The impact on midlevel vision of statistically optimal divisive normalization in V1

Ruben Coen-Cagli

Department of Basic Neuroscience, University of Geneva, Geneva, Switzerland

Odelia Schwartz

Dominick Purpura Department of Neuroscience, Albert Einstein College of Medicine, Bronx, NY, USA
Department of Systems and Computational Biology, Albert Einstein College of Medicine, Bronx, NY, USA

The first two areas of the primate visual cortex (V1, V2) provide a paradigmatic example of hierarchical computation in the brain. However, neither the functional properties of V2 nor the interactions between the two areas are well understood. One key aspect is that the statistics of the inputs received by V2 depend on the nonlinear response properties of V1. Here, we focused on divisive normalization, a canonical nonlinear computation that is observed in many neural areas and modalities. We simulated V1 responses with (and without) different forms of surround normalization derived from statistical models of natural scenes, including canonical normalization and a statistically optimal extension that accounted for image nonhomogeneities. The statistics of the V1 population responses differed markedly across models. We then addressed how V2 receptive fields pool the responses of V1 model units with different tuning. We assumed this is achieved by learning without supervision a linear representation that removes correlations, which could be accomplished with principal component analysis. This approach revealed V2-like feature selectivity when we used the optimal normalization and, to a lesser extent, the canonical one but not in the absence of both. We compared the resulting two-stage models on two perceptual tasks; while models encompassing V1 surround normalization performed better at object recognition, only statistically optimal normalization provided systematic advantages in a task more closely matched to midlevel vision, namely figure/ground judgment. Our results suggest that experiments probing midlevel areas might benefit from using stimuli designed to engage the computations that characterize V1 optimality.

Introduction

Hierarchical processing across distinct areas is a prominent feature of the primate visual cortex. The early areas primary (V1) and secondary (V2) visual cortices provide a paradigmatic example, yet neither the interactions between them nor the computations in V2 are well understood. Experimentally, two main sets of observations about V2 processing suggest that (a) the size of the receptive fields (RFs) increases steadily when moving higher up in the hierarchy, in particular by a factor of two from V1 to V2 (Gattass, Gross, & Sandell, 1981; Shushruth, Ichida, Levitt, & Angelucci, 2009), and (b) the selectivity of V2 receptive fields for complex shape constituents increases over V1, including curved contours (Hegde & Van Essen, 2000), angles (Ito & Komatsu, 2004), combinations of orientations (Anzai, Peng, & VanEssen, 2007), pooling of multiple spatial frequencies (Willmore, Prenger, & Gallant, 2010), and border ownership or side-of-figure signals (Zhou, Friedman, & Von Der Heydt, 2000). This second set of studies is less conclusive though because, as opposed to V1, the appropriate stimulus space to be explored remains unknown.

Therefore, a principled computational approach would be most helpful in clarifying midlevel computations and their role. In recent years, following the idea that cortical processing is optimized to the statistics of the natural environment (Barlow, 1961; Simoncelli & Olshausen, 2001; Zhaoping, 2006), multilayer models of image statistics have been studied that capture higher-order correlations (Garrigues & Olshausen, 2008; Hinton, 2007; Karklin & Lewicki, 2005; Osindero, Welling, & Hinton, 2006; Schwartz, Sejnowski, & Dayan, 2006; Theis, Hosseini, & Bethge, 2012) and
learn without supervision intermediate features, such as extended contours (Hoyer & Hyvaarinen, 2002), corners (Lee, Ekanadham, & Ng, 2008; Malmir & Shiry Ghidary, 2009; Spratling, 2011), and pooling over multiple frequencies (Hyvaarinen, Gutmann, & Hoyer, 2005) and orientations (Lindgren & Hyvaarinen, 2006; Nuding & Zetzsche, 2007; Shan, Zhang, & Cottrell, 2007).

Here, following this general approach, we focused on a specific question that has not been systematically addressed so far: What is the impact of V1 non-linearities (specifically, surround divisive normalization) on the computations that take place downstream and eventually on perception? Divisive normalization describes cortical firing rate responses to a stimulus as the ratio between the output of a linear RF and some measure of the energy of a group of RFs (the normalization pool) (Heeger, 1992). It is considered a canonical computation that is found across cortical areas and sensory modalities (Carandini & Heeger, 2012). In particular, it has been used extensively to capture V1 surround suppression data (Cavanaugh, Bair, & Movshon, 2002; Series, Lorenceau, & Frégnac, 2003), and descriptive models of midlevel neurons often involve a divisively normalized V1-like stage (e.g., Rust, Schwartz, Movshon, & Simoncelli, 2005). Moreover, divisive normalization and related models have been developed from a principled scene-statistics perspective, explaining a range of V1 phenomena and linking divisive normalization to redundancy reduction by cortical neurons (Coen-Cagli, Dayan, & Schwartz, 2012; Karklin & Lewicki, 2009; Schwartz & Simoncelli, 2001; Spratling, 2011). However, residual correlations are often present in the outputs of model V1 neurons, and their precise form can influence downstream processing. Indeed, Freeman, Ziemba, Heeger, Simoncelli, and Movshon (2013) provided evidence that V2 computes average correlations between V1 responses across visual space.

To address this issue, we used a two-stage architecture (Figure 1) in which the outputs of V1-like oriented RFs are non-linearly transformed and then linearly combined across space into V2 RFs; the V2 RFs are then pooled across space and decoded to produce a perceptual judgment. We considered models of V1 that encompassed different types of surround nonlinearity and compared (a) the structure of correlations between the V1 responses, (b) the emerging V2 selectivity, and (c) the performance in perceptual classification tasks. For the V1-stage nonlinearity, we first considered a statistically optimal generalization of divisive normalization, obtained by exact Bayesian inference in a mixture of Gaussian scale mixtures (MGSM) generative model of image statistics (Coen-Cagli et al., 2012; Schwartz et al., 2006; Schwartz & Simoncelli, 2001). We term this a flexible surround model because the

![Figure 1. Overview. We used a two-stage architecture to address three main issues, summarized by the text in red on the left. From the bottom, up: First, we were interested in the V1 output statistics, namely, the structure of correlations in a population of V1 neurons across space and orientation preferences. Second, we considered the selectivity that could emerge downstream, e.g., in V2, where V1 outputs are pooled over a region of space larger than the individual V1 RF. Third, we used a population of V2 units to perform perceptual classification tasks. The goal of this paper is to assess the effects on the three issues mentioned above of using a V1 stage with nonlinearities differently optimized to natural image statistics.](https://jov.arvojournals.org/pdfaccess.ashx?url=/data/journals/jov/933547/ on 11/04/2018)
that when the V1 responses are pooled across spatial neighborhoods larger than the individual V1 RF (i.e., neighborhoods that form the input to the V2 stage), proper surround normalization greatly helps to see the fine detail. Indeed, PCA of the V1 responses across space and orientations revealed V2-like features only when we used the optimal normalization or, to a lesser extent, the canonical one. We then compared the resulting two-stage models on perceptual tasks, such as object recognition, on which multiscale architectures have been tested traditionally (Riesenhuber & Poggio, 1999) and figure/ground judgment; in both cases, we have been tested traditionally (Riesenhuber & Poggio, 1999) and figure/ground judgment; in both cases, we found systematic advantages of using a V1 stage with statistically optimal surround normalization.

Introduction to the V1 modeling framework

In this section, we start with a nontechnical introduction to the models of V1 we used for the simulations. Then, in V1 models, we provide further technical details.

We included in the models operations that are known to capture major properties of V1 physiology: a linear filtering stage that provides selectivity for orientation, a nonlinearity that provides complex cell invariance, and divisive normalization to account for surround suppression (Heeger, 1992; Cavanaugh et al., 2002). Because we were interested in hierarchical computation in the cortex, we focused our efforts particularly on the latter because it is known that downstream targets, such as V2, pool V1 outputs over areas larger than the individual V1 RF (Gattass et al., 1981; Shushruth et al., 2009), and so we argue that V1 surround modulation plays an important role.

In this paper, we focus particularly on nonlinear models of divisive normalization derived from natural scene statistics. Note that the V1 models we use here are not new but rather according to Coen-Cagli et al. (2012).

It has been proposed that divisive normalization, which is observed across different cortical areas and sensory modalities (Carandini & Heeger, 2012), could be beneficial to achieving a better neural code of natural inputs. This view rests on the long-standing idea in neuroscience that cortical neurons are tuned to the properties of the natural sensory environment. In the case of V1, the link between divisive normalization and the structure of natural scenes can be made precise by analyzing images through the lens of oriented V1-like visual RFs: The outputs of RFs that are close in space generally display strong statistical dependencies, which can be related to divisive normalization in two ways.

One approach is to assume that neurons aim to code natural visual inputs efficiently by reducing the RFs’ dependencies; indeed, divisive normalization achieves this objective (e.g., Schwartz & Simoncelli, 2001). We adopt here a different but related approach: We assume that V1 responses represent inferences in a generative model of natural images (Dayan & Abbott, 1999). Generally speaking, this means we describe the outputs of different RFs as a set of dependent variables that arise from a transformation of another set of independent variables; then neural responses amount to inverting the transformation. Divisive normalization would appear naturally if the dependencies between RFs could be described by a multiplication, and the latter turns out to be a good first approximation (e.g., Wainwright, Simoncelli, & Willsky, 2001). Intuitively, the dependence between RFs arises from a global property of the image (for instance, contrast) that is shared between neighboring image regions and therefore similarly affects neighboring RFs. For instance, to generate dependencies between two RF outputs (say, at two different spatial locations), one starts with two independent variables corresponding to the neurons at the two locations and multiplies each of them by a shared variable that introduces the dependency. The neural output essentially reverses this procedure and thus amounts to dividing the RF output by the shared variable.

Recently, many authors have recognized that more sophisticated models of RFs coordination are needed due to the nonstationarity (Parra, Spence, & Sajda, 2001) of natural images, i.e., the fact that perceptually different regions of images (e.g., textures vs. occlusions) contain markedly different statistical dependencies. Ideally, a good model of images should learn without supervision the statistical structure of different image regions. Several variants of mixture models have been proposed that use different mixture components to account for different statistical dependencies between pixels or RFs. These approaches have led to large improvements in unsupervised learning of interesting image structure (Karklin & Lewicki, 2005; Kivinen, Sudderth, & Jordan, 2007; Schwartz et al., 2006), denoising (Guerrero-Colon, Simoncelli, & Portilla, 2008; Hammond & Simoncelli, 2008), synthesis (Theis et al., 2012), and modeling of cortical response properties (Karklin & Lewicki, 2009).

We have recently proposed (Coen-Cagli, Dayan, & Schwartz, 2009; Coen-Cagli et al., 2012; Schwartz, Sejnowski, & Dayan, 2009) that one particularly important distinction for cortical neurons is between image regions that elicit statistically dependent RF outputs (e.g., the fur of a zebra) and regions that produce independent RF outputs (e.g., two RFs sitting across the border between zebra and back-
ground). If the aim of divisive normalization is to undo the RFs' dependencies, then it should operate only when the input image is consistent with dependent RF outputs (e.g., a large homogeneous grating) but not when it is consistent with independent RF outputs (e.g., a vertical grating in the RF center surrounded by a diagonal annulus). We showed, using RFs that cover both the center and surround of the neural RF, that this flexible divisive normalization explains a large range of experimental findings on V1 surround modulation, including some that are not captured by the canonical divisive normalization (Coen-Cagli et al., 2012). In the next section, we formalize this intuition with a Bayesian mixture model: Neural responses still reflect inferences in the model, but now there are two steps of inference. For any given input image, the model uses Bayesian inference first to estimate the extent to which the RF outputs are dependent or independent and then to estimate the neural response, which includes surround normalization only in the former case, not the latter.

An advantage of using this modeling approach is that it allows us to manipulate the degree to which the V1 surround normalization is statistically optimal (i.e., how accurately it represents inference in the generative model). We can then address quantitatively the consequences of such manipulations on the statistics of the V1 outputs and their effects on downstream computation and perceptual judgments.

V1 models

In this section, we describe three models for the V1 stage that differ in how visual inputs to the receptive field surround affect the firing rate: the flexible surround divisive normalization model, the canonical surround divisive normalization model, and the complex cell energy model (denoted no surround in the following). We start by describing the bank of linear filters we used to model center and surround RFs; then we describe the typical dependencies between RF outputs to natural images and provide details of the generative model. Then we present, for each of the three models above, the equations for the neural response to a given input image, and we explain how they correspond to optimal or suboptimal inferences. We note that the V1 modeling is not novel to this paper, and all the technical aspects were covered in detail in Schwartz et al. (2006) and Coen-Cagli et al. (2009, 2012). Nonetheless, for completeness, we provide in this section (and illustrate in Figure 2), the equations that are necessary to understand the mechanics of the V1 models; we also provide further derivations in Appendices I and II.

V1 model receptive fields

Given an image, we define the output of a RF as the inner product between the image and an oriented, bandpass linear filter kernel. We extract filters from the first level of a steerable pyramid (Portilla & Simoncelli, 2000; Simoncelli, Freeman, Adelson, & Heeger, 1992). The diameter of the RFs, defined by the size of an optimally oriented stimulus that is required to obtain 95% of the maximum output, is nine pixels. The peak spatial frequency is one/six cycles/pixel. We chose a value of the orientation bandwidth that approximates the median value found in V1 by Ringach, Shapley, and Hawken (2002): 23.5° half-width at 70% of the height of the tuning curve.

We consider RFs with orientation \( \theta \) (possible values are 0°, 45°, 90°, 135°), phase \( \phi \) (possible values are 0° and 90° to form a quadrature pair), and position \( x,y \) (possible values are minus six, zero, or six pixels, corresponding to a three by three grid with a six-pixel separation, centered at \( x = 0, y = 0 \)), and we denote the output of the generic RF by \( z_{0,\theta,\phi,x,y} \). In this paper, we distinguish RFs at the center of the input image (\( x = 0, y = 0 \)) and in the surround. For convenience, we denote the generic center RF output by

\[
k_{0,\theta,\phi} = z_{0,\theta,\phi,0,0}
\]

(1)

and the generic surround RF output by:

\[
s_{0,\theta,\phi,x,y} = z_{0,\theta,\phi,x,y} \quad \text{excluding } x = 0, y = 0
\]

(2)

In addition, we denote the vector of outputs of center RFs comprising all orientation and phases by:

\[
k = \left\{ k_{0,\theta,\phi} \right\}_{\theta = 0,45,90,135,\phi = 0,90}
\]

(3)

and we denote the vector of outputs of surround RFs comprising all positions and phases at a given orientation \( \theta \) by:

\[
s^{(\theta)} = \left\{ s_{\theta,\phi,x,y} \right\}_{\phi = 0,90, x,y = -6,0,6} \quad \text{excluding } x = 0, y = 0
\]

(4)

Figure 2A illustrates, with a cartoon, two RFs and the corresponding notation; it also illustrates the spatial layout by the red dots corresponding to the positions of the RF centers.

Models of statistical dependencies between V1 RFs

A general property of natural images is that adjacent locations contain highly redundant information. For instance, image regions that elicit strong (or weak) outputs from one RF typically also elicit strong (or...
weak) outputs from the second RF (Schwartz & Simoncelli, 2001). Intuitively, the dependence arises from global properties of the image (for instance, contrast) that are shared between neighboring image regions and therefore similarly affect neighboring RFs. This observation led to the formulation of a generative model of joint RF outputs, known as Gaussian scale mixture (GSM) (Wainwright et al., 2001). The goal of the generative model is to capture these statistical dependencies observed in RF outputs to natural scenes. This is done by starting with two variables that are independent and introducing a joint multiplicative dependency between them. This results in a dependency in which the two variables are likely to be high or low together in their absolute values.

More formally, let us consider, as in the previous section, two vectors of RF outputs \( \mathbf{k} \) and \( \mathbf{s} \) for the center and surround positions, respectively. Let us also introduce two vectors of Gaussian variables, denoted by \( \mathbf{j} \) and \( \mathbf{r} \), with the same dimensionality as \( \mathbf{k} \) and \( \mathbf{s} \), respectively; the vector formed by concatenating \( \mathbf{j} \) and \( \mathbf{r} \) has zero mean and covariance \( C_{jr} \). Let us then introduce a positive, scalar random variable denoted by \( \nu \) (which we will refer to as the mixer). The GSM describes the RF outputs as the variables obtained by the following multiplication:

\[
\mathbf{k} = \nu \mathbf{\kappa} \\
\mathbf{s} = \nu \mathbf{\sigma}
\]

While the dependencies between the elements of \( \mathbf{\kappa} \) and \( \mathbf{\sigma} \) are fully described by their covariance matrix \( C_{\kappa \sigma} \), the joint probability distribution of the elements of \( \mathbf{k} \) and \( \mathbf{s} \) now have also a higher-order dependence because the multiplication by the common mixer \( \nu \) scales all the variances simultaneously (for this reason, this specific form of higher-order dependence is sometimes also called variance dependence). A limitation of Equation 5 is that it assumes stationary dependencies between RF outputs. However, natural images are highly nonstationary, and more sophisticated models have been introduced to account for the fact that different parts of natural images could produce different RF dependencies (Guerrero-Colon et al., 2008; Hammond & Simoncelli, 2008; Karklin & Lewicki, 2005, 2009; Kivinen et al., 2007; Schwartz et al., 2006). In particular, Coen-Cagli et al. (2009, 2012) proposed that while image patches entirely contained within a single object may be well described by Equation 5, patches that sit at the boundary between different objects produce more independent center and surround dependencies between \( \mathbf{k} \) and \( \mathbf{s} \) is measured by the posterior probability (here denoted simply \( q \), see Equation 16); the V1 response is now a weighted sum of the two extreme cases with weights \( q \) and \( 1-q \), respectively.
surround RF outputs that are better described by
\[
k = v_k \kappa
\]
\[
s^{(o)} = v_o \sigma^{(o)}
\]
(6)

where \(v_k, v_o\) are independent mixer variables, and the Gaussian variables \(\kappa, \sigma^{(o)}\) can be correlated only within each group (with covariance matrices \(C_{\kappa}\) and \(C_{\sigma}\), respectively) but not across groups. We formalized this intuition using a MGSM (Coen-Cagli et al., 2009), with one mixture component accounting for dependent RFs (Equation 5), the other mixture component accounting for independent RFs (Equation 6), and a weighted sum of the two components accounting for the generic, intermediate case.\(^1\) We provide in Appendix I the analytical derivation of the generative model.

For the covariance matrices to actually capture structures that are present in natural images, we need to treat them as free parameters and optimize them. As with other generative models, the MGSM can be fitted by maximum likelihood: Namely, for a given training set of RF outputs \(\{k, s^{(o)}\}_{n=1}^N\) computed from natural image patches, we search for the parameter values that maximize the likelihood of the data under the model, i.e., \(p(\{k, s^{(o)}\}_{n=1}^N | \text{parameters})\). The model parameters are (a) the covariance matrices that summarize the structure of local dependencies between RFs for patches that do contain a global dependence \(C_{\kappa,\sigma}\) and patches that do not \(C_{\kappa,\sigma}\) (or zero) and (b) the prior probability that any given natural image patch does actually contain a global dependence. We maximized the log-likelihood numerically; details are provided in Appendix II and Coen-Cagli et al. (2009). For the training set, we used 75,000 patches randomly sampled from a database of five images (Boats, Goldhill, Mountain, Einstein, Crowd) often used for image compression benchmarks and available online at http://neuroscience.aecom.yu.edu/labs/schwartzlab/code/standard.zip. The patches were 21 by 21 pixels, large enough to cover center and surround RFs.

\section*{V1 responses to visual inputs}

We relate neural responses to estimates of (or inferences about) specific components of the model that generated the RFs’ dependencies. We describe the response of a V1 neuron at the center location with preferred orientation \(\theta\) and phase \(\phi\) as essentially reversing the operation that generated the higher-order dependence between \(k_{\theta,\phi}\) and the other RFs, thus computing an estimate \(\hat{k}_{\theta,\phi}\) of the corresponding local Gaussian variable \(k_{\theta,\phi}\). In principle, because the generative model is multiplicative, estimating the local Gaussian amounts to the reverse process, i.e., division by the mixer. Thus, the framework based on image

\begin{align*}
\lambda(k, s^{(o)}) &= \sqrt{(k, s^{(o)}) \cdot C_{\kappa,\sigma}^{-1} \cdot (k, s^{(o)})} \\
\lambda(k) &= \sqrt{k^{\top} \cdot C_{\kappa}^{-1} \cdot k}
\end{align*}

statistics naturally encompasses, and extends, divisive normalization. However, for any given visual input, only the RF outputs are observed whereas it is not known whether such outputs were generated according to Equation 5 (dependent center and surround) or Equation 6 (independent center and surround) nor the actual value of the mixer variable. Therefore, one has to use Bayesian inference: First, we need to compute the probability that the observed RF outputs are indeed statistically dependent—we denote such probability by \(q(k, s^{(o)})\)—to emphasize that it is a function of the observed RF outputs. Second, we need to estimate the values of the mixers under the two conditions \((v\) for the dependent case, \(v_k\) for the independent case) and perform the division. Eventually, we combine estimates for the two phases to obtain the complex cell response:

\[
R_\theta = \sqrt{k_{\theta,0}^2 + k_{\theta,90}^2}
\]

Because the aim of this paper is to assess the consequences of using statistically optimal V1 surround normalization, we consider three models that differ in the degree to which optimal Bayesian inference is used to perform those estimates and, consequently, in the type of surround modulation. We provide in Appendix I the full derivation for the interested reader and report here only the resulting equations for the V1 response.

\section*{Flexible surround}

First, we consider the Gaussian estimate based on exact inference, which amounts to a flexible surround divisive normalization (Figure 2B):

\[
k_{\theta,\phi}^{\text{flexible surround}} \simeq q(k, s^{(o)}) \frac{k_{\theta,\phi}}{\sqrt{\lambda(k, s^{(o)})}} + (1 - q(k, s^{(o)})) \frac{\hat{k}_{\theta,\phi}}{\sqrt{\lambda(k)}}
\]

where the normalization terms in the denominator are given by

\[
\lambda(k, s^{(o)}) = \sqrt{(k, s^{(o)}) \cdot C_{\kappa,\sigma}^{-1} \cdot (k, s^{(o)})}
\]

\[
\lambda(k) = \sqrt{k^{\top} \cdot C_{\kappa}^{-1} \cdot k}
\]

Note that the first normalization term encompasses both nonspecific suppression from the center and tuned suppression from the surround whereas the second term encompasses only non-specific suppression from the center. An expression for the probability of center and surround being dependent \((q(k,s))\) is derived in Appendix I, but the basic intuition is illustrated in Figure 2B: Images that are very similar (or different) in the center and surround elicit values of \(q\) close to one (or zero).
We call the model of Equation 8 a flexible surround model because the normalization pool can switch, on an image-by-image basis, between a local pool (nonspecific center alone) and a global pool (nonspecific center plus tuned surround). The covariance matrices enter the normalization terms (Equations 9 and 10) and so affect the weight of each RF output to the normalization; proper estimation of the covariances is therefore necessary for the model to result in a statistically optimal estimation of the Gaussian variables. As discussed in Coen-Cagli et al. (2012), an important feature of $C_{\alpha\sigma}$ optimized to natural scenes is the fact that it produces a form of collinear facilitation in the V1 responses. Both the flexibility of the normalization pool and collinear facilitation, play an important role in shaping the statistics of the V1 responses and the selectivity that emerges downstream, which we address in the next section.

**Canonical surround**

The second model we consider is *canonical* surround divisive normalization (Cavanaugh et al., 2002; Heeger, 1992). This corresponds to approximate inference in the MGSM, in which the approximation consists of (a) assuming that $q(k, s^\theta) = 1$ or, in other words, that center and surround RFs are always dependent and therefore to be normalized and (b) that $C_{\alpha\sigma}$ is proportional to the identity matrix rather than being optimized to natural scenes. The estimate of the Gaussian component in this case amounts simply to

$$k_{\theta,\phi}^{(\text{canonical surround})} \sim k_{\theta,\phi} \frac{1}{4} (k, s^\theta)^\top \cdot (k, s^\theta)$$

**No surround**

As a control, we consider also a model that does not include surround modulation at all: namely the energy model for complex cells (Adelson & Bergen, 1985) (which we denote *no surround*). In this case, we simply replace, in Equation 7, the Gaussian estimate $k_{\theta,\phi}$ by the corresponding RF output $k_{\theta,\phi}$:

$$k_{\theta,\phi}^{(\text{no surround})} = k_{\theta,\phi}$$

**V1 output statistics and V2 selectivity**

To analyze the effect of the different surround models on the statistics of V1 outputs, we considered a population comprising units with four orientation preferences ($0^\circ, 45^\circ, 90^\circ, 135^\circ$) and with RF centers densely covering a grid of 18 by 18 pixels. Therefore, the population included a total of 1,296 units. We chose the size of the spatial grid to cover approximately twice the RF size of the individual V1 unit, given the knowledge from physiology that RF size approximately doubles from V1 to V2 (Shushruth et al., 2009). We therefore assume that the statistics over that spatial extent are the most relevant for the computations inside the V2 RF. We collected the responses of such V1 population to 10,000 patches sampled from the same images used for training the MGSM (see the section on models of statistical dependencies between V1 RFs). The size of the patches was 40 by 40 pixels, i.e., large enough to cover the entire population (including the full extent of center and surround linear RFs for each unit).

**Analysis of residual correlations between V1 units**

We computed the correlation matrix of the V1 responses and found both quantitative and qualitative differences between the models. First, the V1 units with no surround were highly correlated all throughout the spatial grid. Figure 3A shows that pairwise correlations were strongest between units with similar orientation and small distance (due to RF overlap) and weakest between orthogonal, distant units. However, even for the latter, the average correlation was substantially larger than zero. This is due to the fact that complex cells convert higher-order dependencies, such as the one induced by global contrast, into correlations. This is also true of other rectifying nonlinearities: In general, higher-order dependencies between two variables can be seen as correlations between nonlinear functions of those variables. Indeed, for both the models that used divisive normalization to reduce the higher-order dependence, output correlations were generally much weaker and the decay with distance much steeper for units with similar orientation preference (Figure 3B and 3C). In addition, the flexible surround normalization introduced negative correlations between units with small distance and orthogonal orientations. Intuitively, this might be due to the fact that when, say, the vertical unit at a given position is surround normalized, the horizontal unit at that same location is unlikely to be also normalized and vice versa. We also found that, for the flexible surround, the correlation between collinear units was markedly larger than between parallel units (Figure 3D and 3F) (note that for the nearest neighbors, correlations are mostly due to RF overlap; hence, the collinearity effect is entirely masked). This is due to the collinear facilitation that results from the weights in the normalization term derived from natural images as explained in the section on V1 responses to visual inputs.
These results illustrate in a simple way the important role of divisive normalization: The normalization removes the global, “uninformative” dependence and makes the finer detail easier to discover. This becomes particularly important when V1 responses are pooled over spatial regions larger than the V1 RFs, such as downstream of V1.

Removal of V1 correlations by PCA and emergence of V2 selectivity

Following the hypothesis that the objective for V2 is to learn without supervision a linear representation that removes the residual correlations in the V1 outputs, we applied PCA to the vectors comprising the V1 responses at all four preferred orientations and 18 by 18 spatial positions (1,296 dimensions in total). Let’s denote the responses of the V1 population to the 10,000 image patches by \( R \) (after subtracting the sample mean), a 1,296 by 10,000 matrix. PCA assumes that the correlations between the dimensions of \( R \) arise from mixing linearly a set of independent sources \( Z \) via an orthonormal matrix of weights \( W \), such that \( R = W \cdot Z \) and \( Z = W^T \cdot R \). We take the columns of \( W \) to define the RFs of the V2 units and \( Z \) their responses to the visual inputs. The weight matrix \( W \) is usually obtained by the eigenvector decomposition of the data covariance: We computed the 1,296 by 1,296 sample covariance matrix.

Figure 3. Effect of divisive normalization on spatial correlations between V1 model responses. Correlations between the responses of the V1 model units to natural image patches for the three V1 models: (A and D) no surround; (B and E) canonical surround; and (C and F) flexible surround. (A through C) Correlation coefficients plotted as a function of the center-to-center distance between V1 RFs. The black line and symbols are for pairs of units with identical orientation preference; gray for pairs with orthogonal orientation preference. Each circle is the correlation coefficient for one pair (we randomly sampled only, at most, 100 pairs at each RF distance for visualization purposes); the continuous lines represent the mean. (D through F) Correlation coefficients between the vertical unit at the center and units at all orientations and spatial positions (sampled every other pixel on an 18 by 18 grid), plotted as a spatial map. The orientation of each bar represents the preference of the unit. The color represents the correlation level (warm color for large positive, cold for large negative). Only statistically significant correlations are shown (\( p < 0.005 \)).
\[ \Sigma = \mathbf{R} \mathbf{R}^\top \] and performed its eigenvector decomposition. The columns of \( \mathbf{W} \) are the eigenvectors of \( \Sigma \) or the principal components (PCs). The collection of weights of each V1 unit to a given PC thus defined the RF of the V2 unit corresponding to that PC.

In Figure 4, we show the first 12 PCs by displaying the weight of each V1 unit to a given PC (for readability, only every other position is shown, hence the nine by nine grid). In the absence of normalization (units on the left), a large portion of the variance (about 43%) is captured by the first PC, which, not surprisingly, essentially measures contrast by assigning similar weight to all positions and orientations. The second component captures global content in one orientation band while the remaining components capture global oriented structure by simply combining all orientations at a given position together and then organizing positive and negative weights across space. Further components not shown in the plot represent higher frequencies and quickly become noisier. The results are qualitatively similar to what would be obtained by PCA on pixels: Because of the translation invariance, the PCs are the Fourier basis.

The units displayed on the right of Figure 4 (flexible surround) show a different pattern of results when the proper surround normalization is taken into account. In this case, the first three PCs capture some global-oriented structure and together only account for 14% of the variance. The next components represent corners (PCs 4 and 5), global orientation discontinuities or texture boundaries (PCs 6 and 7), and Y-junctions (PCs 8 and 9). The units obtained with canonical divisive normalization (center) are somewhat in between as they reveal some extended collinear structure but no corners or Y-junctions as with flexible surround. Figure 5 further illustrates the selectivity of some PCs obtained with flexible surround by showing the maximal patches across the five training images. Visually, the variability of the top patches for each PC is due to two factors: (a) Our model addresses only the linear V2 RF; the PCs respond maximally to conjunctions of edges, but they respond also to the component edges; and (b) the PCs are insensitive to the polarity of the component edges (because the V1 layer is made of complex cells). To obtain a more quantitative comparison, in the following sections, we analyze the performance of the different models on two perceptual tasks.

Perceptual classification tasks

In the previous section, we have illustrated how the form of the nonlinearity adopted in V1 affects the correlations in the responses of a V1 population to natural inputs; we also showed that this affects the linear RF selectivity that emerges following the assumption that downstream neurons aim to decorrelate their inputs. While the differences shown in Figures 3 and 4 are compelling, it is difficult to draw conclusions as to their importance for perception. To address this question, we quantified the ability of the
models to perform perceptual classification tasks. To this aim, we computed the responses of populations of V2 units to visual inputs and used them as inputs to a linear classifier, which we trained with supervision and tested on disjoint subsets of images. Below, we provide the details for each database and task; we note here that there are no extra free parameters in the V1 flexible surround model that we fine-tune with supervision for the classification tasks—the only free parameters are the classifier weights, and there are the same number of them for flexible, canonical, and no surround.

Object recognition

Object recognition tasks are usually formulated as a multiway classification problem, i.e., given an image that contains an object, what category does the object belong to? The answer is produced by a classifier that reads out the responses of the model neural population and decides which class is the most likely to have produced those responses. The classifier is trained on a subset of images with supervision and tested on a disjoint subset. This general approach is commonly used to assess computer vision algorithms, and several labeled databases have been made publicly available for benchmarks. We used two popular examples, Caltech101 (Fei-Fei, Fergus, & Perona, 2004) and NORB (LeCun, Huang, & Bottou, 2004). Our implementation used the steps detailed below, following common practice in object recognition architectures (e.g., Jarrett, Kavukcuoglu, Ranzato, & LeCun, 2009).

The Caltech101 database contains photographic images divided into 102 categories (101 object types plus a background category). We converted the images from RGB to grayscale, resizing the longest side to 164 pixels and padding the other side with zeros. We preprocessed the patches by subtracting the pixel mean and dividing by the pixel standard deviation across the entire dataset. We then computed the responses of all the V1 units with a given orientation preference. This resulted, for each model and image, in a 164 by 164 map that we further cropped to 143 by 143 to eliminate boundary artifacts, low-pass filtered (two by two boxcar), and down-sampled to 71 by 71. We repeated this procedure separately for each of the four V1 preferred orientations. We then created neighborhoods comprising four orientations at nine by nine positions (corresponding to subsampling every other pixel of an 18 by 18 spatial neighborhood, therefore doubling the RF size from V1 to V2), i.e., a 324-dimensional vector at each of the central 52 by 52 positions. We transformed these vectors by PCA to produce the 52 by 52 response maps of the 324 V2 units. For each V2 unit, we then low-pass filtered (six by six boxcar) and down-sampled (four by four factor) and retained the center six by six responses, which we eventually fed into a multinomial logistic regression classifier as in Jarrett et al. (2009). We trained the classifier using 30 randomly selected images per category and then tested it on a disjoint set.
of another 30 images, or as many as available, per category (the least represented category contains 31 images). We optimized the MGSM parameters, without supervision as described above, on a random subset of the images used to train the classifier. For this dataset, it is common practice to compute the classification performance for each class separately and then average across classes. We repeated this for seven random splits of the images into training/test sets.

The second database, NORB, contains photographs of five object categories taken under different illumination conditions and from different points of view. We used similar steps as above except in this case the original image size was 96 by 96 pixels, and we used different neighborhood sizes and down-sampling factors, so the resulting V2 population included 100 V2 units at eight by eight positions. The database provides 24,300 training and as many test images, and so we did not perform cross-validation in this case.

The results are shown in Figures 6 and 7. In agreement with previous observations in the computer vision literature that used canonical normalization (Jarrett et al., 2009), we found that classification based on the V2 responses was better than with the V1 responses directly, and using divisive normalization always helped as the models with canonical and flexible surround both performed substantially better than the model without surround. Interestingly, only a small fraction of the PCs was needed to achieve maximal performance, suggesting that the features learned without supervision are nonetheless informative about object identity. We found also that the performance could be improved by adding simple nonlinearities at the V2 outputs (namely unspecific divisive normalization and z scoring) as is often done in the computer vision literature (Figures 6B and 7B); this however did not change the relationship between the three models.

Figure/ground judgment

Figure/ground organization refers to the perceptual operation of assigning a contour to one of the two regions that it separates. It has been shown that neurons selective for the side of figure (or border ownership) exist in primate V2 (Zhou et al., 2000), and this selectivity relies on global context integration, most likely through feedback signals (Zhang & Von Der Heydt, 2010). However, it has also been shown that this operation relies at least partly on local cues that may be processed in a bottom-up fashion (Fowlkes, Martin, & Malik, 2007; Ren, Fowlkes, & Malik, 2006). Here, we followed the approach of Fowlkes et al. (2007) and formulated the figure/ground judgment as a binary classification problem, i.e., is the figure region left or right of the contour? The work of Fowlkes et al. has shown that, in small image patches containing a contour separating two regions, there are three low-level cues that are informative about figure/ground: namely, convexity, which region is above the other, and relative size. Each of these cues can be used to classify the patches with accuracies that depend on the overall patch size (respectively, between 52% and 60% for convexity, 62% and 64% for lower region, 53% and 68% for relative size). They also reported that simpler properties, such as contrast or brightness of the two regions, are on average matched.
between the two categories in the database and therefore not informative.

To assess the ability of our V2 units to perform this task, we randomly sampled small patches from images taken from the Berkeley Segmentation Dataset (http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/) for which ground truth information about both object segmentation and figure ground assignment are provided (http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/fg/). We then measured the performance of a logistic regression classifier based on the outputs of the two-stage architecture. We first sampled randomly 20 patches, centered on contour elements, of size 39 by 39 pixels from each of the 200 images. Each patch was rotated such that the central segment of the contour was oriented vertically as in Fowlkes et al. (2007). We followed similar steps as for the object recognition tasks, but we used different neighborhood sizes and down-sampling steps. The neighborhoods comprised four orientations and five by five positions (corresponding approximately to subsampling every fourth pixel of a 20 by 20 spatial neighborhood, again resulting in a doubling of RF size from V1 to V2), i.e., a 100-dimensional vector at each of the center three by three positions. These vectors were transformed by PCA and fed into a logistic regression classifier based on the outputs of the two-stage architecture. We first sampled randomly 20 patches, centered on contour elements, of size 39 by 39 pixels from each of the 200 images. Each patch was rotated such that the central segment of the contour was oriented vertically as in Fowlkes et al. (2007). We followed similar steps as for the object recognition tasks, but we used different neighborhood sizes and down-sampling steps. The neighborhoods comprised four orientations and five by five positions (corresponding approximately to subsampling every fourth pixel of a 20 by 20 spatial neighborhood, again resulting in a doubling of RF size from V1 to V2), i.e., a 100-dimensional vector at each of the center three by three positions. These vectors were transformed by PCA and fed into a logistic regression classifier as in Fowlkes et al. (2007). We first sampled randomly 20 patches, centered on contour elements, of size 39 by 39 pixels from each of the 200 images. Each patch was rotated such that the central segment of the contour was oriented vertically as in Fowlkes et al. (2007).

Figure 8A shows the cross-validated performance obtained using the different V1 models and different numbers of PCs. The performance level achieved by the flexible surround (76%) was larger than the values reported in Fowlkes et al. (2007), using a combination of the local cues described above (ranging between 52% and 74%), and in Ren et al. (2006), using local shapemes (approximately 65%). Interestingly, (a) only the flexible surround, not the canonical surround, performed significantly better than no surround in this case, and (b) only the V2 populations based on flexible surround and no surround, but not on canonical surround, performed better than the corresponding V1 populations. Furthermore, many fewer components were needed to reach the maximal performance when using the flexible surround or no surround than when using the canonical surround. This suggests that the PCs learned by flexible and no surround models are more informative for this task. We confirmed this in two ways: First, Figure 8A shows that removing the top 40 to 60 PCs from those models decreases classification performance close to chance as opposed to the canonical surround model whose V2 stage contains a large amount of information in the high-rank (low-variance) modes of the covariance matrix. Second, we quantified the importance of each PC for the classification by the significance of the learned classifier weight (i.e., the p value for the hypothesis that the weight is different from zero): This measure correlated with the
PC rank significantly more for the flexible surround (68% c.i. of the median correlation [.37 .41]) than for canonical ([.24 .25]) and no surround ([.20 .27]).

**Discussion**

We addressed the influence of V1 surround normalization on the computations that take place downstream, using a modeling framework that explains normalization from a principle of statistical optimality. We focused on three specific issues: the correlations between V1 responses, the emerging selectivity in V2, and classification performance on perceptual tasks. On all three, we found both qualitative and quantitative differences when using versions of surround normalization that were optimal (flexible surround) or suboptimal (canonical surround) or when using no surround.

First, both versions of surround normalization reduced substantially the correlations between V1 responses to natural images. However, only the flexible surround produced negative correlations between V1 units with largely different orientation preferences and stronger correlations between collinear than noncollinear units. Second, differences in the V1 correlation structure implied differences in the feature selectivity that emerged at the next cortical stage in the models. With the flexible surround, we found V2 units selective for corners, Y-junctions, and texture boundaries; this is in contrast to no surround (and to some extent canonical)

**Correlations between V1 responses to natural images**

Many theoretical models of V1 RFs are based on principles of efficiency or redundancy reduction (Barlow, 1961; Simoncelli & Olshausen, 2001; Zhaoping, 2006), according to which V1 (and, more generally, sensory) neurons aim to produce responses to natural inputs that are as independent as possible. Indeed different flavors of this idea have shown that V1-like linear RFs can be learned from natural images (e.g., Bell & Sejnowski, 1997; Olshausen & Field, 1996). However, natural images contain both (a) additional structure that extends well beyond the V1 RF size (e.g., beyond the size of the image patches used to train the above models) and (b) higher-order dependencies that cannot be removed by linear RFs (Schwartz & Simoncelli, 2001). Our starting point here was the simple observation that such
structure introduces dependencies between the responses of V1 neurons with spatially separated RFs over regions that include neurons that are likely to project to V2 and therefore determine V2 input statistics. Therefore, using models based on major V1 properties (orientation selectivity, phase invariance, divisive normalization), we provided a characterization of the types of dependencies that one might expect to find experimentally in V1 neurons. In Figure 3A, we have shown that the two above-mentioned sources of dependence produce strongly correlated responses between model V1 complex cells even when their RFs are completely nonoverlapping and have orthogonal orientation preferences. We have also shown that canonical surround divisive normalization greatly reduces the output correlations (Figure 3B) by removing a higher-order dependence between linear RFs that is induced by a global, “uninformative” image property, such as contrast (Schwartz & Simoncelli, 2001). A similar effect of divisive normalization on noise correlations in a spiking model of V1 neurons was reported by Tripp (2012), and recently, Beck, Latham, and Pouget (2011) have linked normalization to marginalization of uninformative nuisance variables.

Our main finding, however, is that further differences in the structure of correlations emerge when we consider flexible surround normalization optimized for natural images. In this case, we observed two effects: First, we observed negative correlation between V1 units with overlapping RFs and orthogonal orientation preference (Figure 3C) as a consequence of the flexible normalization. On any given input, if, say, the vertical neuron is surround suppressed, then the overlapping horizontal neuron is more likely not to be surround suppressed. Second, we found stronger correlations between non-overlapping V1 units with collinear RFs than between parallel (but not collinear) units; in our previous work (Coen-Cagli et al., 2012; Schwartz & Coen-Cagli, 2013), we suggested that this difference could contribute to perceptual effects of collinear facilitation and could be enhanced by perceptual learning and attention. Both these distinctive results could be tested experimentally using current technology by recording simultaneously from V1 populations and characterizing signal correlations (i.e., correlations between the responses to an ensemble of images) with large natural images that cover the center and surround of the population RFs.

V2 selectivity by decorrelation of V1 responses

We studied the selectivity that emerges downstream of V1, say, in V2, under the assumption that V2 neurons aim to decorrelate the inputs they receive from V1. We considered a direct implementation of this idea by learning the linear pooling of V1 outputs into V2 RFs via PCA, a technique that achieves full decorrelation. As expected from the different structures of V1 correlations discussed in the previous section, the V2 RFs exhibited qualitatively different properties depending on the type of V1 surround normalization considered. In Figure 4, we illustrated that among the first PCs following canonical surround some contain pooling of a particular orientation over extended regions of space, resulting in collinear units. However, the flexible surround produced a richer repertoire of features (Figures 4 and 5), some of which have been observed experimentally in V2: selectivity for corners, Y-junctions, and texture boundaries. We also noted that using the V1 complex cell model with no surround led to PCs analogous to those obtained by PCA directly on pixels, i.e., the Fourier basis (a known result due to the translation-invariance of the covariance matrix).

Clearly, our framework encompasses only the initial stage of a full V2 model, namely, the emergence of the linear RF selectivity. Adjacent or partly overlapping V2 linear RFs will exhibit further dependencies, analogous to the V1-stage RFs. Our framework proposes one possible way to eliminate the V1 dependence in two steps: First, the surround divisive normalization removes the higher-order dependence at the V1 stage; second, the residual correlation is removed downstream by pooling the responses across space and performing PCA. This roughly corresponds to a generative model in which top-level, independent Gaussian variables (the PCs or V2 features) generate the correlations in the outputs of the first level (V1 complex cells) via linear combinations. Those, in turn, generate the underlying local Gaussian variables (V1 simple cells) by randomly splitting the energy between the two phases. Finally, the local kurtotic variables (oriented RFs) are obtained by multiplication of the local Gaussians by the positive mixer. A more complete hierarchical model could iterate these steps again by including V2 nonlinearities (e.g., rectification and normalization), followed by pooling over larger regions and decorrelation leading into the next stage. This is an important future direction to explore. A crucial issue is that, given the expansion in the number of features per spatial location compared to V1, there is no natural way to combine such features across space to form the normalization pool for any given V2 unit. This requires a more sophisticated generative model, for instance, one in which the composition of the normalization pools can be inferred online as in Schwartz et al. (2006).

The emergence of V2-like features, in particular corner units and pooling over orientations, in models of image statistics has been reported by few authors in recent years (Lee et al., 2008; Lindgren & Hyvaarininen, 2006; Malmir & Shiry Ghidary, 2009; Nuding & Zetzsche, 2007; Shan et al., 2007; Spratling, 2011). Our goal was to clarify the role of V1 surround norma-
zation because this operation is a major feature of V1 cortical neurons (Carandini & Heeger, 2012) as well as of accurate models of image statistics (Theis et al., 2012). The normalization was included more or less explicitly in Shan et al. (2007), Nuding and Zetzsche (2007), Malmir and Shiry Ghidary (2009), and Sprotling (2011) but not in Lee et al. (2008) and Lindgren and Hyvarinen (2006) in which maximization of sparseness was the objective (although in deep networks the division could be implemented by concatenation of a layer performing the logarithm followed by one performing subtraction). There are also many other differences between those models, making it difficult to discern the exact role of surround normalization. Our results show rather directly that the form of the normalization influences the statistics of the inputs to V2 and, therefore, V2 selectivity. This is probably true also on time scales much shorter than the development of V2 RFs, and we suggest therefore that experiments probing midlevel visual areas might benefit from using stimuli rich enough to engage the full V1 surround nonlinearity.

Perceptual classification performance

We quantified the differences between the V1 normalization schemes by testing how well the respective V2 units could be used to perform perceptual classification tasks. We found that the two-stage architectures based on V1 models with surround normalization perform substantially better than those without surround in two object recognition tasks (Figures 6 and 7). This result confirms a previous finding by Jarrett et al. (2009), but our approach based on image statistics provides further insight into why this is the case. In our framework, surround normalization allows V1 neurons to accurately represent natural images by discarding global, uninformative image structure, such as contrast; in more technical terms, normalization amounts to Bayesian inference in the MGSM generative model of images. The better performance achieved by models that include normalization can thus be explained by the reduced V1 output correlations (Figure 4; also Tripp [2012] provided similar results in orientation discrimination with noisy neurons) and the fact that features learned downstream without supervision are more informative about object identity (Figures 6 and 7 show that only the first few, high-variance V2 PCs are informative). Shan and Cottrell (2008) used a similar approach and also noticed that informative features could be learned in a two-stage system with a nonlinearity that produced Gaussianization of the V1 outputs; they also suggested that the improved classification performance can be partly explained by the expansion of dimensionality in the features represented by the second stage (a sort of “Kernel trick”). We note also that the object recognition performance levels we reported fall short of the state of the art (e.g., Boureau, Bach, LeCun, & Ponce, 2010; Coates, Lee, & Ng, 2011; Jarrett et al., 2009; Pinto, Cox, & Dicarlo, 2008). A number of known factors are likely to play a role, such as using a finer sampling of orientations and spatial scales (Pinto et al., 2008) and positions (Coates, Lee, & Ng, 2011), using sparse features (Boureau et al., 2010), using max rather than average pooling (Jarrett et al., 2009); moreover, in our implementation, there are no free parameters to be learned with supervision (other than the classifier). However, the maximal performance we obtain is similar to that reported using implementations whose details are more comparable with ours (e.g., in Jarrett et al. [2009], a two-stage system initialized with Gabor filters at four orientations remains well below 60% on Caltech101).

We further explored the idea that it is the statistical optimality of normalization that matters, and we showed that indeed flexible (optimal) normalization led to slight but significant improvements in performance over canonical (suboptimal) normalization. Interestingly, when tested on a task closer to midlevel visual processing, i.e., figure/ground judgment, we found that only the flexible, not the canonical, surround performed better than no surround (Figure 8). The best performance achieved by the flexible surround model (76%) surpassed that reported by Fowlkes et al. (2007), using combinations of carefully designed low-level local cues, and in Ren et al. (2006), using local “shapemes.” Note that the figure/ground task we considered was originally conceived by Fowlkes et al. (2007) to test the validity of local cues: Therefore, the image patches never included entire objects but only small parts so as to remain globally ambiguous about which side is figure. Indeed human observers performed well below 100%. An important future direction to explore, in the context of a generative model of images, is the integration of global information from the V2 RF surround, which is thought to mediate border-ownership signals in single V2 neurons (Zhou et al., 2000; Zhang & Von Der Heydt, 2010) and was included in previous mechanistic models (Nishimura & Sakai, 2007; Ren et al., 2006; Zhaoping, 2005). The fact that the flexible normalization improves the performance in both object recognition and figure/ground, as it can cope with the different nature of the images used, resonates with the suggestion of Han and Vasconcelos (2010) to use intelligent normalization. In that case, though, the normalization is trained with supervision to optimize discrimination as opposed to the unsupervised training of the MGSM.

The main finding that canonical normalization did not help in figure/ground judgment can be explained by the fact that this task relies on pixels around the
The border between two distinct regions (figure and ground), which are more likely to be independent (hence the need for the flexible surround). Therefore, RF outputs are expected to be more independent on average than on the object datasets (in which most of the images contain a single object); indeed, we found that the mean probability that center and surround were dependent was 0.78 and 0.7 on Caltech101 and NORB, respectively, and 0.62 on the figure/ground dataset. Thus, performing surround normalization on every input (as per the canonical surround) is a worse approximation on the figure/ground dataset; as a consequence, the high-variance PCs are less informative for this task, and information is spread across many low-order modes of the covariance (Figure 8B). Such bias is not due to the particular dataset but rather to the nature of the task: Because figure-ground discrimination is a task that is more closely related to midlevel vision, we propose that it is an informative test for V2 models.

Keywords: image statistics, divisive normalization, midlevel vision, visual cortex

Acknowledgments

This work was supported by the NIH grant CRCNS-EY021371, the Army Research Office grant 58760LS, and the Alfred P. Sloan Foundation (OS). We are very grateful to P. Dayan, A. Kohn, and E. Simoncelli for discussion.

Commercial relationships: none.
Corresponding author: Ruben Coen-Cagli.
Email: ruben.coencagli@unige.ch.
Address: Department of Basic Neuroscience, University of Geneva, Geneva, Switzerland.

Footnote

1 Note that the binary mixture is an approximation to the full complexity of natural images because image patches for which the center and surround are independent may contain different structures in the surround (e.g., occlusions or adjacency or transparency involving multiple objects and textures). More complex models have addressed such heterogeneity (e.g., Karlkin & Lewicki, 2009; Schwartz et al., 2006). Here instead we have chosen to explore the consequences of a much simpler, and biologically more plausible, implementation that can still capture a vast repertoire of effects of V1 surround modulation.

References


Olshausen, B., & Field, D. (1996). Emergence of


Appendix I

MGSM generative model

A binary MGSM defines the distribution of the observable variables (in our case, the RF outputs $k$, $s^{(0)}$) as a probabilistic mixture of two base distributions. In the section on models of statistical dependencies between V1 RFs, we introduced the two mixture components, in which RFs are defined by Equations 5 and 6, respectively, but we did not write down explicitly the MGSM distribution. To do so formally, we introduce a binary assignment variable $\xi$ with prior distribution $p(\xi)$, such that $\xi = \xi_1$ corresponds to the first mixture component and $\xi = \xi_2$ to the second; the joint distribution of $k$, $s^{(0)}$ can be written as

$$p(k, s^{(0)}) = p(\xi_1)p(k, s^{(0)}|\xi_1) + p(\xi_2)p(k, s^{(0)}|\xi_2)$$

Inference in the MGSM proceeds in two steps. First, for an input image patch, the posterior probability $p(\xi|k, s^{(0)})$ that the surround shares a common mixer with the center is inferred (in the main text, we denoted it, for simplicity, $q(k, s^{(0)})$); this is proportional via Bayes rule to Equation 16. Then, we compute the expected value of the Gaussian components $k_{\theta,\phi}k_{\theta,\phi}$ (which we relate to V1 neural responses via Equation 7).

$$E[k_{\theta,\phi}|k, s^{(0)}] = \int dk_{\theta,\phi}p(k_{\theta,\phi}|k, s^{(0)})$$

where the conditional expected values under $\xi = \xi_1$ and $\xi = \xi_2$ are obtained directly by solving the integrals:

$$E[k_{\theta,\phi}|k, s^{(0)}, \xi_1] = \int dk_{\theta,\phi}p(k_{\theta,\phi}|k, s^{(0)}, \xi_1)$$

$$E[k_{\theta,\phi}|k, s^{(0)}, \xi_2] = \int dk_{\theta,\phi}p(k_{\theta,\phi}|k, s^{(0)}, \xi_2)$$

The model parameters are the prior on the hidden assignment variable, $p(\xi)$, and the covariances $\Theta \triangleq \{C_{\kappa\eta}, C_{\kappa}, C_{\eta}\}$. To find the optimal parameter values, we used a generalized expectation maximization algorithm. Expectation maximization amounts to maximizing a lower bound on the log-likelihood of the data (the RF outputs to a large collection of natural image patches as described in the section on models of statistical dependencies between V1 RFs) by iterating

$$\lambda(s^{(0)}) = \sqrt{\lambda(s^{(0)})}C^{-1}\lambda(s^{(0)})$$

where $C$ is the modified Bessel function of the second kind; $m_k = dim(k)$; $m_s = dim(s^{(0)})$; $m = m_k + m_s$; the terms $\lambda(k, s^{(0)})$ and $\lambda(k)$ are defined in Equations 9 and 10 respectively; and

$$k_{\theta,\phi} = p(\xi_1|k, s^{(0)})E[k_{\theta,\phi}|k, s^{(0)}, \xi_1] + p(\xi_2|k, s^{(0)})E[k_{\theta,\phi}|k, s^{(0)}, \xi_2]$$

We point out that the inference is solved analytically and thus is of small computational cost.

Appendix II: MGSM learning

The model parameters are the prior on the hidden assignment variable, $p(\xi)$, and the covariances $\Theta \triangleq \{C_{\kappa\eta}, C_{\kappa}, C_{\eta}\}$. To find the optimal parameter values, we used a generalized expectation maximization algorithm. Expectation maximization amounts to maximizing a lower bound on the log-likelihood of the data (the RF outputs to a large collection of natural image patches as described in the section on models of statistical dependencies between V1 RFs) by iterating
two steps until convergence: First, given the current parameters' values, approximate the true posterior distribution of the hidden variable \( p(\xi_1 | \mathbf{k}, \mathbf{s}^{(0)}) \); second, maximize the so-called complete-data log-likelihood, namely, the expectation of \( \log p(\mathbf{k}, \mathbf{s}^{(0)}, \xi) \) under the posterior of \( \xi \). In the E step, we use Bayes rule to compute the estimate, \( Q \), of the posterior distribution over \( n_1 \), given the observed RF outputs and the previous estimates of the parameters, which we denote by \( p(n_1)^{\text{old}} \) and \( \Theta^{\text{old}} \):

\[
Q(\xi_1) = p(\xi_1 | \mathbf{k}, \mathbf{s}^{(0)}; \Theta^{\text{old}}) \\
\propto p(\xi_1)^{\text{old}} p(\mathbf{k}, \mathbf{s}^{(0)} | \xi_1; \Theta^{\text{old}}) \quad (24)
\]

\[
Q(\xi_2) = p(\xi_2 | \mathbf{k}, \mathbf{s}^{(0)}; \Theta^{\text{old}}) \\
\propto (1 - p(\xi_1)^{\text{old}}) p(\mathbf{k}, \mathbf{s}^{(0)}; | \xi_2; \Theta^{\text{old}}) \quad (25)
\]

In the M–step we increase the complete-data Log Likelihood, namely:

\[
f = Q(\xi_1) \log \left[ p(\xi_1) p(\mathbf{k}, \mathbf{s}^{(0)} | \xi_1; \Theta) \right] \\
+ Q(\xi_2) \log \left[ (1 - p(\xi_1)) p(\mathbf{k}, \mathbf{s}^{(0)} | \xi_2; \Theta) \right] \quad (26)
\]

Solving \( \partial f / \partial p(\xi_1) = 0 \), we obtain \( p(\xi_1)^* \doteq \arg\max_{p(\xi_1)} f = Q(\xi_1) \). For the other terms, we obtained analytical expressions for the gradient, e.g.,

\[
\frac{\partial f}{\partial C_{k\sigma}} = Q(\xi_1) \left( \frac{C_{k\sigma}}{2} - \frac{1}{2\lambda(k, \mathbf{s}^{(0)})} \frac{B(-m_{k\sigma}; \lambda(k, \mathbf{s}^{(0)}))}{B(1 - \frac{m_{k\sigma}}{2}; \lambda(k, \mathbf{s}^{(0)}))} \times (k, \mathbf{s}^{(0)}) (k, \mathbf{s}^{(0)})^T \right) \quad (27)
\]

(similar expressions hold for the other partial derivatives) but could not solve the maximization analytically and adopted a numerical procedure to maximize \( f \) with reference to the covariance matrices. We used partial M steps for each covariance matrix separately (i.e., expectation conditional maximization).