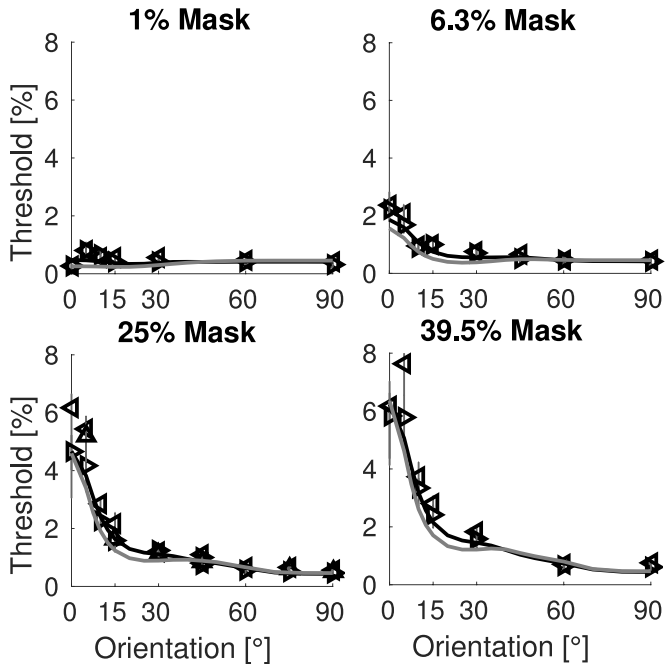

Figure 11. Results for oblique masking experiments, spatial frequency for both signal and mask was $3\frac{cyc}{deg}$. As in previous figures, symbols represent data and lines the predictions of our model. The black line uses parameters specifically fit to the oblique masking data; the gray line is the prediction using the parameters estimated using all data at the long presentation time of 1497 ms.

parameters fitted to the three presentation times. The model can reproduce the data for each presentation time. However, the different presentation times require different parameters since the curves simulated from a single parameter set do not capture the data adequately. Especially the width of the dip and its position relative to the detection threshold differ between presentation times.

Oblique masking

Next we compare our model to oblique masking data, which represent the psychophysical reason for replacing the channel wise nonlinearity with normalization across channels (Foley, 1994). Here the task is to detect the presence of a horizontal grating, while all observation intervals contain an additional "oblique mask," i.e., another grating of the same spatial frequency and spatial envelope, but with a different orientation. All oblique masking experiments were performed with the 1497-ms presentation time and 3° × 3°, $3\frac{cyc}{deg}$ targets.

Results of these experiments are presented in Figure 11. While the masking effect of nearby orientations is slightly underestimated by the model the overall fit of the model to the data is good.

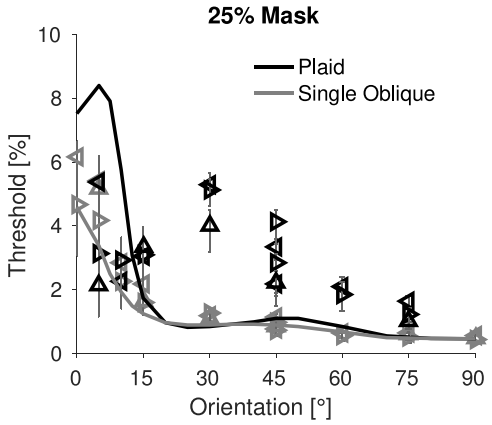

Figure 12. Results for Plaid masking experiments, here only for the 25% contrast mask. The plaid data and model predictions are plotted as the black symbols and line again. Additionally we replot the data and prediction for a single oblique mask from Figure 11 in gray.

Plaid masking

The next type of data we compare our model to is plaid masking data. Here, the task is the same as for oblique masking, but the one oblique mask is now replaced with two masks rotated away in opposite directions from the signal orientation, which are together called a plaid.

Results of these experiments are displayed in Figure 12. Characteristic for these experiments is that at relatively high contrast (here 25%) plaids 30°–45° (and even further away from the signal orientation) substantially mask the signal, while each of the two gratings composing the plaid alone hardly mask the signal. Thus, the two gratings' masking capabilities combine strongly superadditively. To show this superadditivity, we replotted the oblique masking data in the figure.

Our model fails to replicate the super additive masking effect of plaid masks, as most probably all other spatial vision models based on the multiresolution theory do (Derrington & Henning, 1989). A clearly favoured explanation of this effect has not yet emerged although it is strong and reliable. For some weaker forms of plaid masking where the signal and mask are separated in spatial frequency, linear summations over channels can explain plaid masking (Holmes & Meese, 2004). For the effects of plaids of the same spatial frequency only speculations exist though. One is that plaid masking is a perceptual effect created because observers frequently perceive high contrast plaids as "checkerboards" oriented between the orientations of the plaid components (Georgeson & Meese, 1997). A different one is that the recurrent dynamics of V1 might create activity at orientations different from the signal orientations, especially at the orientation between the two plaid components (Carandini & Ringach, 1997).

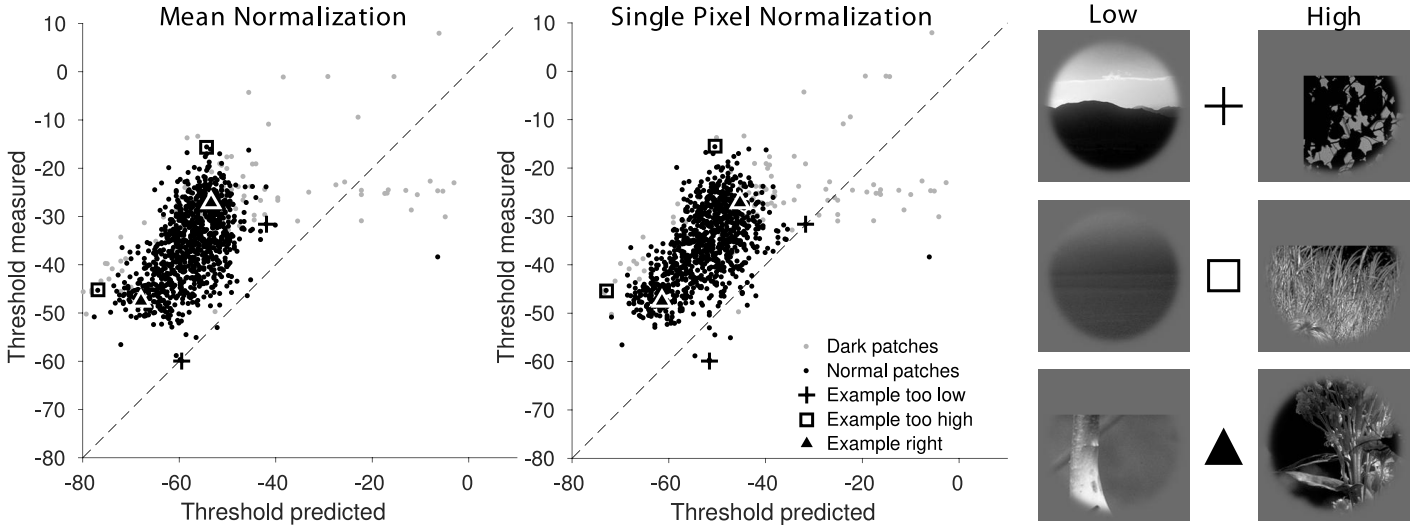

Figure 13. Results for the natural image masking database. First we plot the measured thresholds against the predictions of our model setting the spatial extent of the normalization pool either to the whole image or to a single pixel. Patches darker than $4\frac{cd}{m^2}$ are plotted in gray, all others in black. Additionally, we marked one low and one high threshold example patch each, where the measured threshold was higher, lower, or roughly equal to the prediction.

However, neither of these suggestions can be easily incorporated into the kind of model we propose here.

Natural scene masking database

To include some evaluation of our model on more natural stimuli than gratings, we evaluate our model on a natural image-masking database (Alam et al., 2014). The database consists of the detection thresholds for log-Gabor filtered noise targets masked by 1080 natural image patches taken from 30 black and white digital photographs.

To apply our model, we used a single exemplar of the noise, which accompanies the database and calculated its detectability on the different patches imitating the conditions the subjects saw in the experiment as closely as possible. As subjects were allowed to move their eyes and our model cuts out a rather small foveal area, we simulated not only a fixation at the exact center of the patch and signal, but also at the eight points moved 0.5° up and down and/or left and right from the center. Following the overarching theme of optimality, we display the lowest of the nine thresholds obtained this way. For the parameters, we chose the parameters for the long, 1.5 s Hanning window as the natural image patches were displayed for an even longer time of 5 s.

To convert the images to luminance values, we used the formula provided with the database, although it returns values smaller than the minimum luminance of the monitor reported in the paper. Thus, the data for dark patches seems to be unreliable. Also, the original paper excluded patches with low average luminance. Consequently, we follow the lead of the original paper and exclude patches with an average nominal luminance below $4\frac{cd}{m^2}$ from further analysis. These excluded patches are still displayed in Figure 13 as gray dots.

The results of our model are displayed in Figure 13. We find that the model generally overestimates the sensitivity of observers on the natural image stimuli, but produces thresholds highly correlated to the measured ones and thus seems to represent a sensible upper bound on these data. Models designed and adjusted specifically to fit this database can produce higher correlations with the data (Alam et al., 2014, 2015). Nonetheless, for generalization from grating-based experiments, the predictions seem to be quite accurate. Also, we err in the explainable direction. It seems plausible that highly trained observers perform better on simple grating stimuli without any random variation than less trained observers on natural image patches whose exact properties they were not extensively familiar with.

Surprisingly, we find that the single pixel normalization scheme, which was problematic for predicting the classical grating data, yields a higher correlation to human thresholds ($r = 0.5801$) than the mean normalization scheme ($r = 0.5196$), which was better at predicting the grating data. Tentatively, we assume that there is a local normalization scheme of medium size, which still fits the grating data and produces an equally good prediction as the local normalization.

One possible explanation for why our model predicts too low thresholds for the natural image stimuli might be that subjects are worse at decoding the noise signals

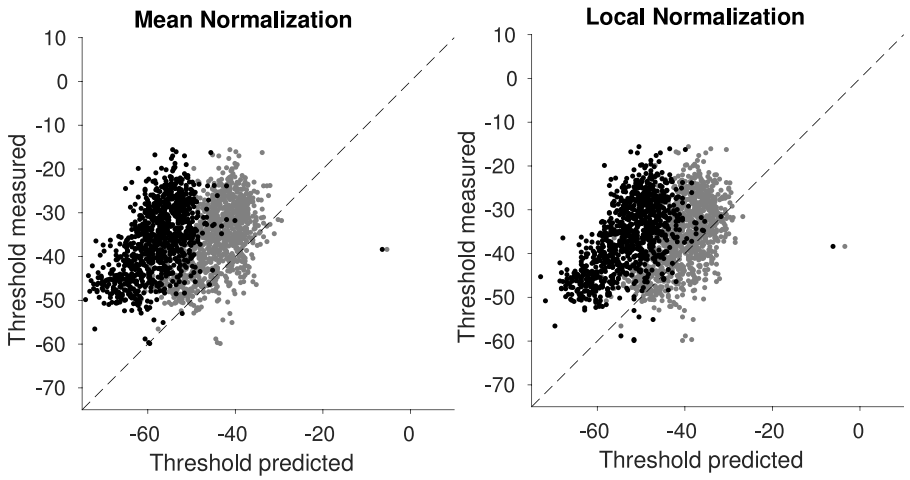

Figure 14. Using a weaker decoder for predicting the early natural image-masking database. The gray dots represent the predictions from the model when differences between the two images are summed disregarding their signal strength. The black symbols reproduce for the optimal decoder from Figure 13.

on the natural image masks than they are in the simpler classical grating experiments. In Figure 14 we show one specific weaker decoder. Namely, it weights any difference only by its sign instead of its signal-to-noise ratio. This limitation results in a decoder that simply adds all image differences, but ignores how well the specific channel differentiates the two images. This scheme is equivalent to taking the Minkowski-1-norm of the difference between the images drawing the connection to earlier models. Clearly such a simpler, worse decoder moves our predictions much closer to the measurements. However, we do not claim that this specific decoder mimics human behavior, as many other bad decoders would certainly increase the predicted thresholds equally. Nonetheless, this illustrates the point that a realistic but suboptimal decoding could explain the weaker performance of subjects in this natural image-masking task.

Different parameter sets

To further investigate the models' internal processing, we shall have a look at how the parameters needed to be changed to fit the different presentation times and data types. As described in detail in Appendix A, we first fit the longest presentation time for which we also have the oblique masking data to fix the orientation bandwidth of the normalization pool and then fit the parameters of the final normalization for the different presentation times and for ModelFest.

The parameter fits are given in Table 1. First, note that the linear contribution to the noise N_f is 0 for all datasets. We set this because we noticed that the exponent q can compensate for vastly different N_f such that all of them explain the data equally well (see Appendix A). Additionally, there is a presentation time dependent scaling of the input in our model. Thus, the constant C cannot be compared directly across presen-

Parameter	Meaning	19 ms	79 ms	1497 ms	Oblique	ModelFest
N_c	Constant noise variance	**1.4389**	**0.6450**	**0.4763**	**0.4235**	**0.0070**
N_f	Noise variance factor	0*	0*	0*	0*	0*
C	Nonlinearity, constant	**0.0031**	**0.0046**	**0.0027**	**0.0014**	**0.0147**
p	Nonlinearity, exponent	**2.7996**	**2.0253**	**1.8667**	**1.3732**	**1.2090**
q	Difference exponents	**0.3767**	**0.3676**	**0.3032**	**0.3755**	0.3032
ω_θ	Normalization pool orientation	0.2008	0.2008	0.2008	**0.2008**	0.2008
σ_θ	Filter standard deviation orientation	0.2965	0.2965	0.2965	0.2965	0.2965
ω_f	Normalization pool frequency	1	1	1	1	1
σ_f	Filter standard deviation frequency	0.5945	0.5945	0.5945	0.5945	0.5945
$\omega_x = \omega_y$	Normalization pool space	—	—	—	—	—

Table 1. Parameter values used for the different experiments. *Note*: The bold values were fit for the data in the experiment; the others were kept at the values we estimated from the 1497-ms presentation time, as we had the most oblique masking data to constrain the parameters at that time.

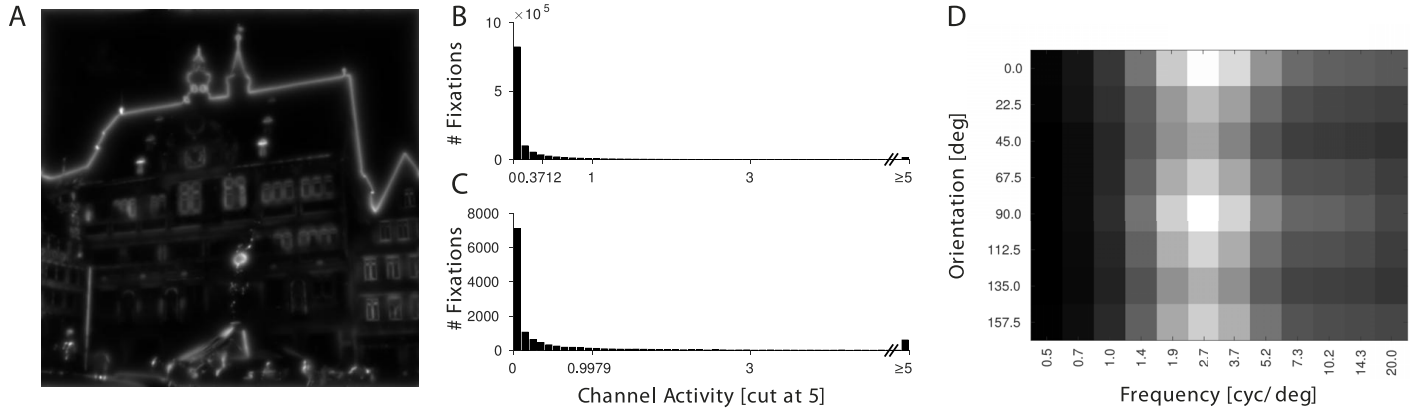

Figure 15. Some information on the output of the model. (A) Square root of the sum of the outputs of all channels for the example photograph of the town hall of Tübingen as an example unrelated to panels (B), (C), and (D). The output of our model highlights edges. (B) Histogram of all channel activities over fixation locations in natural images. As the highest channel activity we observe is 194, we cut the histogram at 5 to make some distribution visible. The activity distribution is extremely skewed, i.e., our model produces a sparse code. (C) As in (B), but only for the most active channel (vertical with 2.7 $\frac{cyc}{deg}$ peak sensitivity) to show that each channel is sparsely active. (D) Mean activation over all fixation locations.

tation times. Consequently, only the exponents p, q and possibly the noise strength N_C can be compared between presentation times. Furthermore, the parameters we fitted for ModelFest depend on the data augmentation we used to achieve a good fit from the thresholds only, and the oblique masking data were fit to a considerably different kind of data. Thus, we shall restrict our discussion to the parameter sets fit to the contrast discrimination data at the three presentation times.

For these three presentation times we see that q, which regulates the high contrast behavior of the model, changes little with the presentation time. This corresponds to the empirical statement that the power law behavior at high contrasts has a similar log-log slope for all presentation times. The exponent p changes such that longer presentation times require a lower exponent. This change fits the empirical observation of a less pronounced dip at longer presentation times (see Figure 10). Additionally we can observe that the noise variance N_C decreases with presentation time fitting the absolute decrease in thresholds for longer presentation times. This could be interpreted as averaging away noise over time. However, caused by the different scaling of contrast applied before the decomposition and the different C it is not entirely clear whether this conclusion should be taken seriously based on these data.

Analysis of the models representation

Additional to the theories developed based on psychophysical or neural measurements, researchers developed normative theories to characterize what the information extracted from natural stimulation for animals or humans should be. Our model was not designed to maximize coding efficiency or to fit natural stimuli. Thus, it is interesting to have a more detailed look at what responses to natural stimuli look like and which normative principles our model follows.

As a first qualitative analysis on the model output, we looked at the responses our model produces to natural images. Simply summing the responses from all channels, we found that our model indeed highlights edges. This fits the earliest accounts of the responses of primary visual cortex neurons (Hubel & Wiesel, 1968). As an example, we show the summed response for the example photograph of the Tübingen town hall in Figure 15A. To allow a better display, we show the square root of the sum. Note also that the town hall is easily recognizable from this representation.

Sparseness

To get some more quantitative information about the typical responses of our model, we analyzed the responses of our model to some natural images, for which eye movement data are available from an earlier study (Engbert, Trukenbrod, Barthelme, & Wichmann, 2015). In this study 35 observers explored 15 natural scenes and 15 photographs of texture surfaces for 10 s each to memorize them. During this experiment they produced 24,582 fixations. At each of these fixations we extracted the activity at the fixated pixel from an image we had processed by the model as a whole without the foveal window. This might give us some hint what the internal representation in our model looks like for natural foveal stimulation of human observers.

First, we looked at the range of activations observed and found an extremely skewed distribution (see Figure

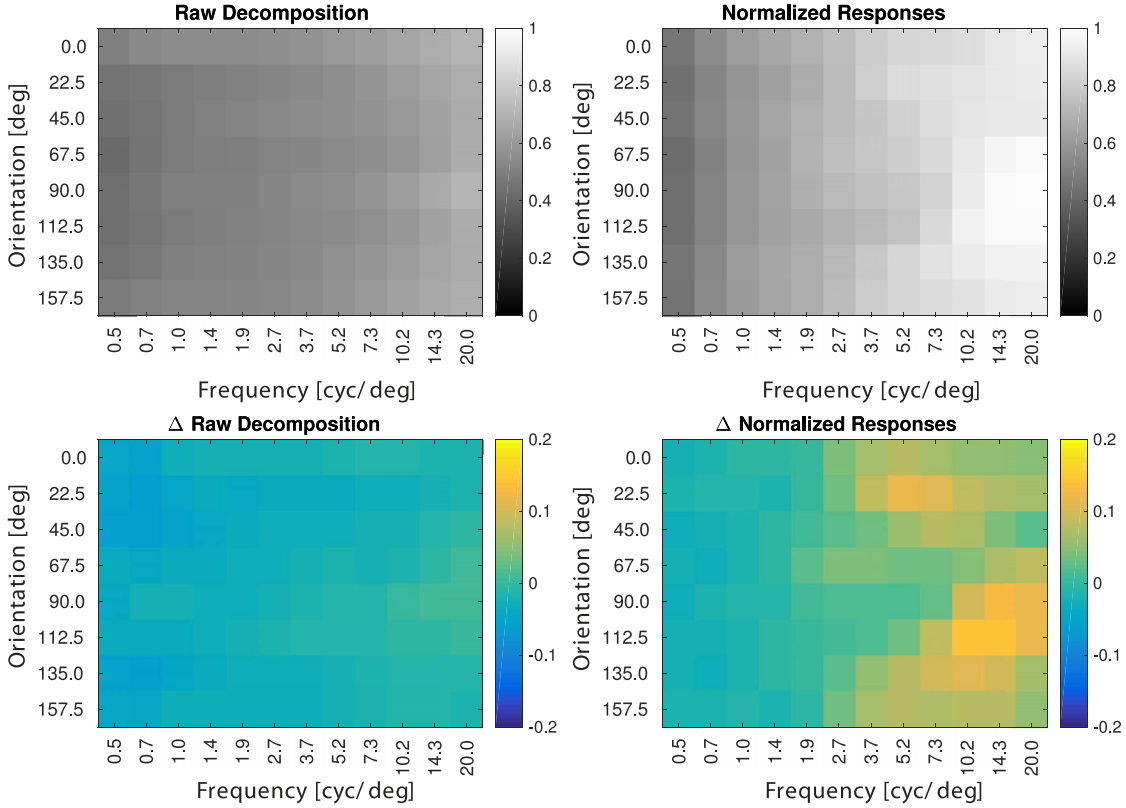

Figure 16. Lifetime sparseness for the different spatial frequency and orientation channels. Left shows the sparseness of the linear filter responses (before nonlinearities and normalization). Right shows the sparseness of the final responses. In the top row we show the sparseness of activities at fixated locations. In the lower row we show the difference between the sparseness at fixated locations and the sparseness at nonfixated control locations.

15B): Maximal activations were almost 200, while 98.7% of the channel activities observed was smaller than 5. This effect is caused by skewed distributions in each channel. To illustrate this distortion we show the activity histogram of the most active channel in Figure 15C. Even this most active channel is rarely active. These observations fit well with theoretical arguments for using a sparse code (Olshausen & Field, 1996) and physiological observations showing sparse neuronal responses (Buzsáki & Mizuseki, 2014).

To quantify the sparsity of the model responses, we used the formula developed first by Rolls and Tovee (1995) and refined and applied to primate primary visual cortex by Vinje and Gallant (2000):

$$S = 1 - \frac{\left(\frac{1}{n}\sum_{i=1}^{n} r_i\right)^2}{\frac{1}{n}\sum_{i=1}^{n} r_i^2} \frac{1}{1-\frac{1}{n}}. \quad (8)$$

S measures the proportion of the sum of squares explained by the mean response and subtracts it from 1. After dividing by $1-\frac{1}{n}$ this yields a measure which conveniently scales from 0 to 1 from a constant response to a perfectly sparse response, which reacts exactly to one stimulus and is 0 for all others. Applying this formula to our model responses, we follow Froudarakis et al. (2014) in separating *population sparseness* (whether the population response to a stimulus is sparse) from *lifetime sparseness* (whether an individual channel is sparsely active over the presentation of all stimuli).

For population sparseness, we find an average value of 33.86% for the raw decomposition and 52.31% for the normalized responses, which is more sparse than average neuronal populations in mouse V1 (mean = 0.26, maximum $\leq$ 0.6) as measured by Froudarakis et al. (2014), but within the range observed. Due to the small numbers of simultaneously recorded neurons in typical primate recordings, we lack data to compare our model to for monkey primary visual cortex.

Investigating lifetime sparseness, we find high values for the sparseness of the channels as displayed in Figure 16. On average the channels after the raw decomposition have $S = 55.07\%$, which increases to an even higher S of 73.85% after normalization. These are both much higher than the lifetime sparseness measured in mouse V1 by Froudarakis et al. (2014), which was 35% on average.

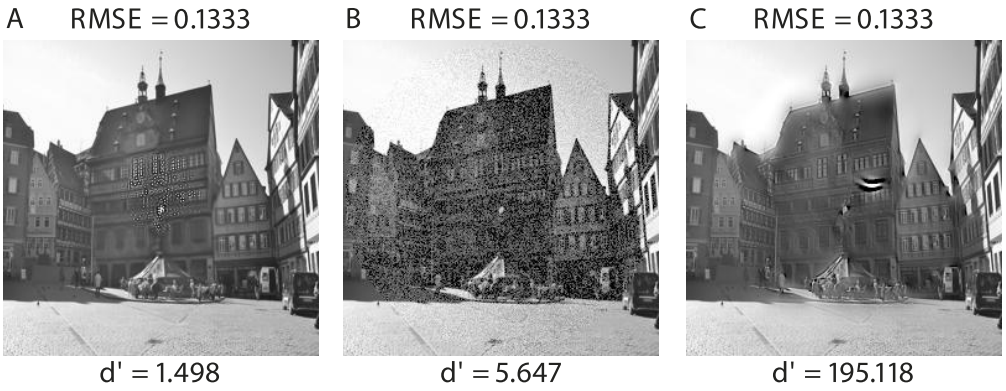

Figure 17. Stimuli with optimized differentiability from the original Tübingen town hall image, with a given RMSE in the windowed contrast image. Luminance images are displayed assuming a gamma of 2.2. In the model these images were simulated to cover 2° × 2° of visual angle. (A) Minimized differentiability, (B) Gaussian noise over the area within the foveal window, and (C) Maximized differentiability.

Furthermore, our model also reproduces the observation that natural stimulation—viewing natural images—elicits a sparser code. Patches extracted around fixated locations yield higher lifetime sparseness in high spatial frequency channels than do control patches, which we extracted at the measured fixation locations, but from different images from the stimulus set (see Figure 16).

We also computed the average activations produced by the channels in our model. The results are displayed in Figure 15D. After the normalization the fall off for higher spatial frequencies inherent in natural images (Field, 1987) is not observed any more. In contrast, the higher content for the cardinal axes (0° and 90° in our notation) persists after the normalization (Furmanski & Engel, 2000; Li, Peterson, & Freeman, 2003). This activation pattern qualitatively fits reasonably well to the distribution of neurons in primary visual cortex, fitting the idea that the distribution of neuronal preferences reflects the distribution of activations produced by natural stimulation (Field, 1987; Laughlin, 1983).

Optimized stimuli

One additional benefit of (successful) image-computable models of human vision is that they should allow the generation of image modifications leading to minimal and/or maximal perceptual differences, exploiting the idea of maximally differentiating (MAD) stimuli (Wang & Simoncelli, 2008). In the following we illustrate the viability of MAD applied to our image-computable spatial vision model, comparing changes in the model responses to the default and simple root mean squared error (RMSE) metric.

For our illustration we optimized the images to be as easy or as hard to differentiate from the image of the Tübingen town hall as possible with a given RMSE after conversion to luminance and application of the foveal window. The exact optimization scheme we applied is described in detail in Appendix C.

In Figure 17 we show three images with equal RMSE from the original Tübingen town hall example image: one with minimized differentiability, one with simple Gaussian noise, and one with maximized differentiability. The optimization clearly produced stimuli which are predicted to be considerably more or less differentiable from the original image but all have the same RMSE.

In the image with maximized differentiability, we can observe two aspects of the model: First, a single, local signal is predicted to be more easily detectable. Second, the optimized signal is similar to the filter shape of a single channel of medium spatial frequency where contrast sensitivity is highest.

In the image with minimized differentiability, the RMSE is realized as a high frequency nonoriented and distributed noise on the image. This indeed becomes practically invisible when viewed such that the image covers the 2° × 2° simulated in the model (around a 1.4 m distance if you printed this paper on A4, such that the images are 5 × 5 cm).

These generated stimuli demonstrate that our model is capable of producing predictions for suprathreshold stimuli and their differences, which are interpretable and testable. This makes our model potentially applicable for image quality assessment, and, as discussed below, allows more thorough tests of the model to be performed.

Discussion

We describe an image-computable model of spatial vision. When applied to classical psychophysical

results, it is consistent with the broad range of contrast detection, discrimination and oblique (orientation) masking data fitted by earlier, more abstractly implemented, nonimage-computable models. In addition, we tested our model on the ModelFest dataset on which it also performs well. Alas, our model—like all previous models—fails to account for human plaid masking data.

While developing our model, we uncovered two crucial ingredients for a successful image-computable spatial vision model: First, when including nonlinear interactions between channels after the decomposition, strong oscillations in the response are observed unless we sample the spatial frequency and orientations axes more densely than required from signal processing considerations. In the human visual system, this problem appears to be solved by not having discrete channels like in engineered subband transforms, but by having a continuous distribution of cells covering the relevant spatial frequencies and orientations.

Second, different temporal presentation modes require—systematically—different parameters of the model: Shorter presentation times require higher exponents for both the signal and the normalization pool, yielding stronger nonlinearities for short presentation times. This finding confirms an earlier conjecture by Wichmann (1999), based on much simpler models, and might explain differences in estimated exponents between different labs and studies. For the parameter q, the difference in the exponent between numerator and denominator—we find little dependence on the presentation time once we assume only a constant noise. Finally, variance of the constant noise decreases with presentation time. All these changes in parameters are consistent with the following picture: Channels show an onset response with stronger nonlinearity, followed by a less nonlinear sustained response. Over time human observers appear to be able to average some of the noise.

When we applied our model to the natural image-masking database by Alam et al. (2014), we found that our model predicted the data reasonably well, but almost always predicted lower thresholds than observed in their experiment. Potential reasons for the discrepancy include the following: First, our optimal decoder knows both signal and mask exactly, which is unlikely to be true for human observers with either stochastic or hitherto unseen natural images as masks. Thus, the overestimation of performance of our model may in part be due to our too knowledgeable decoder. Second, our model is solely fit to data from very experienced psychophysical observers, and we do not know about the experience of the observers in the natural image-masking study (c.f. Jäkel & Wichmann, 2006).

To investigate whether our model conforms to normative notions derived from efficient coding, we analyzed its response to natural images at positions foveated by human observers. We found our model to be sparse as expected from theoretical considerations. Furthermore, average responses still contain a bias for cardinal orientations as observed in natural images, but the 1/f decline over spatial frequency associated with natural images is obliterated by normalization.

Finally we created MAD stimuli to compare our model to the RMSE. These stimuli illustrate the behavior of our model. Additionally, such stimuli might be used to psychophysically test our model in the future, which is the intended purpose of MAD stimuli (Wang & Simoncelli, 2008). Especially once one wants to test different, more complex models against each other, analyses like this are invaluable.

Comparison to earlier models

As we specifically designed our model to be an image-computable version of the standard spatial vision model, it naturally shares many properties with earlier models and implementations.

The model by Foley (1994) first introduced the spatial frequency and orientation channel decomposition followed by divisive normalization, which is at the heart of our model. However, Foley implemented decoding as a Minkowski norm of the difference between responses instead of explicitly modelling noise and optimal factorial decoding as we do here. Another model using the simpler Minkowski norm decoding scheme is the model by Itti et al. (2000). This model is also an important precursor of our model, as it showed that different tasks like spatial frequency and orientation discrimination could be explained by a *single* model of the style we use here. Finally, the most closely related abstract, nonimage-computable model is the model by Goris et al. (2013). The remaining conceptual differences of our model to the Goris model are, on the one hand, that we did not include noise correlations or adaptation present in the Goris model, but, on the other hand, we added the spatial extend of the normalization pool, orientation, etc., to move our model from 1D to 2D.

Of the few image-based spatial vision models, the two most closely related ones to ours are the models by Teo and Heeger (1994) and by Watson and Solomon (1997), which both implement a spatial frequency decomposition and divisive normalization. However, they use the simplistic Minkowski norm decoding and were implemented with the technology of their time, which made diverse compromises for speed necessary. For example, the Watson and Solomon (1997) model represented only three spatial frequency channels of

which one always hit the spatial frequency of the target. Also both models were compared to a rather small range of data and some views and questions like natural scene statistics and optimal coding were not yet discussed at the time of these models.

Most other image-computable models of spatial vision do not aim to mimic the internal processes involved in spatial visual processing, but simply optimize prediction with computationally less demanding processes. Especially the most modern models of this kind (Bradley et al., 2014) predict human performance quite well and can even include peripheral limitations. However, these models are designed for different purposes than our model, providing no output similar to the output of the first steps of the human visual system and allow no direct tests of hypotheses about the early visual processing either.

Potentially controversial details

Phase invariance

Our model provides phase invariant output, which represents the information perfect complex cells would convey. This is computationally efficient and provides all information necessary for the psychophysical tasks we model. Additionally, this output nicely fits with other psychophysical data which explicitly shows phase independence for the detection of multiple sufficiently separate components (Graham & Nachmias, 1971) and that phase perception can be explained based on detection of local contrast changes (Badcock, 1984, 1988). However, neuronal data show that the distinction between simple and complex cells is gradual, and both types express some sensitivity to relative phase (Mechler, Reich, & Victor, 2002). Furthermore humans show more dependence on phase information for object recognition than predicted from contrast reduction caused by phase noise (Wichmann, Braun, & Gegenfurtner, 2006). Consequently, a more complete model might add decoding from phase dependent output to mimic simple cells, or even include both simple and complex cells.

Tuning and complexity of the normalization pool

The spatial vision community is divided whether the normalization pool is orientation specific. In our purely divisive normalization implementation—without a subtractive normalization—orientation specific normalization is required to be consistent with our data; the same is true for the model by Itti et al. (2000). The models by Foley (1994) and by Teo and Heeger (1994) argue for an orientation unspecific normalization, in line with neurophysiology (Heeger, 1992).

In our data and the data of Itti et al. (2000), orthogonal gratings barely mask each other, even at high mask contrasts—thus a nontuned divisive normalization does not fit such data. In the data by Foley (1994), however, orthogonal gratings mask the signal grating. Similarly physiologists sometimes find that orthogonal gratings considerably attenuate neuronal responses (Heeger, 1992)—cross-orientation inhibition. However, at least the suppressive surround is sometimes found to be tuned (Cavanaugh, Bair, & Movshon, 2002b). One possible explanation for this discrepancy *in the data* is the temporal presentation of the stimuli during the experiments. Our data and the Itti et al. (2000) data were collected using static gratings presented for an extended period of time, whereas the data of Foley (1994) and Foley and Boynton (1994) were collected using very short presentation times. Thus, the normalization pool may initially be broadly tuned, but narrows during prolonged presentation.

Furthermore, the normalization we implemented does not cover all interactions reported between channels. There are well known facilitatory effects of collinear flankers (Polat & Sagi, 1993). Most commonly these are interpreted as facilitatory effects between channels, but alternatively these could be explained by collector units further up in the hierarchy of visual processing (Solomon & Morgan, 2000). Similar ideas were also proposed to explain the unexpectedly strong masking produced by amplitude modulated gratings at their modulation frequency (Henning, Hertz, & Broadbent, 1975). Such explanations based on further processing of the filter responses are compatible with our model being a correct model of the first transformations in spatial vision. If the interpretation as facilitatory effects in the earliest representation is correct, however, it should be included in future spatial vision models.

High contrast signals

Another aspect differing between models is how they treat high contrast signals. In our model we implement a higher numerator exponent in the normalization, which yields nonsaturating responses in the individual channels as in the model by Watson and Solomon (1997). The alternative approach followed by Teo and Heeger (1994) is to simulate multiple types of channels covering different contrast ranges (in their case four). This second approach models the responses in closer agreement to neuronal data, as neurons undeniably saturate. From a psychophysical perspective this seems to add little, however, as channels differing only in their absolute sensitivity cannot be targeted specifically by any stimuli and are thus modelled quite adequately as a single channel. Only if the cells or channels for higher contrasts had different tuning curves or interactions

with other channels than the low contrast ones, it would be necessary to separate them. Consequently, we interpret the V1 neurons for different contrast levels as the neuronal implementation of a single channel using multiple neurons to avoid saturation.

Decoding stage

In our decoding ideas we follow modern abstract models like the Goris et al. (2013) model and explicitly model the noise on individual channels and propose optimal or near optimal readout of the channel responses in a Bayesian sense (c.f. Beck et al., 2008; Ma, Shen, Dziugaite, & van den Berg, 2015). The idea that observers in basic psychophysical tasks are (only) limited by an internal noise source has recently been challenged. Beck, Ma, Pitkow, Latham, and Pouget (2012) instead propose that performance is limited by imperfections of the readout mechanism. For explaining the systematic discrepancy between our model and natural image database data by Alam et al. (2014), we follow this interpretation. It appears that (highly experienced) observers during simple contrast detection and discrimination experiments were more sensitive than subjects producing the natural image-masking data. We suggest that this might be caused by better decoding rather than more available information, similar to the suggestion that perceptual learning improves decoding rather than the original representation (Diaz, Queirazza, & Philiastides, 2017). In our model the decoding is optimal for the classical grating experiments, as these experiments are set up to make decoding as easy as possible for humans.

Variance and type of internal noise

The noise model used in early spatial vision models has always been a matter of discussion, partly because the psychophysical data collected during classical detection and discrimination tasks appear not to constrain the standard model sufficiently. Even fundamental questions as whether the noise variance changes with signal strength were not finally answered by psychophysics yet, although some attempts were made (Georgeson & Meese, 2006; Kingdom, 2016; Kontsevich, Chen, & Tyler, 2002; Wichmann, 1999). Based on our maximum-likelihood estimation we cannot, unfortunately, answer the question whether the noise grows with the signal or not. Our model can explain the data with constant noise equally well as with noise variance growing linearly with the signal. The underlying reason for this is that changing the q-parameter can compensate for a growing noise. In terms of the neural implementation this corresponds to the statement that adding more neurons tuned to high contrasts can compensate for neurons being noisier when responding strongly. This insight explains why we cannot differentiate how the noise should change with increasing contrast based on psychophysics—at least not based on the data currently available. Also it might serve as a reminder that the nonlinearity we employ in our psychophysical model does not directly map to the nonlinearity of neurons, although they use the same basic form.

If the connection to neuronal processing was closer, we could use the typically employed noise forms from physiology. In physiology, noise is typically modelled as Poisson noise or variations of it with different factors between mean response and variance, or with additional variance shared between units (Goris, Movshon, & Simoncelli, 2014). For our model, however, it is unclear how many neurons a channel response at a single pixel represents, and on which level of the model the noise relevant for a task is induced. Thus we believe that modelling the noise as Gaussian is warranted for simplicity.

We include no noise correlations in our model—it was simply unnecessary to add this additional "complication" in order to fit our psychophysical data. Including noise correlations in our model is computationally far from trivial, caused by the sheer number of activities which could be correlated. Furthermore, having to decide which channel responses should be correlated would add many additional degrees of freedom not constrained by psychophysical data. This does not argue against noise correlations, of course, but only that adding more uncorrelated noise adequately mimics the effects of these correlations for our purposes.

Processing heterogeneity

Like all previous spatial vision models—image-computable or not—we did not model the diversity of V1 neurons (and, presumably, psychophysical channels). For computational efficiency all the channels in our model have the same bandwidths, i.e., all neurons have the same receptive field, scaled and rotated to adjust their preferred spatial frequency and orientation. In contrast, V1 neurons have diverse bandwidths (De Valois, Albrecht, & Thorell, 1982; Ringach et al., 2002), which seems to be adaptive for natural scenes (Goris et al., 2015). Also all channels in our model cover the image with constant and equal density, although in truth V1 neurons seem to be sparser and the number of neurons differs between different spatial frequencies and orientations, which manifests itself in the psychophysical oblique effects (Furmanski & Engel, 2000; Li et al., 2003). Changing the density of neurons might be adaptive to concentrate resources on frequent stimuli and to implicitly represent the prior distribution over stimuli (Laughlin, 1983). However, as for the simple

Gaussian noise approximation discussed above, our simplified model appears sufficient to capture human behavior in response to classical psychophysically employed stimuli.

Limitations of the presented model

Although we tried to closely represent the concepts realized in classical spatial vision models, we did not include all ideas in our model for computational simplicity. Phrased negatively, we excluded substantial areas of spatial vision, as we discuss below.

Temporal dynamics, color, and stereo

We restrict our model to static, gray-scale luminance images projected onto a single cyclopean fovea. This excludes any kind of temporal changes beyond the very coarse separation by presentation duration we made in our model. A true processing of stimuli over time would go beyond our current computational capabilities. Nonetheless, it is worth highlighting that temporal processing was investigated and seems to be explainable by two or maximally three temporal channels (Watson, 1986; Watson & Nachmias, 1977). However, we are not aware of a combination of these models for temporal processing with masking or discrimination models. Furthermore, luminance images exclude color processing, which requires considerably more complex models of the optics to include chromatic aberrations (Bedford & Wyszecki, 1957; Charman & Jennings, 1976) and of the retinal sampling, adaptation and processing, which differ between color channels (Brainard, 2015). Additionally, cortical processing of color is understood less completely (Gegenfurtner, 2003). Finally, luminance images contain no depth information, which relieves us from explicitly modelling 3D scenes, the optical effects on objects outside the focal plane and binocular vision. Modelling binocular vision is possible, but results in considerably more complex psychophysical models (Baker, Meese, & Georgeson, 2007; Georgeson, Wallis, Meese, & Baker, 2016; Legge, 1984a, 1984b; Meese, Georgeson, & Baker, 2006). The additional complexity arises, because human observers do not only non-trivially combine the binocular input into one combined image, but can also perceive disparity (spatial shift between eyes) and luster (contrast differences between eyes). Under dichoptic presentation, these additional channels can lead to interesting unintuitive results (e.g., May & Zhaoping, 2016).

Adaptation

Our model includes no adaptation effects yet. This means that some classical psychophysical datasets are not within the scope of our model (Blakemore & Campbell, 1969, for example). Some abstract models (Foley & Chen, 1997; Goris et al., 2013; Meese & Holmes, 2002, for example) contain adaptation and discuss which parts of the model adapt to what kind of stimuli. However, adaptation would at least require additional input besides the stimuli to be discriminated and depends considerably on the duration of the adaptation stimulus and the interval between adaptation and test stimuli. Thus adaptation in our image-based model would require substantial additional work, and would perhaps best be tackled after a model with adequate temporal dynamics exists.

Peripheral vision

We restrict ourselves to a purely foveal model, and thus to a model with uniform processing and sensitivity. Peripheral vision differs from foveal vision already in the optical quality (Jennings & Charman, 1981; Navarro, Williams, & Artal, 1993; Williams, Artal, Navarro, McMahon, & Brainard, 1996) and retinal processing—at least by the sampling density (Curcio, Sloan, Packer, Hendrickson, & Kalina, 1987; Curcio & Allen, 1990). Additionally, the interactions between channels, which we model in our normalization step, are different in the periphery (Xing & Heeger, 2000). More generally, higher level restrictions like crowding (Whitney & Levi, 2011) play a larger role in the periphery, presumably due the stronger information reduction and the faster growth in peripheral receptive field size (Gattass, Sousa, & Gross, 1988; Rosenholtz, 2016). Hence, a detailed modelling of the periphery would require a considerable effort beyond our current model.

Additional tasks

In this paper we evaluate our model exclusively on discrimination data, and we cover a broad range of psychophysical data, but, of course, not all of it. Obvious omissions are data from direct estimation tasks ("What was the orientation of the grating?") as well as classification tasks ("Was the grating tilted left or right?"), because our model cannot deal with data from such tasks in its present form. Clearly, such tasks are important and have been used to investigate models of early visual processing (Meng & Qian, 2005; Solomon, Felisberti, & Morgan, 2004). Such tasks could be implemented as a different type of decoding based on the model representation, an avenue we are planning to pursue. To explain biases in human perception explanations of these effects might require the inclusion of prior beliefs about the categories (Girshick, Landy, & Simoncelli, 2011) or deviations from optimal decoding.

Applications in and beyond spatial vision

On the one hand we hope to facilitate investigations into the details of spatial visual processing, using our model as a starting point or basis. Further developments are still necessary, not least to address the limitations and controversial design choice we discuss above. To do this, image-computable models can be applied to a much wider range of existing data and allow the generation of optimized stimuli to differentiate different models, as we demonstrated in the Optimized stimuli section. In addition, image-computable models allow direct comparisons to normative theories, as we have started on a small scale in this paper. Whatever normative ideas might arise in the future, it can be assessed whether our spatial vision model optimizes the proposed measures.

On the other hand, going beyond early spatial visual processing, our model might help with the development of mechanistic models of mid- or high-level visual processing by providing a psychophysically sound basis in which to represent images beyond pixels. This tool, we conjecture, might improve the match between mid- and high-level vision models and human perception. One clear target for such endeavors are convolutional DNNs in object recognition (Kriegeskorte, 2015; LeCun, Bengio, & Hinton, 2015; Yamins & DiCarlo, 2016).

Finally a working spatial vision model might have practical applications as an image quality metric as was the original intention of Teo and Heeger (1994). Later image quality metrics like the structural similarity metric (SSIM, Wang, Simoncelli, & Bovik, 2003; Wang, Bovik, Sheikh, & Simoncelli, 2004) claim to go beyond error visibility, but arguably getting error visibility right would be a good start as well. As it is currently demonstrated for the Normalized Laplacian Pyramid (Laparra et al., 2016), such image quality metrics can then be used to optimize the display of images to make it match the perception of the original (Laparra et al., 2017).

Keywords: model, spatial vision, image-computable, psychophysics

Acknowledgments

We would like to thank David Janssen for fruitful discussion, and Ralf Engbert, Lars Rothkegel, and Hans Trukenbrod for their feedback on our manuscript. Additionally we would like to thank Joshua Solomon and an anonymous reviewer for their helpful reviews. This work was funded, in part, by the German Federal Ministry of Education and Research (BMBF) through the Bernstein Computational Neuroscience Program Tübingen (FKZ: 01GQ1002), and the Deutsche Forschungsgemeinschaft through a grant to F. A. W. (grant WI 2103/4-1).

Commercial relationships: none.
Corresponding author: Heiko H. Schütt.
Email: heiko.schuett@uni-tuebingen.de.
Address: Neural Information Processing Group, University of Tübingen, Tübingen, Germany.

Footnotes

[1] Technically, we should use a von Mises distribution for orientation, which wraps the tails of the normal distribution around as orientation is a circular dimension. However, as the normalization pool we find is narrow, the difference between a Gaussian and the von Mises distribution is negligible.

[2] 0 represents normalization exclusively by channels with the same orientation, and ∞ represents equal weighting of all orientations.

References

Alam, M. M., Patil, P., Hagan, M. T., & Chandler, D. M. (2015). A computational model for predicting local distortion visibility via convolutional neural network trained on natural scenes. In *2015 IEEE international conference on image processing* (pp. 3967–3971). Quebec City, Canada: IEEE.

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, doi:10.1167/14.8.22. [PubMed] [Article]

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*, 183–193.

Badcock, D. R. (1984). Spatial phase or luminance profile discrimination? *Vision Research*, *24*(6), 613–623.

Badcock, D. R. (1988). Discrimination of spatial phase changes: Contrast and position codes. *Spatial Vision*, *3*(4), 305–322.

Baker, D. H., Meese, T. S., & Georgeson, M. A. (2007). Binocular interaction: Contrast matching and contrast discrimination are predicted by the same model. *Spatial Vision*, *20*(5), 397–413.

Baldwin, A. S., Meese, T. S., & Baker, D. H. (2012). The attenuation surface for contrast sensitivity has the form of a witch's hat within the central visual

field. *Journal of Vision*, *12*(11):23, 1–17, doi:10.1167/12.11.23. [PubMed] [Article]

Barlow, H. B. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. In *NPL symposium on the mechanization of thought process, No. 10* (pp. 535–559). London, UK: HM Stationary Office.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J.,... Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. *Neuron*, *60*(6), 1142–1152.

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not noisy, Just wrong: The role of suboptimal inference in behavioral variability. *Neuron*, *74*(1), 30–39.

Bedford, R. E., & Wyszecki, G. (1957). Axial chromatic aberration of the human eye. *JOSA*, *47*(6), 564–565.

Bird, C. M., Henning, G. B., & Wichmann, F. A. (2002). Contrast discrimination with sinusoidal gratings of different spatial frequency. *JOSA A*, *19*(7), 1267–1273.

Blakemore, C., & Campbell, F. W. (1969). On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images. *The Journal of Physiology*, *203*(1), 237–260.

Bradley, C., Abrams, J., & Geisler, W. S. (2014). Retina-Vl model of detectability across the visual field. *Journal of Vision*, *14*(12):22, 1–22, doi:10.1167/14.12.22. [PubMed] [Article]

Brainard, D. H. (2015). Color and the cone mosaic. *Annual Review of Vision Science*, *1*(1), 519–546.

Buzsáki, G., & Mizuseki, K. (2014). The log-dynamic brain: How skewed distributions affect network operations. *Nature Reviews Neuroscience*, *15*(4), 264–278.

Campbell, F. W., Cleland, B. G., Cooper, G. F., & Enroth-Cugell, C. (1968). The angular selectivity of visual cortical cells to moving gratings. *The Journal of Physiology*, *198*(1), 237–250.

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

Campbell, F. W., & Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. *The Journal of Physiology*, *197*(3), 551.

Carandini, M., & Heeger, D. J. (2011). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*, 51–62.

Carandini, M., & Ringach, D. L. (1997). Predictions of a recurrent model of orientation selectivity. *Vision Research*, *37*(21), 3061–3071.

Carter, B. E., & Henning, G. B. (1971). The detection of gratings in narrow-band visual noise. *Journal of Physiology*, *219*(2), 355–365.

Cavanaugh, J. R., Bair, W., & Movshon, J. A. (2002a). Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *Journal of Neurophysiology*, *88*(5), 2530–2546.

Cavanaugh, J. R., Bair, W., & Movshon, J. A. (2002b). Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons. *Journal of Neurophysiology*, *88*(5), 2547–2556.

Charman, W. N., & Jennings, J. A. M. (1976). Objective measurements of the longitudinal chromatic aberration of the human eye. *Vision Research*, *16*(9), 999–1005.

Curcio, C. A., & Allen, K. A. (1990). Topography of ganglion cells in human retina. *The Journal of Comparative Neurology*, *300*(1), 5–25.

Curcio, C. A., Sloan, K. R., Packer, O., Hendrickson, A. E., & Kalina, R. E. (1987). Distribution of cones in human and monkey retina: Individual variability and radial asymmetry. *Science*, *236*(4801), 579–582.

Daugman, J. G. (1980). Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, *20*(10), 847–856.

Derrington, A. M., & Henning, G. B. (1989). Some observations on the masking effects of two-dimensional stimuli. *Vision Research*, *29*(2), 241–246.

De Valois, R. L., Albrecht, D. G., & Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research*, *22*(5), 545–559.

Diaz, J. A., Queirazza, F., & Philiastides, M. G. (2017). Perceptual learning alters post-sensory processing in human decision-making. *Nature Human Behaviour*, *1*(2), 35.

Engbert, R., Trukenbrod, H. A., Barthelme, S., & Wichmann, F. A. (2015). Spatial statistics and attentional dynamics in scene viewing. *Journal of Vision*, *15*(1):14, 1–17, doi:10.1167/15.1.14. [PubMed] [Article]

Fechner, G. T. (1860). *Elemente der psychophysik* [Translation: *Elements of psychophysics*]. Leipzig: Breitkopf und Härtel.

Field, D.J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A*, *4*(12), 2379–2394.

Foley, J. M. (1994). Human luminance pattern-vision

mechanisms: Masking experiments require a new model. *JOSA A*, *11*(6), 1710–1719.

Foley, J. M., & Boynton, G. M. (1994). New model of human luminance pattern vision mechanisms: Analysis of the effects of pattern orientation, spatial phase, and temporal frequency. In *Proceedings of the SPIE: Computational Vision Based on Neurobiology* (Vol. 2054, pp. 32–42). Bellingham, WA: SPIE.

Foley, J. M., & Chen, C.-C. (1997). Analysis of the effect of pattern adaptation on pattern pedestal effects: A two-process model. *Vision Research*, *37*(19), 2779–2788.

Foley, J. M., & Legge, G. E. (1981). Contrast detection and near-threshold discrimination in human vision. *Vision Research*, *21*(7), 1041–1053.

Freeman, W. T., & Adelson, E. H. (1991). The design and use of steerable filters. *IEEE Transactions on Pattern analysis and machine intelligence*, *13*(9), 891–906.

Froudarakis, E., Berens, P., Ecker, A. S., Cotton, R. J., Sinz, F. H., Yatsenko, D.,... Tolias, A. S. (2014). Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. *Nature Neuroscience*, *17*(6), 851–857.

Furmanski, C. S., & Engel, S. A. (2000). An oblique effect in human primary visual cortex. *Nature Neuroscience*, *3*(6), 535–536.

Gattass, R., Sousa, A. P., & Gross, C. G. (1988). Visuotopic organization and extent of V3 and V4 of the macaque. *Journal of Neuroscience*, *8*(6), 1831–1845.

Gegenfurtner, K. R. (2003). Cortical mechanisms of colour vision. *Nature Reviews Neuroscience*, *4*(7), 563–572.

Geisler, W. S., & Albrecht, D. G. (1995). Bayesian analysis of identification performance in monkey visual cortex: Nonlinear mechanisms and stimulus certainty. *Vision Research*, *35*(19), 2723–2730.

Georgeson, M. A., & Meese, T. S. (1997). Perception of stationary plaids: The role of spatial filters in edge analysis. *Vision Research*, *37*(23), 3255–3271.

Georgeson, M. A., & Meese, T. S. (2006). Fixed or variable noise in contrast discrimination? The jury's still out.... *Vision Research*, *46*(25), 4294–4303.

Georgeson, M. A., Wallis, S. A., Meese, T. S., & Baker, D. H. (2016). Contrast and lustre: A model that accounts for eleven different forms of contrast discrimination in binocular vision. *Vision Research*, *129*, 98–118.

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932.

Goris, R. L. T., Movshon, J. A., & Simoncelli, E. P. (2014). Partitioning neuronal variability. *Nature Neuroscience*, *17*(6), 858–865.

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

Goris, R. L. T., Simoncelli, E. P., & Movshon, J. A. (2015). Origin and function of tuning diversity in macaque visual cortex. *Neuron*, *88*(4), 819–831.

Graham, N. (1989). *Visual pattern analyzers*. Oxford, UK: Oxford University Press.

Graham, N., & Nachmias, J. (1971). Detection of grating patterns containing two spatial frequencies: A comparison of single-channel and multiple-channels models. *Vision Research*, *11*(3), 251–259.

Hahn, L. W., & Geisler, W. S. (1995). Adaptation mechanisms in spatial vision-i. Bleaches and backgrounds. *Vision Research*, *35*(11), 1585–1594.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*(2), 181–197.

Henning, G. B., Hertz, B. G., & Broadbent, D. E. (1975). Some experiments bearing on the hypothesis that the visual system analyses spatial patterns in independent bands of spatial frequency. *Vision Research*, *15*(8–9), 887–897.

Holmes, D. J., & Meese, T. S. (2004). Grating and plaid masks indicate linear summation in a contrast gain pool. *Journal of Vision*, *4*(12):7, 1080–1089, doi:10.1167/4.12.7. [PubMed] [Article]

Hood, D. C. (1998). Lower-Level Visual Processing and Models of Light Adaptation. *Annual Review of Psychology*, *49*(1), 503–535.

Hood, D. C., & Finkelstein, M. (1986). Sensitivity to light. In *Handbook of perception and human performance. Vol. 1: Sensory processes and perception*. New York, NY: John Wiley and Sons.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243.

Itti, L., Koch, C., & Braun, J. (2000). Revisiting spatial vision: Toward a unifying model. *JOSA A*, *17*(11), 1899–1917.

Jennings, J. A. M., & Charman, W. N. (1981). Off-axis image quality in the human eye. *Vision Research*, *21*(4), 445–455.

Jäkel, F., & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal*

of Vision, *6*(11):13, 1307–1322, doi:10.1167/6.11.13. [PubMed] [Article]

Kelly, D. H. (1979). Motion and vision II stabilized spatio-temporal threshold surface. *Journal of the Optical Society of America*, *69*(10), 1340.

Kingdom, F. A. (2016). Fixed versus variable internal noise in contrast transduction: The significance of Whittle's data. *Vision Research*, *128*, 1–5.

Kontsevich, L. L., Chen, C.-C., & Tyler, C. W. (2002). Separating the effects of response nonlinearity and internal noise psychophysically. *Vision Research*, *42*(14), 1771–1784.

Kortum, P. T., & Geisler, W. S. (1995). Adaptation mechanisms in spatial vision-ii. Flash thresholds and background adaptation. *Vision Research*, *35*(11), 1595–1609.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Laparra, V., Ballé, J., Berardino, A., & Simoncelli, E. P. (2016). Perceptual image quality assessment using a normalized Laplacian pyramid. *Electronic Imaging*, *2016*(16), 1–6.

Laparra, V., Berardino, A., Ballé, J., & Simoncelli, E. P. (2017). Perceptually optimized image rendering. *Journal of the Optical Society of America* A, *34*(9), 1151–1525.

Laughlin, S. (1983). Matching coding to scenes to enhance efficiency. In D. O. J. Braddick & A. C. Sleigh (Eds.), *Physical and biological processing of images* (pp. 42–52). Berlin and Heidelberg, Germany: Springer.

LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, *521*, 436–444.

Legge, G. E. (1984a). Binocular contrast summation—I. Detection and discrimination. *Vision Research*, *24*(4), 373–383.

Legge, G. E. (1984b). Binocular contrast summation—II. Quadratic summation. *Vision Research*, *24*(4), 385–394.

Legge, G. E., & Foley, J. M. (1980). Contrast masking in human vision. *JOSA*, *70*(12), 1458–1471.

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

Li, B., Peterson, M. R., & Freeman, R. D. (2003). Oblique effect: A neural basis in the visual cortex. *Journal of Neurophysiology*, *90*(1), 204–217.

Ma, W. J., Shen, S., Dziugaite, G., & van den Berg, R. (2015). Requiem for the max rule? *Vision Research*, *116*, 179–193.

May, K. A., & Solomon, J. A. (2015a). Connecting psychophysical performance to neuronal response properties I: Discrimination of suprathreshold stimuli. *Journal of Vision*, *15*(6):8, 1–26, doi:10.1167/15.6.8. [PubMed] [Article]

May, K. A., & Solomon, J. A. (2015b). Connecting psychophysical performance to neuronal response properties II: Contrast decoding and detection. *Journal of Vision*, *15*(6):9, 1–21, doi:10.1167/15.6.9. [PubMed] [Article]

May, K. A., & Zhaoping, L. (2016). Efficient coding theory predicts a tilt aftereffect from viewing untilted patterns. *Current Biology*, *26*(12), 1571–1576.

Mechler, F., Reich, D. S., & Victor, J. D. (2002). Detection and discrimination of relative spatial phase by V1 neurons. *Journal of Neuroscience*, *22*(14), 6129–6157.

Meese, T. S., Georgeson, M. A., & Baker, D. H. (2006). Binocular contrast vision at and above threshold. *Journal of Vision*, *6*(11):7, 1224–1243, doi:10.1167/6.11.7. [PubMed] [Article]

Meese, T. S., & Holmes, D. J. (2002). Adaptation and gain pool summation: Alternative models and masking data. *Vision Research*, *42*(9), 1113–1125.

Meng, X., & Qian, N. (2005). The oblique effect depends on perceived, rather than physical, orientation and direction. *Vision Research*, *45*(27), 3402–3413.

Nachmias, J., & Sansbury, R. V. (1974). Grating contrast: Discrimination may be better than detection. *Vision Research*, *14*(10), 1039–1042.

Naka, K. I., & Rushton, W. A. H. (1966). S-potentials from colour units in the retina of fish (Cyprinidae). *The Journal of Physiology*, *185*(3), 536.

Navarro, R., Williams, D. R., & Artal, P. (1993). Modulation transfer of the human eye as a function of retinal eccentricity. *JOSA A*, *10*(2), 201–212.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *JOSA A*, *2*(9), 1508–1532.

Phillips, G. C., & Wilson, H. R. (1984). Orientation bandwidths of spatial mechanisms measured by masking. *JOSA A*, *1*(2), 226–232.

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

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

Ringach, D. L., Shapley, R. M., & Hawken, M. J. (2002). Orientation selectivity in macaque V1: Diversity and laminar dependence. *The Journal of Neuroscience*, *22*(13), 5639–5651.

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

Rolls, E. T., & Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology*, *73*(2), 713–726.

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Review of Vision Science*, *2*(1), 437–457.

Rovamo, J., Luntinen, O., & Näsänen, R. (1993). Modelling the dependence of contrast sensitivity on grating area and spatial frequency. *Vision Research*, *33*(18), 2773–2788.

Rovamo, J., Mustonen, J., & Näsänen, R. (1994). Modelling contrast sensitivity as a function of retinal illuminance and grating area. *Vision Research*, *34*(10), 1301–1314.

Rovamo, J., & Virsu, V. (1979a). An estimation and application of the human cortical magnification factor. *Experimental Brain Research*, *37*(3), 495–510.

Rovamo, J., & Virsu, V. (1979b). An estimation and application of the human cortical magnification factor. *Experimental Brain Research*, *37*(3), 495–510.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience*, *4*(8), 819–825.

Schütt, H. H., Harmeling, S., Macke, J. H., & Wichmann, F. A. (2016). Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, *122*, 105–123.

Sharpe, L. T., Stockman, A., Jagla, W., & Jägle, H. (2005). A luminous efficiency function, V*(λ), for daylight adaptation. *Journal of Vision*, *5*(11):3, 948–968, doi:10.1167/5.11.3. [PubMed] [Article]

Simoncelli, E. P., Freeman, W. T., Adelson, E. H., & Heeger, D. J. (1992). Shiftable multiscale transforms. *IEEE Transactions on Information Theory*, *38*(2), 587–607.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural images statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193–1215.

Snowden, R. J., & Hammett, S. T. (1998). The effects of surround contrast on contrast thresholds, perceived contrast and contrast discrimination. *Vision Research*, *38*(13), 1935–1945.

Solomon, J. A., Felisberti, F. M., & Morgan, M. J. (2004). Crowding and the tilt illusion: Toward a unified account. *Journal of Vision*, *4*(6):9, 500–508, doi:10.1167/4.6.9. [PubMed] [Article]

Solomon, J. A., & Morgan, M. J. (2000). Facilitation from collinear flanks is cancelled by non-collinear flanks. *Vision Research*, *40*(3), 279–286.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153.

Strang, G., & Nguyen, T. (1996). *Wavelets and filter banks* (2nd ed.). Wellesley, MA: Wellesley-Cambridge Press.

Teo, P. C., & Heeger, D. J. (1994). Perceptual image distortion. In *IS&T/SPIE 1994 international symposium on electronic imaging: Science and technology* (pp. 127–141). Bellingham, WA: SPIE.

Thibos, L. N., Hong, X., Bradley, A., & Cheng, X. (2002). Statistical variation of aberration structure and image quality in a normal population of healthy eyes. *Journal of the Optical Society of America A*, *19*(12), 2329–2348.

Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, *287*(5456), 1273–1276.

Virsu, V., & Rovamo, J. (1979). Visual resolution, contrast sensitivity, and the cortical magnification factor. *Experimental Brain Research*, *37*(3), 475–494.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), 600–612.

Wang, Z., & Simoncelli, E. P. (2008). Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, *8*(12):8, 1–13, doi:10.1167/8.12.8. [PubMed] [Article]

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *Conference record of the thirty-seventh Asilomar Conference on signals, systems and computers, 2004*. (Vol. 2, pp. 1398–1402). Pacific Grove, CA: IEEE.

Watson, A. B. (1986). Temporal sensitivity. In K. R. Boff, L. Kaufmann, & J. P. Thomas (Eds.),

Handbook of perception and human performance (Vol. 1, pp. 6-1–6-43). New York: Wiley.

Watson, A. B. (1987). The cortex transform: Rapid computation of simulated neural images. *Computer Vision, Graphics, and Image Processing*, *39*(3), 311–327.

Watson, A. B. (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, doi: 10.1167/13.6.18. [PubMed] [Article]

Watson, A. B., & Ahumada, A. J. (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision*, *5*(9):6, 717–740, doi:10.1167/5.9.6. [PubMed] [Article]

Watson, A. B., Borthwick, R., & Taylor, M. (1997). Image quality and entropy masking. In *Electronic Imaging '97* (pp. 2–12). Bellingham, WA: SPIE.

Watson, A. B., & Nachmias, J. (1977). Patterns of temporal interaction in the detection of gratings. *Vision Research*, *17*(8), 893–902.

Watson, A. B., & Solomon, J. A. (1997). Model of visual contrast gain control and pattern masking. *JOSA A*, *14*(9), 2379–2391.

Watson, A. B., & Yellott, J. I. (2012). A unified formula for light-adapted pupil size. *Journal of Vision*, *12*(10):12, 1–16, doi:10.1167/12.10.12. [PubMed] [Article]

Weber, E. H. (1834). *De Pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae* [Translation: *Of pulsing, resorption, hearing and touch the anatomical and physiological notes*]. Leipzig: C.F. Koehler.

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

Wichmann, F. A. (1999). *Some aspects of modelling human spatial vision: Contrast discrimination*. Unpublished doctoral dissertation, University of Oxford, Oxford, UK.

Wichmann, F. A., Braun, D. I., & Gegenfurtner, K. R. (2006). Phase noise and the classification of natural images. *Vision Research*, *46*(8–9), 1520–1529.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313.

Williams, D. R., Artal, P., Navarro, R., McMahon, M. J., & Brainard, D. H. (1996). Off-axis optical quality and retinal sampling in the human eye. *Vision Research*, *36*(8), 1103–1114.

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

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.

Appendix A: Fitting

In the main text our presentation follows the order in which the experimental findings depend on each other, building up from grating detection experiments to masking by arbitrary natural images. As this is not the order in which we fitted the parameters to data, we explain the setting of parameters in this appendix in the order we fixed the parameters.

As our model computes a percent correct pc_i for any pair of stimuli to be discriminated, we can compute the likelihood L—the probability of observing the data given the model parameters—directly from the observed number of correct trials k_i and the total number of trials n_i in each specific experimental condition using the Binomial distribution $\mathcal{B}$:

$$L(\theta|\text{data}) = P_M(\text{data}|\theta) = \prod_{i=1}^{N} \mathcal{B}(k_i|n_i, pc_i) \quad (9)$$

$$= \prod_{i=1}^{N} \binom{n_i}{k_i} (pc_i)^{k_i} (1 - pc_i)^{n_i - k_i} \quad (10)$$

As it is usually done, we computed the log-likelihood l from this and removed constant factors from the equation:

$$l(\theta|\text{data}) = \log(L(\theta|\text{data})) = \sum_{i=1}^{N} \log(\mathcal{B}(k_i|n_i, pc_i)) \quad (11)$$

$$= \sum_{i=1}^{N} \log\left(\binom{n_i}{k_i}\right) + \sum_{i=1}^{N} (k_i \log(pc_i) + (n_i - k_i)\log(1 - pc_i)) \quad (12)$$

$$= C + \sum_{i=1}^{N} (k_i \log(pc_i) + (n_i - k_i)\log(1 - pc_i)) \quad (13)$$

For this log-likelihood we calculated a gradient over the parameters of the nonlinearity as detailed in

Appendix B and optimized using a BFGS algorithm as implemented in MATLAB's "fminunc" function.

As not all data are constrained in each condition, for which we needed a separate parameter fit, we had to fit the parameters in a successive fashion.

We first fixed the parameters of the preprocessing for all presentation times. For the optical distortions we fixed the pupil size to 4 mm diameter as a rough estimate for the environment of psychophysical measurements. Using preliminary parameter estimates from the literature, we then fixed the initial neural weighting of spatial frequencies to fit the detection data for each presentation duration.

Next we fixed the parameters of the log-Gabor-decomposition. We set the filter bandwidths to 40° and 1.4 octaves for orientation and frequency respectively, based on rough estimates from earlier measurements. We then set the range of spatial frequencies to $.5 - 20$ $\frac{cyc}{deg}$ roughly covering the visible range of frequencies. For the number of channels we set the model to use eight orientations and 12 frequencies to reduce the ripple artefacts in the output to a bearable range as described in the main text. At this stage, we also fixed the bandwidth of the normalization pool, setting the standard deviation of the Gaussian to be $\sigma_F = 0.5$ octaves.

Next we fixed the bandwidth of the normalization pool in orientation based on the oblique masking data for the 1497-ms presentation time, for which we had most data. To do so, we computed the likelihood for a grid of parameter values over the normalization bandwidth ω_θ, p, q, and C.

One computational trick we used to reduce the number of parameters to evaluate was to fit the overall noise variance independently of the other parameters. This can be done very efficiently, because scaling the noise for all pixels and all channels by the same factor c_ε does not change the optimal decoding scheme. Thus, the signal-to-noise ratio (SNR) with a changed overall noise size can be computed using only the final SNR from the original evaluation. We used this trick to replace the two parameters N_F and N_C with the single parameter $\frac{N_F}{N_C}$.

We then used a grid search to optimize parameters for each presentation time. In this grid search we used the parameters listed in Table 2. These parameter values cover the range for p, q, and C densely. For σ_θ we chose $\left\{0, \frac{3}{8}, \frac{7}{8}, \frac{12}{8}, \infty\right\} \times \sigma_\theta$—the orientation bandwidth of the filter—covering the range of qualitative behaviors for this parameter.[2] Similarly we set the linear noise factor N_F to $\{0, 0.1, 1, 10\} \times N_C$. By saving the likelihood value for each image combination separately, we could extract this cube for different parts of the data.

The results of the grid search are displayed in Figure 18, displaying the maximum likelihood found in the slice which sets the given parameter to the plotted value. Different lines give the values for the different $\frac{N_F}{N_C}$ values. The different panels are based on different subsets of the data.

Parameter	Levels
p	1.00, 1.50, 1.60, 1.70, 1.80, 1.90
	2.00, 2.10, 2.20, 2.30, 2.40, 2.50, 2.60, 2.70, 2.80, 2.90
	3.00, 3.25, 3.50, 3.75, 4.00, 4.50, 5.00
q	0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0
$C \times 10^2$	0.1000, 0.1292, 0.1668, 0.2154, 0.2783
	0.3594, 0.4642, 0.5995, 0.7743, 1.0000
ω_θ	0, 0.1112, 0.2594, 0.4447, ∞
$\frac{N_F}{N_C}$	0, 0.1, 1, 10

Table 2. Parameter values evaluated in the grid search for parameters. *Note*: For each parameter combination an optimal factor to the final variance was fit as a final noise factor.

Based on the displayed results on the grid search, we drew the following conclusions:

- As displayed in Panel A, the clearly best bandwidth of the normalization pool ω_θ is the one slightly smaller than the bandwidth of the filter. ($\omega_\theta = \frac{7}{8}\sigma_\theta$)
- From Panel B: The composition of the noise and q are coupled. When the linear contribution to the noise grows, larger values of q are needed to compensate this. However, any $\frac{N_F}{N_C}$-ratio explains the data equally well, when we use the adequate q. To remove this ambiguity, we set N_F to zero.
- Finally, from Panels C and D: p and C are reasonably well constrained by the data. However, the oblique masking data and the contrast discrimination favor slightly different values for p and C (not shown). These result in the two parameter values we display in the main paper. The two parameters differ only slightly, however, and make reasonably similar predictions as we saw in the main paper.

We evaluated the same range of parameter values for the other presentation times and for single pixel normalization. For the other presentation times, we can draw the same conclusions as above. For the single pixel normalization, however, we find a pronounced inconsistency of oblique masking and contrast discrimination. The oblique masking requires a much higher p value than the discrimination data. Consequently a parameter which optimizes the results for both conditions is considerably worse in the single pixel normalization model than in the mean normalization model.

As our grid was a bit coarse, we used the best parameter from the grid search as a starting point for some further optimization with a BFGS algorithm employing the gradients from Appendix B:

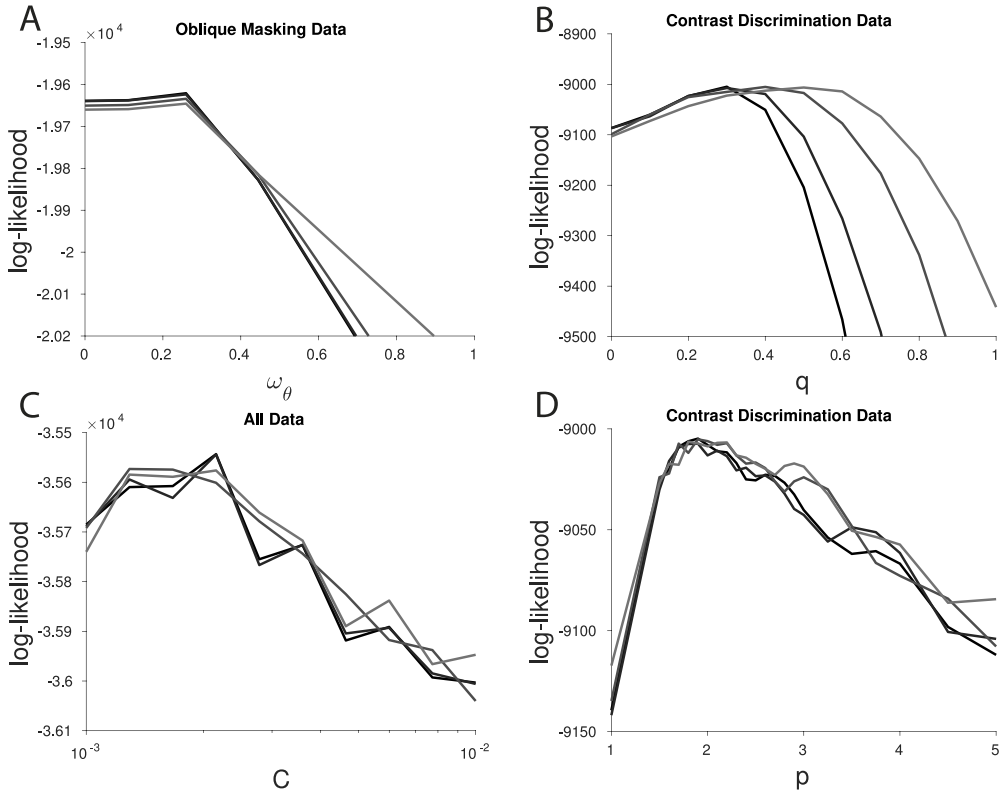

Figure 18. Evaluation of the grid search over the parameter space for the 1497-ms presentation time. Each panel shows the maximum likelihood reached with the given parameter value. (A) Bandwidth of the normalization pool evaluated over the oblique masking data. Maximum is at $\frac{7}{8}\sigma_\theta$. $\omega_\theta = \infty$ is plotted at $\omega_\theta = 1$. (B) Exponent q evaluated on the contrast discrimination data. (C) Constant C evaluated over all data. (D) Exponent p evaluated on the contrast discrimination data.

- First, we fit the bandwidth ω_θ to the oblique masking data, fixing N_F to 0 and p, q, and C to the optimal values from the grid.
- Using the estimate for ω_θ from this optimization, we fitted four parameter values for p, q, C, and N_C:
  - One for each presentation time to the corresponding contrast discrimination data, starting the optimization at the optimal value from the grid search for that presentation time.
  - One to the oblique masking data for the 1497-ms presentation time, starting at the best grid point again.
  - One for the ModelFest dataset starting at the optimal parameter for the 1497-ms presentation time from the grid search.

For an additional comparison on the ModelFest data, we fitted the ModelFest data adjusting only N_C starting from the parameter for 1497 ms and 79 ms respectively. To fit these, we again calculated performance from the signal-to-noise ratios, reducing the computational cost for this step.

Finally, we fitted a parameter set for the ModelFest dataset specifically, although the estimates from the classical data were decent already. As we did not have individual percent correct values for these data, we transformed the given thresholds to surrogate blocks of trials with percent correct. We assumed three blocks of 100 trials each: One with 86 correct trials at the threshold, one with 100 correct trials at 1.5 times the threshold, and a block with 50 correct trials a factor 3 below threshold. Using this surrogate data we then fitted the normalization and noise parameters (N_c, N_f, C, p, and q) as for the classical data.

As a last rather cosmetic step, we refit the neuronal filter we employ with the final parameters to fit the data for detection well, which was necessary as the processing of the model does distort the csf (higher exponents exaggerate the differences between different input strengths).

To give the reader a better understanding of what the different parameter values mean, Figure 19 shows the effect on the contrast discrimination results, when the different parameters are varied separately. Clearly the parameters N_f, N_c, and C merely move the function around hardly changing its shape. In contrast changing p—i.e., both exponents—controls how peaked the dip in the contrast sensitivity function is. Changing q—i.e., only the numerator exponent—strongly affects detection performance and how steeply the discrimination thresholds rise with pedestal contrast for high pedestal contrasts.

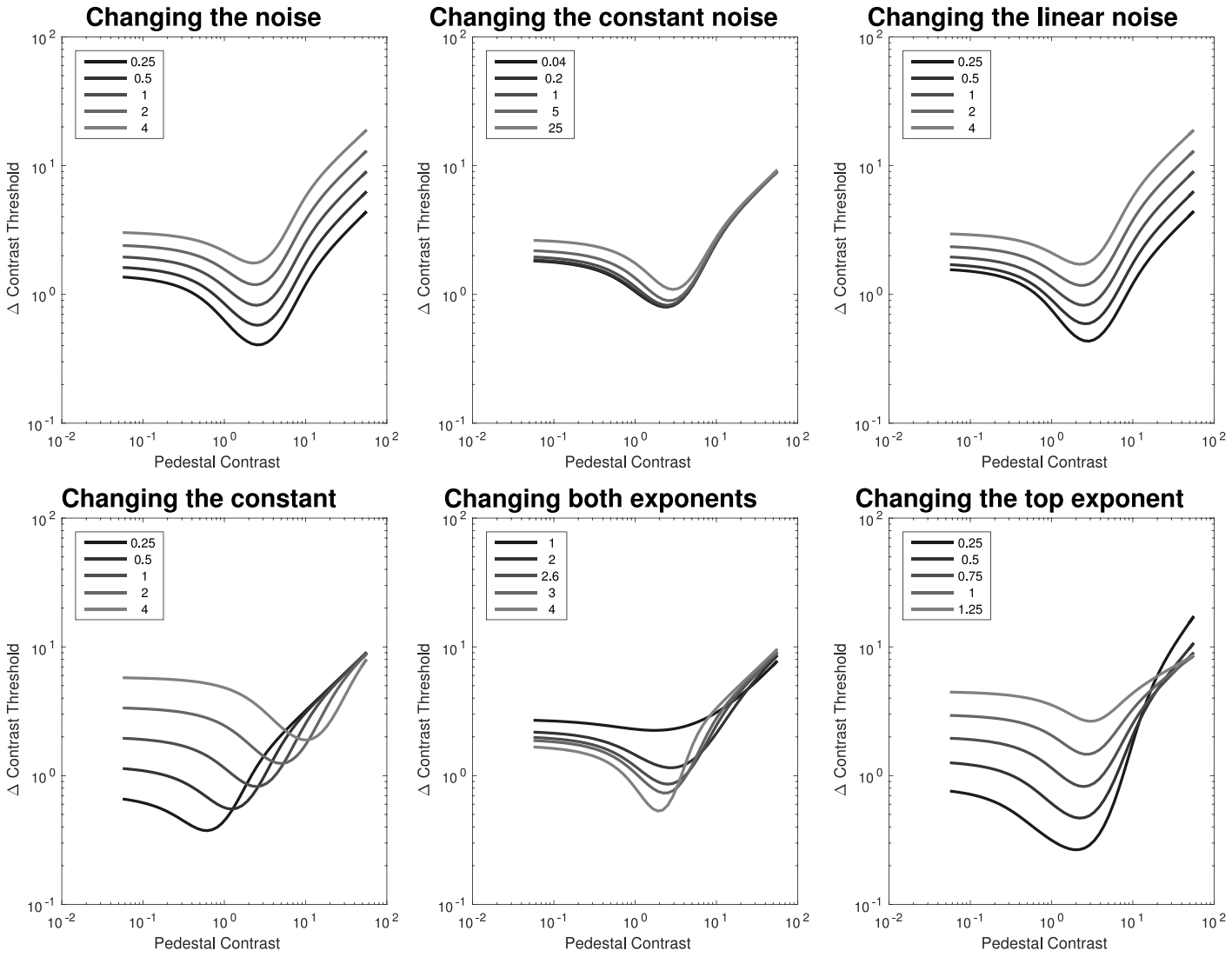

Figure 19. Effects of changing the parameters on contrast discrimination curves. Each panel shows the curve of contrast discrimination thresholds ("the dipper") when a single parameter of the model is changed, leaving the others at their fitted values.

Appendix B: Derivatives of the model

For parameter optimization we derived a gradient of the model likelihood with respect to all parameters. To compute this, we also compute derivatives for the signal-to-noise ratio, percent correct, and quite a few of the internal model states against each other. For a mathematically proficient reader these might thus provide some insight into the internal dependencies of the model. Also these calculations illustrate that the derivatives of the stages in our model can be computed as for the now popular deep neural network models.

Our presentation of the model derivatives follows the calculation in backward order in analogy to the back prop algorithms for deep neural networks; i.e., we start with the likelihood and go back to the parameters using the chain rule consecutively. Computation can be implemented in forward order with equal ease.

For each step we will first calculate the derivatives with respect to the parameters used in the step directly and then the one to the input of the processing step, which allows the calculation of derivatives with respect to parameters used in the previous processing step.

Likelihood from signal-to-noise ratio

We start with the log-likelihood, which depends on the lapse rate λ and the signal-to-noise ratio $\frac{d'}{\sqrt{\eta}}$ from the model. As a first step we calculate the derivative with respect to p_c, the percent correct predicted by the model without lapses:

$$\frac{\partial l(correct)}{\partial p'_c} = \frac{\partial}{\partial p'_c}\log(\lambda + (1-2\lambda)p'_c) = \frac{1-2\lambda}{\lambda + (1-2\lambda)p'_c} \quad (14)$$

$$\frac{\partial l(incorrect)}{\partial p'_c} = \frac{\partial}{\partial p'_c}\log(1-\lambda-(1-2\lambda)p'_c) = \frac{-(1-2\lambda)}{1-\lambda-(1-2\lambda)p'_c} \quad (15)$$

The derivative of the predicted percent correct p_c with regard to the signal-to-noise ratio $\frac{d}{\sqrt{\eta}}$ is simply the density of the normal distribution at the signal-to-noise ratio:

$$\frac{\partial p'_c}{\partial\left(\frac{d}{\sqrt{\eta}}\right)} = \phi\left(\frac{d}{\sqrt{\eta}}\right) \quad (16)$$

Decoding

Next we analyses the decoding stage. This stage receives two arrays of model responses $\{r_i^{(1)}\}$ and $\{r_i^{(2)}\}$ both indexed with an index i from an index-set $\mathcal{I}$ over position, orientation, and frequency. As the output we consider the signal-to-noise ratio $\frac{d}{\sqrt{\eta}}$. To calculate the derivative of the signal-to-noise ratio $\frac{d}{\sqrt{\eta}}$ with respect to any parameter used earlier in the model x, we use the following formulas:

$$\frac{\partial}{\partial x}\frac{d}{\sqrt{\eta}} = \frac{1}{\sqrt{\eta}}\frac{\partial d}{\partial x} + d\eta^{\frac{-3}{2}}\frac{\partial \eta}{\partial x} = \frac{1}{\sqrt{\eta}}\sum_{i\in\mathcal{I}}\frac{\partial d_i}{\partial x} + d\eta^{\frac{-3}{2}}\sum_{i\in\mathcal{I}}\frac{\partial \eta_i}{\partial x} \quad (17)$$

The two parameters of the decoding stage are the size of the constant noise N_c and the factor for the noise variance N_f for which we calculate the derivatives first:

$$\frac{\partial \eta_i}{\partial N_c} = 1 \qquad \frac{\partial \eta_i}{\partial N_F} = r_i \qquad \frac{\partial r_i}{\partial N_C} = \frac{\partial r_i}{\partial N_F} = 0 \quad (18)$$

For any other parameters x which changes r_i, we can calculate the derivatives from the derivative $\frac{\partial r_i}{\partial x}$:

$$\forall i \in \mathcal{I}: \frac{\partial \eta_i}{\partial x} = \frac{\partial}{\partial x}\frac{\left(r_i^{(1)} - r_i^{(2)}\right)^2}{n_i^{(1)} + n_i^{(2)}}(n_i^{(1)} + n_i^{(2)}) = \frac{\partial}{\partial x}\left(r_i^{(1)} - r_i^{(2)}\right)^2 \quad (19)$$

$$= 2\left(r_i^{(1)} - r_i^{(2)}\right)\left(\frac{\partial r_i^{(1)}}{\partial x} - \frac{\partial r_i^{(2)}}{\partial x}\right) \quad (20)$$

$$\frac{\partial d_i}{\partial x} = \frac{\partial}{\partial x}\frac{\left(r_i^{(1)} - r_i^{(2)}\right)^2}{\sqrt{n_i^{(1)} + n_i^{(2)}}} \quad (21)$$

$$= 2\frac{r_i^{(1)} - r_i^{(2)}}{\sqrt{n_i^{(1)} + n_i^{(2)}}}\left(\frac{\partial r_i^{(1)}}{\partial x} - \frac{\partial r_i^{(2)}}{\partial x}\right) - \frac{\left(r_i^{(1)} - r_i^{(2)}\right)^2}{2\left(n_i^{(1)} + n_i^{(2)}\right)^{\frac{3}{2}}}\left(\frac{\partial n_i^{(1)}}{\partial x} + \frac{\partial n_i^{(2)}}{\partial x}\right) \quad (22)$$

$$= 2\frac{r_i^{(1)} - r_i^{(2)}}{\sqrt{n_i^{(1)} + n_i^{(2)}}}\left(\frac{\partial r_i^{(1)}}{\partial x} - \frac{\partial r_i^{(2)}}{\partial x}\right) - N_f\frac{\left(r_i^{(1)} - r_i^{(2)}\right)^2}{2\left(n_i^{(1)} + n_i^{(2)}\right)^{\frac{3}{2}}}\left(\frac{\partial r_i^{(1)}}{\partial x} + \frac{\partial r_i^{(2)}}{\partial x}\right) \quad (23)$$

using in the last step:

$$\frac{\partial n_i}{\partial x} = N_f\frac{\partial r_i}{\partial x} \quad (24)$$

Normalization

Thus, in the Normalization stage we require the derivatives of the response r_i, which we again first compute for the parameters of this stage p, q, and C and then for the bandwidths of the normalization ω and the filter σ.

For $a_i = 0$ all derivatives are 0 because r_i is then 0 independent of all parameters; for $a_i > 0$:

$$\frac{\partial r_i}{\partial p} = \log(a_i)r_i - \frac{a_i^{p+q}}{(C^p + b_i)^2}\left[\log(C) + \sum_{i\in\mathcal{I}}(G * \log(a))(x_i)\right] \quad (25)$$

$$\frac{\partial r_i}{\partial q} = \log(a_i)r_i \quad (26)$$

$$\frac{\partial r_i}{\partial C} = -r_i p\frac{C^{p-1}}{C^p + b_i} \quad (27)$$

For computing the derivatives with respect to the σs we need to compute the derivatives of r_i towards a_i and b_i as well as the derivatives of the filter values:

$$\frac{\partial r_i}{\partial a_i} = (p+q)\frac{a_i^{p+q-1}}{C^p + b_i} + \frac{\partial r_i}{\partial b_i}\frac{\partial b_i}{\partial a_i} \quad (28)$$

$$\frac{\partial r_i}{\partial b_i} = -r_i \frac{1}{C^p + b_i} \quad (29)$$

A Gaussian $G(x|\sigma)$ in x without normalization (as the ones in frequency space to define the log-Gabors) has the following derivative with respect to its standard deviation σ

$$\frac{\partial G}{\partial \sigma}(x) = \frac{(x-\bar{x})^2}{\sigma^3} \exp\left(-\frac{(x-\bar{x})^2}{2\sigma^2}\right) \quad (30)$$

For the normalization of a Gaussians $G_n = G_n(\cdot|\sigma)$, we used the sum over the grid in each case, i.e.,:

$$\forall i \in \mathcal{I}: \qquad G_n(x_i) = \frac{G(x_i)}{\sum_{j\in\mathcal{I}} G(x_j)} \quad (31)$$

The derivative of this normalized Gaussians G_n with respect to its *SD* σ is thus given by:

$$\frac{\partial G_n}{\partial \sigma}(x_i) = \frac{1}{\sum_{j\in\mathcal{I}} G(x_j)} \frac{\partial G}{\partial \sigma}(x_i) - \frac{G(x_i)}{\left(\sum_{j\in\mathcal{I}} G(x_j)\right)^2} \sum_{j\in\mathcal{I}} \frac{\partial G}{\partial \sigma}(x_j) \quad (32)$$

$$= \frac{1}{\sum_{j\in\mathcal{I}} G(x_j)} \left(\frac{\partial G}{\partial \sigma}(x_i) - G_n(x_i) \sum_{j\in\mathcal{I}} \frac{\partial G}{\partial \sigma}(x_j)\right) \quad (33)$$

For a convolution, when only one of the two convolved functions f depends on the variable x:

$$\frac{\partial (f(x) * g(y))}{\partial x} = \frac{\partial f(x)}{\partial x} * g(y) \quad (34)$$

$$\frac{\partial (g(y) * f(x))}{\partial x} = g(y) * \frac{\partial f(x)}{\partial x} \quad (35)$$

Thus, we can compute the derivatives of $B = \{b_i\}_{i\in I}$ interpreted as the four dimensional array of normalization inputs for each channel at each position: For any of the standard deviations $\omega_{x,y,\phi,f}$ we can decompose the 4D Gaussian into the four one-dimensional convolutions and compute the four derivatives as follows:

$$\frac{\partial B}{\partial \omega_x} = \frac{\partial G(\omega_x)}{\partial \omega_x} * G(\omega_y, \omega_\phi, \omega_f) * A^p \quad (36)$$

$$\frac{\partial B}{\partial \omega_y} = \frac{\partial G(\omega_y)}{\partial \omega_y} * G(\omega_x, \omega_\phi, \omega_f) * A^p \quad (37)$$

$$\frac{\partial B}{\partial \omega_\phi} = \frac{\partial G(\omega_\phi)}{\partial \omega_\phi} * G(\omega_x, \omega_y, \omega_f) * A^p \quad (38)$$

$$\frac{\partial B}{\partial \omega_f} = \frac{\partial G(\omega_f)}{\partial \omega_f} * G(\omega_x, \omega_y, \omega_\phi) * A^p \quad (39)$$

For any parameter, except the parameters of the normalization pool:

$$\frac{\partial B}{\partial A} = G(\omega_x, \omega_y, \omega_\phi, \omega_f) * \frac{\partial A^p}{\partial A} \quad (40)$$

$$= G(\omega_x, \omega_y, \omega_\phi, \omega_f) * pA^{p-1} \quad (41)$$

Decomposition

As we did not fit the filters to data, we do not require the derivatives to their bandwidths for fitting. These derivatives can be computed nonetheless as follows:

The derivative of the absolute value we apply between the decomposition and the nonlinearity with respect to a parameter which influences real $\Re$ and imaginary $\Im$ part of a complex number z is:

$$\frac{\partial |f(x)|}{\partial x} = \frac{\Re(f(x))}{|f(x)|} \frac{\partial \Re(f(x))}{\partial x} + \frac{\Im(f(x))}{|f(x)|} \frac{\partial \Im(f(x))}{\partial x} \quad (42)$$

This is not a proper complex derivative, but only a real derivative by interpreting the complex z as $\in \mathbb{R}^2$

Finally, to compute the derivatives of the filter output against the filter parameters, we can use the following formula

$$\frac{\partial \mathcal{F}(f)}{\partial x} = \mathcal{F}\left(\frac{\partial f}{\partial x}\right), \quad (43)$$

because the Fourier-transform is a linear operator.

The filtering in Fourier space is an element wise multiplication. Thus, the derivative of a channel response $f(x, y)$ can be computed from the derivatives of the filters in Fourier space $g(x, y)$ and the image $I(x, y)$:

$$\frac{\partial f(x)}{\partial \sigma} = \frac{\partial}{\partial \sigma} \mathcal{F}^{-1}(\mathcal{F}(I(x,y)) g(x,y)) = \mathcal{F}^{-1}\left(\mathcal{F}(I(x,y)) \frac{\partial g(x,y)}{\partial \sigma}\right) \quad (44)$$

Preprocessing

Our preprocessing is an affine transformation. Thus the derivatives with respect to the original inputs can be computed from derivatives with respect to the preprocessed image simply by applying the same filters with flipped phase and adding back the mean luminance.

Appendix C: Optimizing stimuli

To compare different models, one interesting method is to optimize stimuli which are especially different or similar according to one model while keeping similarity according to another model constant (Wang & Simoncelli, 2008). As analyses of this type are a strength of image-computable models, we use it in the main text to show the advantages of an image-computable model. In this appendix we explain the details of the optimization procedure we employ to get the stimuli.

We aim to find luminance images ($I_1, I_2 \in R^{N\times N}$) which have a given Root Mean Square Error ($RMSE_0$) from a given start image I_0 after conversion to contrast and cut out of the fovea and are either maximally easy (I_1)or maximally hard (I_2) to differentiate from I_0 according to the model.

Furthermore we require two constraints on the images to yield displayable and interesting stimuli: (a) All pixels must be in the range [0, L_m] for a maximal displayable luminance L_m. (b) Pixels for which the foveal window $w \in R^{N\times N}$ is 0 shall be equal to the corresponding pixels in I_0.

For notation we shall use:

- N for the size of the square images
- $I' = w \cdot \left(I / \left(\frac{1}{N^2}\sum_{j,k=1}^{N} I_{0jk}\right)\right)$ for the image I after conversion to contrast and application of the foveal window w. Here "/" means element-wise division. Note that we always use the mean luminance of I_0 for this conversion.
- $RMSE(I'_1, I'_0) = \sqrt{MSE(I'_1, I'_0)} = \sqrt{\sum_{j,k=1}^{N}(I'_{1jk} - I'_{0jk})^2}$ to denote the root mean squared error of two (converted) images I_0 and I_1.
- $d'(I_1, I_0)$ to denote the discriminability d' of I_1 and I_0 according to our model. Additionally we write $d'(I'_1, I'_0) := d'(I_1, I_0)$ for converted images overloading notation.

To allow a conversion back from a contrast image I' to a luminance image I, we set (suppressing indices):

$$I(I') = \begin{cases} I_0 & w \le 0.001 \\ \frac{1}{N^2}\sum_{j,k=1}^{N} I_{0jk})(I'/w) & w > 0.001 \end{cases}, \quad (45)$$

i.e., wherever the foveal window is 0 we set the luminance image to be equal to I_0. As we enforce I_1 and I_2 to be equal to I_0 there, this yields correct results. To avoid numerical issues with the division by w, we extend this enforced equality to pixels with $w \le 0.001$.

This yields the following optimization problem in mathematical shorthand:

Minimize $d'(I_1, I_0)$ / Maximize $d'(I_2, I_0)$

subject to:

$$0 \le I_1, I_2 \le L_m \quad (46)$$

$$RMSE(I'_2, I'_0) = RMSE(I'_1, I'_0) = RMSE_0 \quad (47)$$

$$\forall j,k = 1 \ldots N : w_{jk} \le 0.001 \Rightarrow I_{1jk} = I_{2jk} = I_{0jk} \quad (48)$$

To solve this problem with nonlinear equality constraints approximately, we relax the constraints quadratically and add one common parameter β which shall increase during optimization to increase the penalty for missing the constraints. Finally, we add another regularize $\sum_{j,k=1}^{N}(1 - w_{jk})^{\beta}(I'_{0jk} - I'_{1jk})^2$ which pushes the optimization to yield similar images near the edges of the window w.

For I_1 this yields the following relaxed optimization problem:

Minimize:

$$d'(I_1, I_0) + \beta^4(MSE(I'_1, I'_0) - RMSE_0^2)^2 \quad (49)$$

$$+\beta^2 \sum_{j,k=1}^{N}(1 - w_{jk})^{\beta}(I'_{0jk} - I'_{1jk})^2 \quad (50)$$

subject to:

$$0 < I_1 < L_m \quad (51)$$

We then use a gradient decent algorithm to solve this optimization problem starting from Gaussian noise added to the area where $w > 0.001$ with the correct $RMSE_0$ and cut to fit the displayable luminance range. We then apply a gradient descent during which we test at each point whether it is better than the previous one. Depending on the outcome of this, we adjust the stepsize, adding 30% every time we update successfully and dividing by 2 every time we fail. We increase β by 1 every time the change predicted by the current gradient and step size is smaller than 0.001. When $\beta = 100$ and the predicted change is smaller than 10^{-6}, we end the optimization. If at any time a pixel leaves the allowed luminance range, we set it back inside the range by the smallest possible numerical value.