INSTITUTE OF PHYSICS PUBLISHING NETWORK: COMPUTATION IN NEURAL SYSTEMS

Network: Comput. Neural Syst. **14** (2003) 579–593 PII: S0954-898X(03)53779-4

Optimal coding through divisive normalization models of V1 neurons

Roberto Valerio and Rafael Navarro

Instituto de Óptica 'Daza de Valdés' (CSIC), Serrano 121, 28006, Madrid, Spain

E-mail: r.valerio@io.cfmac.csic.es

Received 13 September 2002, in final form 12 February 2003
Published 23 June 2003
Online at stacks.iop.org/Network/14/579

Abstract

Current models of the primary visual cortex (V1) include a linear filtering stage followed by a gain control mechanism that explains some of the nonlinear behaviour of neurons. This nonlinear stage consists of a divisive normalization in which each linear response is squared and then divided by a weighted sum of squared linear responses in a certain neighbourhood plus a constant. Simoncelli and Schwartz (1999 *Adv. Neural Inform. Process. Syst.* **11** 153–9) have suggested that divisive normalization reduces the statistical dependence between neuron responses when the weights are adapted to the statistics of natural images, which is consistent with the efficient coding hypothesis. Nevertheless, there are still important open issues, such as, for example, how to obtain the values for the parameters that minimize statistical dependence? Does divisive normalization give a total independence between responses? In this paper, we present the general mathematical formulation of the first of these two questions. We arrive at an expression which permits us to compute, numerically, the parameters of a quasi-optimal solution adapted to an input set of natural images. This quasi-optimal solution is based on a Gaussian model of the conditional statistics of the coefficients resulting from projecting natural images onto an orthogonal linear basis. Our results show, in general, lower values of mutual information, that is, responses are more independent than those provided by previous approximations.

1. Introduction

Different authors have shown that nonlinear behaviour of V1 neurons could be explained by including a gain control mechanism, known as divisive normalization, after a linear filtering stage (Bonds 1989, Geisler and Albrecht 1992, Heeger 1992, Carandini *et al* 1997). In this nonlinear stage, the linear inputs, c_i, are squared and then divided by a weighted sum of squared neighbouring responses in space, orientation and scale, plus a regularizing constant:

$$r_i = \frac{c_i^2}{d_i^2 + \sum_j e_{ij} c_j^2}. \tag{1}$$

0954-898X/03/030579+15$30.00 © 2003 IOP Publishing Ltd Printed in the UK

Network Downloaded from informahealthcare.com by Staats- und Universitaetsbibliothek Hamburg on 12/27/14
For personal use only.

RIGHTSLINK

The hypothesis that sensory systems are adapted to the signals to which they are exposed implies that the parameters of the divisive normalization, that is, the constant d_i^2 and the weights $\{e_{ij}\}$ in equation (1), are related to the statistics of natural images. More specifically, the efficient coding hypothesis states that an efficient group of neurons should be able to encode as much information as possible, or in other words, that all their responses should be statistically independent. Different versions of this hypothesis were formulated by Attneave (1954), Barlow (1961), who proposed that early sensory neurons remove statistical redundancy in the input signal, and other authors (Atick 1992, Field 1994, Laughlin 1981, Rieke *et al* 1995, Van Hateren 1992).

Very recently, Schwartz and Simoncelli (2001) presented a statistically derived divisive normalization model. In addition to its utility to characterize the nonlinear response properties of neurons in sensory systems, and thus to demonstrate that early neural processing is well matched to the statistical properties of the stimuli, they showed empirically that the statistical normalization model strongly reduces pairwise statistical dependences between responses.

It is well known that linear decompositions, such as those based on independent component analysis (ICA), which has been used to explain images in terms of sparse and statistically independent components (Olshausen and Field 1996, 1997, Bell and Sejnowski 1997, Van Hateren and van der Schaaf 1998, Lewicki and Olshausen 1999), cannot completely eliminate higher-order statistical dependences (e.g. Wegman and Zetzsche 1990, Simoncelli 1997, Simoncelli and Schwartz 1999). Figure 1 shows a typical conditional histogram of the resulting coefficients c_i of applying an orthogonal wavelet linear transform to a natural image. We can see that coefficients are strongly decorrelated, since the expected value of the ordinate is approximately zero, independent of the abscissa. However, the 'bowtie' shape of the histograms reveals that these coefficients are not statistically independent, since the variance of the ordinate scales with the squared value of the abscissa. All pairs of coefficients taken either in space, frequency or orientation always show this type of dependence (e.g. Wegman and Zetzsche 1990, Simoncelli 1997, 1999), while the strength varies depending on the specific pair chosen. The form of the histograms is robust across a wide range of images and different pairs of coefficients. In addition, this is a property of natural images and is not of the particular basis functions chosen (Simoncelli and Olshausen 2001). Several distributions have been proposed to model these conditional statistics of the coefficients obtained by projecting natural images onto orthogonal linear basis (e.g., Wainwright *et al* 2001a, Schwartz and Simoncelli 2001). Here we will consider the Gaussian model of Schwartz and Simoncelli (2001). Other closely related models have been proposed out of this context. Among them are the Gaussian scale mixture (GSM) models (Wainwright and Simoncelli 2000, Wainwright *et al* 2001b), which model a group of nearby wavelet coefficients as the product of two independent random variables: a scalar random variable and a zero-mean Gaussian random vector. Although GSM models permit us, in theory, to get statistically independent coefficients through a nonlinear normalization, these models are not considered here because, first, this normalization procedure differs from the classical divisive normalization in V1 models and, second, the Gaussian-distributed output coefficients obtained by this procedure do not seem to be a proper model of the sparse responses of V1 neurons (for a discussion on sparse coding in V1 see, e.g., Olshausen and Field 1997, Olshausen 2002).

In summary, from the efficient coding hypothesis it follows that, ideally, neural responses should be statistically independent. However, the responses of linear basis functions to natural images exhibit statistical dependences, even when the basis functions are chosen to obtain maximal statistical independence. Moreover, these higher-order dependences cannot be removed through further linear processing. Simoncelli and colleagues suggested that these dependences might be dramatically reduced by considering a statistically derived nonlinear

Network Downloaded from informahealthcare.com by Staats- und Universitaetsbibliothek Hamburg on 12/27/14
For personal use only.

RIGHTSLINK

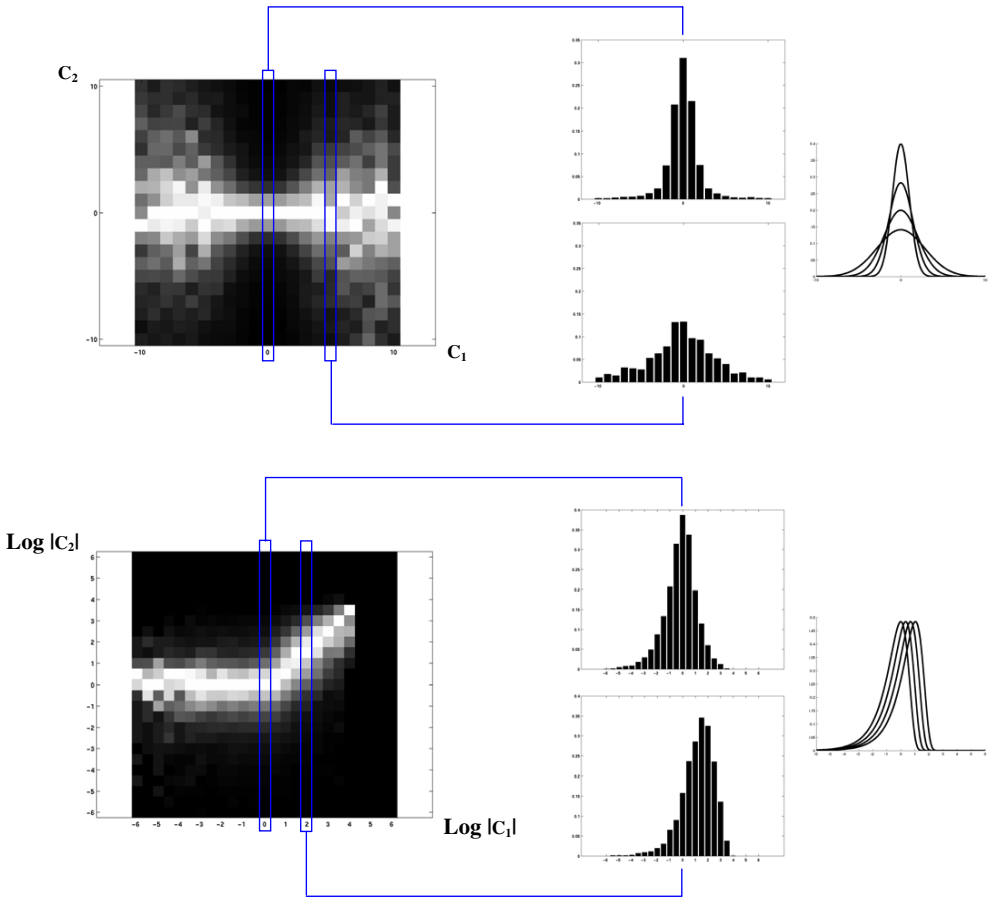

Figure 1. Upper panel: conditional histogram of a wavelet coefficient (c_2) in the lowest scale vertical subband and its right down spatial neighbour (c_1) of the 'Boats' image, two cross sections of this histogram and a sample of Gaussian densities. Lower panel: the same but using logarithmic variables.

divisive normalization model of V1 neurons. They have shown quite convincing results in a series of publications (Simoncelli and Schwartz 1999, Wainwright *et al* 2001a, Schwartz and Simoncelli 2001), but there are still different open issues including the lack of a rigorous mathematical demonstration and the lack of a method for obtaining the optimal values of parameters to get maximum independence, or even some apparent inconsistencies with physiological models (Heeger 1992, Teo and Heeger 1994, Watson and Solomon 1997), such as the fact of not considering the coefficient c_i in the denominator in equation (1).

In this paper, we depart from a general formulation of the problem of obtaining the values for the divisive normalization parameters that minimize the statistical dependence of responses. We use the mutual information (MI), also called Kullback–Leibler divergence (Kullback and Leibler 1951, Kullback 1959), as a measure of statistical independence. Then we particularize the general expression, assuming a Gaussian model for the conditional statistics of the linear coefficients, which has been shown to fit reasonably well the conditional histograms of natural images (Schwartz and Simoncelli 2001). Thus, we arrive at an approximate expression, easier to solve numerically, which permits us to obtain quasi-optimum parameters, so that the divisive

Network Downloaded from informahealthcare.com by Staats- und Universitaetsbibliothek Hamburg on 12/27/14
For personal use only.

RIGHTSLINK

normalization provides a value of mutual information close to the global minimum. Even though this value is greater than zero, it is very small. In addition, we have compared this optimal solution with previous approximations using natural images. Our method generally provided the lowest mutual information between responses.

2. Optimal divisive normalization

We will use optimal divisive normalization to refer to that defined by the values of the parameters (constant d_i^2 and weights $\{e_{ij}\}$ in equation (1)) that yields the minimal mutual information, or equivalently minimizes statistical dependence, between normalized responses for a set of natural images. This 'training set' can have an arbitrary number of images.

In this section, the goal is to arrive at an explicit mathematical expression of these optimal values of parameters. For this purpose, we start formulating the optimal condition in the general case. Then we focus on the particular case of Gaussian conditional statistics of the coefficients obtained by projecting natural images onto an orthogonal linear basis (section 2.2). This is a realistic assumption for natural images. Under this Gaussian assumption, we arrive at a quasi-optimal condition (section 2.3) that is much easier to compute and that is close to the global minimum of the mutual information.

2.1. General optimum condition

For a general formulation, we depart from the definition of the mutual information, also called Kullback–Leibler divergence (Kullback and Leibler 1951, Kullback 1959) of the normalized responses r_i:

$$\mathrm{MI}(r_1, r_2, \ldots, r_n) = \int_0^{+\infty} \cdots \int_0^{+\infty} p(r_1, r_2, \ldots, r_n) \times \log\left(\frac{p(r_1, r_2, \ldots, r_n)}{p(r_1) \cdot p(r_2) \cdots p(r_n)}\right) \mathrm{d}r_1\, \mathrm{d}r_2 \cdots \mathrm{d}r_n. \tag{2}$$

Then we use the change of variable theorem (Papoulis 1991) to express the conditional probability density of the normalized responses $p(r_1, r_2, \ldots, r_n)$ in terms of that of the squared linear inputs $p(c_1^2, c_2^2, \ldots, c_n^2)$:

$$p(r_1, r_2, \ldots, r_n) = \frac{p(c_1^2, c_2^2, \ldots, c_n^2)}{|\det[\mathbf{J}(c_1^2, c_2^2, \ldots, c_n^2)]|} \tag{3}$$

where $\mathbf{J}(c_1^2, c_2^2, \ldots, c_n^2)$ represents the Jacobian matrix. The Jacobian determinant of equation (3) can be computed as:

$$|\det[\mathbf{J}(c_1^2, c_2^2, \ldots, c_n^2)]| = \left|\det\left[\frac{\partial(r_1, r_2, \ldots, r_n)}{\partial(c_1^2, c_2^2, \ldots, c_n^2)}\right]\right| = \frac{r_1 \cdot r_2 \cdots r_n}{c_1^2 \cdot c_2^2 \cdots c_n^2} |\det[\mathbf{Id} - \mathbf{R} \cdot \mathbf{E}]| \tag{4}$$

where **Id** denotes the identity matrix, **E** is the matrix of weights $\{e_{ij}\}$ and **R** is the diagonal matrix of the normalized responses r_i.

Substituting equations (3) and (4) into (2) and changing the integration variables we get:

$$\mathrm{MI}(r_1, r_2, \ldots, r_n) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} p(c_1, c_2, \ldots, c_n) \times \log\left(\frac{c_1^2 \cdot c_2^2 \cdots c_n^2 \cdot p(c_1^2, c_2^2, \ldots, c_n^2)}{r_1 \cdot r_2 \cdots r_n \cdot p(r_1) \cdot p(r_2) \cdots p(r_n) \cdot |\det[\mathbf{Id} - \mathbf{R} \cdot \mathbf{E}]|}\right) \mathrm{d}c_1\, \mathrm{d}c_2 \cdots \mathrm{d}c_n. \tag{5}$$

Network Downloaded from informahealthcare.com by Staats- und Universitaetsbibliothek Hamburg on 12/27/14
For personal use only.

RIGHTSLINK

Therefore the general expression for the set of parameters that minimizes the mutual information is:

$$\{d_i^2, e_{ij}\} = \arg\min_{\{d_i^2, e_{ij}\}} \mathrm{MI}(r_1, r_2, \ldots, r_n)$$
$$= \arg\max_{\{d_i^2, e_{ij}\}} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} p(c_1, c_2, \ldots, c_n) \log(r_1 \cdot r_2 \cdots r_n \cdot p(r_1)$$
$$\times\, p(r_2) \cdots p(r_n) |\det[\mathbf{Id} - \mathbf{R} \cdot \mathbf{E}]|)\, \mathrm{d}c_1\, \mathrm{d}c_2 \cdots \mathrm{d}c_n. \tag{6}$$

We can go one step beyond, since the probability densities of the normalized outputs $p(r_i)$ can be expressed in terms of the corresponding conditional densities of the linear coefficients c_i, as follows. Let us apply the change of variable theorem again to obtain the conditional densities of the normalized output responses r_i, given the squared linear coefficients $\{c_j^2\}$ $(j \neq i)$:

$$p(r_i|\{c_j^2\}) = \frac{2p(c_i|\{c_j^2\})}{\left|\partial r_i(c_i)/\partial c_i\right|} = \frac{1}{\sqrt{r_i}\sqrt{(1 - r_i e_{ii})^3}} \cdot \sqrt{d_i^2 + \sum_{j \neq i} e_{ij} c_j^2} \cdot p(c_i|\{c_j^2\}). \tag{7}$$

Now, we can apply the Bayes theorem and integrate with respect to $\{c_j^2\}$ $(j \neq i)$ to get the desired expression of the probability densities of the outputs:

$$p(r_i) = \int_0^{+\infty} \cdots \int_0^{+\infty} p(r_i|\{c_j^2\}) \cdot p(\{c_j^2\})\, \mathrm{d}c_1^2 \cdots \mathrm{d}c_{i-1}^2\, \mathrm{d}c_{i+1}^2 \cdots \mathrm{d}c_n^2. \tag{8}$$

Equation (8) is especially useful when there is an analytical expression for the conditional densities of the inputs.

2.2. *Approximate solution*

Several distributions have been proposed in the context of V1 models to model the conditional statistics of the coefficients obtained by projecting natural images onto an orthogonal linear basis (e.g., Schwartz and Simoncelli 2001, Wainwright *et al* 2001a). Here we will consider the Gaussian model of Schwartz and Simoncelli (2001). If we take two vertical slices of the conditional histogram in figure 1, the conditional distribution of coefficient c_2 looks like a Gaussian centred at zero and whose width increases with the amplitude of c_1. This statistical behaviour can be approximated by the following Gaussian probability density function (Schwartz and Simoncelli 2001):

$$p(c_i|\{c_j^2\}) = \frac{1}{\sqrt{2\pi\left(a_i^2 + \sum_{j \neq i} b_{ij} c_j^2\right)}} \exp\left\{-\frac{c_i^2}{2\left(a_i^2 + \sum_{j \neq i} b_{ij} c_j^2\right)}\right\}. \tag{9}$$

Note that $\left(a_i^2 + \sum_{j \neq i} b_{ij} c_j^2\right)$, where a_i^2 and $\{b_{ij}\}$ $(i \neq j)$ are free parameters, is the variance of the zero-mean Gaussian conditional density.

This probability density is consistent with the empirical observation that the standard deviation of c_i scales with the absolute value of the neighbouring coefficients $\{c_j\}$ $(j \neq i)$. The parameters a_i^2 and $\{b_{ij}\}$ $(i \neq j)$ can be determined, for example, by maximum-likelihood (ML) estimation (Schwartz and Simoncelli 2001). Operating, we obtain the following equation:

$$\{a_i^2, b_{ij}\} = \arg\min_{\{a_i^2, b_{ij}\}} \mathbb{E}\left\{\log\left(a_i^2 + \sum_{j \neq i} b_{ij} c_j^2\right) + \frac{c_i^2}{a_i^2 + \sum_{j \neq i} b_{ij} c_j^2}\right\} \tag{10}$$

where $\mathbb{E}$ denotes the expected value. In practice we can compute $\mathbb{E}$ for each subband, averaging over all spatial positions of a given set of natural images.

Network Downloaded from informahealthcare.com by Staats- und Universitaetsbibliothek Hamburg on 12/27/14
For personal use only.

RIGHTSLINK

The similarity between the exponent in equation (9) and equation (1) is completed by simply choosing $d_i^2 = a_i^2$, $e_{ij} = b_{ij}$ $(i \neq j)$ and $e_{ii} = 0$. In fact, this was the normalization proposed by Schwartz and Simoncelli (2001). In other words, they directly adopted the parameters of the Gaussian model, a_i^2 and b_{ij} $(i \neq j)$, as the normalization parameters.

In a previous work (Valerio and Navarro 2002), we demonstrated that the choice of parameters $d_i^2 = a_i^2$, $e_{ij} = b_{ij}$ $(i \neq j)$ and $e_{ii} = 0$ proposed by Schwartz and Simoncelli (2001) theoretically eliminates statistical dependences between normalized responses r_i and linear input coefficients $\{c_j\}$ $(j \neq i)$, but not between normalized responses. In effect, if we evaluate the minimum conditions for these parameters:

$$\frac{\partial \mathrm{MI}(r_1, r_2, \ldots, r_n)}{\partial d_i^2} \Rightarrow \mathbb{E}\left\{ r_i \frac{1}{c_i^2} [-\text{element}_{ii} \text{ of } (\mathbf{Id} - \mathbf{E} \cdot \mathbf{R})^{-1} + 1] \right\} \leqslant 0 \tag{11}$$

$$\frac{\partial \mathrm{MI}(r_1, r_2, \ldots, r_n)}{\partial e_{ij}} \Rightarrow \mathbb{E}\left\{ \begin{array}{l} r_i \dfrac{c_j^2}{c_i^2} [-\text{element}_{ii} \text{ of } (\mathbf{Id} - \mathbf{E} \cdot \mathbf{R})^{-1} + 1] \\ \quad + r_i \cdot \text{element}_{ji} \text{ of } (\mathbf{Id} - \mathbf{R} \cdot \mathbf{E})^{-1} \end{array} \right\} \geqslant 0 \tag{12}$$

$$\frac{\partial \mathrm{MI}(r_1, r_2, \ldots, r_n)}{\partial e_{ii}} \Rightarrow 0. \tag{13}$$

Expressions (11) and (12) are only approximately zero for the case of natural images. Note that with natural images $p(r_1, r_2, \ldots, r_n)$ tends to show a sharp peak at the origin. Therefore we can approximate the expected values by the values at the origin. This can explain the good numerical results reported so far, which show a drastic reduction of mutual information when using that choice of parameters. However, we have to bear in mind that this is a rough approximation, and this choice of parameters is not the solution of equation (6) in general. Observing the sign of the partial derivatives of the mutual information (equations (11) and (12)), one expects the optimal values to be $d_i^2 \gtrsim a_i^2$ and $e_{ij} \lesssim b_{ij}$ $(i \neq j)$. On the other hand, the considered choice of parameters, in particular $e_{ii} = 0$, that is eliminating its own linear coefficient c_i from the neighbourhood considered, is not fully consistent with experimental findings in physiological experiments and hence would limit its plausibility as a biological model. In fact, the mutual information of the normalized responses, $\mathrm{MI}(r_1, r_2, \ldots, r_n)$ (equation (2)), does not depend on the value of the normalization parameters e_{ii} (it can be demonstrated that $\frac{\partial \mathrm{MI}(r_1, r_2, \ldots, r_n)}{\partial e_{ii}} = 0$, no matter what the statistics of the linear inputs), so that these parameters can be fixed to any value.

2.3. Quasi-optimum condition

In the above subsection we have shown that the choice of parameters $d_i^2 = a_i^2$ and $e_{ij} = b_{ij}$ $(i \neq j)$, inspired in a Gaussian model, is an approximate solution of equation (6), easy to be computed as has been empirically demonstrated (Schwartz and Simoncelli 2001, Valerio and Navarro 2002). Here, we will take advantage of that fact. The basic idea is to simplify equation (6) assuming that the probability densities of the nonlinear responses provided by the exact solution can be approximated by those corresponding to the above approximate solution. This means that we approximate $p(r_i)$ by particularizing its definition in equation (8) to the case when $d_i^2 = a_i^2$ and $e_{ij} = b_{ij}$ $(i \neq j)$, that is to say:

$$p(r_i) = \frac{1}{\sqrt{2\pi r_i (1 - r_i e_{ii})^3}} \exp\left\{ -\frac{1}{2} \frac{r_i}{1 - r_i e_{ii}} \right\}. \tag{14}$$

This approximation gives an extraordinary fit to real data, as can be checked, for example, by calculating the Kullback–Leibler divergence between the actual histograms and the corresponding discretized distributions (see section 3).

Network Downloaded from informahealthcare.com by Staats- und Universitaetsbibliothek Hamburg on 12/27/14
For personal use only.

RIGHTSLINK

In that case, equation (6) converts to:

$$\{d_i^2, e_{ij}\} = \arg \max_{\{d_i^2, e_{ij}\}} \mathbb{E}\left\{\frac{1}{2}\sum_{k=1}^{n} \log \frac{r_k}{(1 - r_k e_{kk})^3} - \frac{1}{2}\sum_{k=1}^{n} \frac{r_k}{1 - r_k e_{kk}} + \log(|\det[\mathbf{Id} - \mathbf{R} \cdot \mathbf{E}]|)\right\}, \tag{15}$$

which is computationally much more efficient because it eliminates the necessity of estimating the marginal densities of the nonlinear responses.

It is noteworthy that, for natural images, the normalized responses r_i are typically close to zero (marginal densities tend to show a sharp peak at zero), which implies that $|\det[\mathbf{Id} - \mathbf{R} \cdot \mathbf{E}]|$ can be approximated by the product of the matrix elements in the main diagonal: $(1 - r_1 e_{11})(1 - r_2 e_{22}) \cdots (1 - r_n e_{nn})$. Then, after operating on equation (15) we arrive at

$$\begin{aligned}\{d_i^2, e_{ij}\} &= \arg \max_{\{d_i^2, e_{ij}\}} \mathbb{E}\left\{\frac{1}{2}\log\left(\frac{r_i}{1 - r_i e_{ii}}\right) - \frac{1}{2}\frac{r_i}{1 - r_i e_{ii}}\right\} \\ &= \arg \min_{\{d_i^2, e_{ij}\}} \mathbb{E}\left\{\log\left(d_i^2 + \sum_{j \neq i} e_{ij} c_j^2\right) + \frac{c_i^2}{d_i^2 + \sum_{j \neq i} e_{ij} c_j^2}\right\}. \end{aligned} \tag{16}$$

Equations (16) and (10) are equivalent, which means that under these approximations we arrive again at the approximate solution of section 2.2. This is another proof that equation (15) is a good approximation to equation (6), that is, the general solution to obtain a maximum statistical independence through divisive normalization of the linear coefficients.

2.4. Implementation

It is not computationally efficient to try to implement directly equation (15), mainly due to the log-determinant term, which involves a high computational cost. A rather more convenient solution can be obtained by applying a Taylor approximation to the log-determinant (Marcus and Minc 1992, Martin 1993), although there are other possible choices, such as the Chebyshev approximation (Abramowitz and Stegun 1972):

$$\log(|\det[\mathbf{Id} - \mathbf{R} \cdot \mathbf{E}]|) \approx -\sum_{k=1}^{q} \frac{\mathrm{tr}[(\mathbf{R} \cdot \mathbf{E})^k]}{k} \tag{17}$$

where tr[ ] represents the trace function.

One or two terms ($q = 1$ or 2) in equation (17) are usually enough to get a reasonable approximation of the log-determinant, because the elements of the matrix $\mathbf{R} \cdot \mathbf{E}$ are typically much less than 1 for natural images (note that, in the Taylor series approximation, the error introduced is of the same order of magnitude as the first neglected term). In this way, the numerical problem becomes much more tractable.

3. Results

The following results have been obtained using a 'training set' of six black and white natural images with 512 × 512 pixel format ('Boats', 'Elaine', 'Goldhill', 'Lena', 'Peppers' and 'Sailboat'). This is a reduced but representative set with a variety of images, and hence we believe that the results have sufficient generality. The computation of the optimal parameters of the divisive normalization can be considered as an adaptation process, so that one might expect to obtain even better results, that is lower mutual information, when the 'training set' is more homogeneous.

Network Downloaded from informahealthcare.com by Staats- und Universitaetsbibliothek Hamburg on 12/27/14
For personal use only.

RIGHTSLINK

The linear stage to obtain the input coefficients c_i consisted of applying a four-level orthogonal wavelet decomposition based on Daubechies filters of order 4 (db8), that is, with 4 vanishing moments and 8 coefficients (Daubechies 1992). This gives rise to 12 subbands (horizontal, vertical and diagonal for each of the 4 scales) plus an additional low-pass channel. The conditional statistics of the resulting coefficients were adjusted to the Gaussian model of equation (9). The model parameters were fitted to the histograms using the mathematical expectation in equation (10) over all coefficients of the 6 input images of the 'training set', but independently for each subband of the wavelet pyramid. Both in the Gaussian model and in the divisive normalization we have considered a 12-coefficient neighbourhood, $\{c_j\}$ $(j \neq i)$, of adjacent coefficients to the central one c_i along the four dimensions (8 in a square box in the 2D space, plus 2 adjacent neighbours in frequency and 2 in orientation). A linear search method was used to solve the corresponding minimization problems, using the additional constraint of positivity of the free model parameters, a_i^2 and $\{b_{ij}\}$ $(i \neq j)$, to improve convergence. As an example, figure 2 shows the values of the Gaussian model parameters for the lowest scale vertical subband. In general, $\{b_{ij}\}$ $(i \neq j)$ values are much less than 1 and, as expected, the highest values correspond to those neighbouring coefficients in space, orientation and scale that have the strongest statistical dependence with a certain coefficient.

Once we get this parametric fit of the conditional histograms, we can compute the approximately optimal, or quasi-optimal, parameters of the divisive normalization, applying equation (15), and using equation (17) to approximate the log-determinant by a second-order Taylor series. As we mentioned in the previous section, the mutual information of the normalized responses does not depend on the value of the normalization parameters e_{ii}, so that these parameters can be fixed to any value (zero for example, for the sake of simplicity). The linear search of these quasi-optimal parameters begins at the approximate solution of Schwartz and Simoncelli (2001) (section 2.2): $d_i^2 = a_i^2$, $e_{ij} = b_{ij}$ $(i \neq j)$ and $e_{ii} = 0$, where a_i^2 and b_{ij} $(i \neq j)$ are the parameters calculated previously of the Gaussian model for the conditional statistics of the linear coefficients c_i. As an example, figure 2 permits us to compare the values of the parameters of the Gaussian model and the divisive normalization for the lowest scale vertical subband. As one would expect (see section 2.2), the quasi-optimal values verify: $d_i^2 \gtrsim a_i^2$ and $e_{ij} \lesssim b_{ij}$ $(i \neq j)$.

When we apply the divisive normalization with the resulting quasi-optimal parameters, we obtain responses similar to those of the bottom panel of figure 3. Intuitively, this nonlinear transform has the effect of randomizing the image representation in order to reduce statistical dependences between coefficients belonging to the same structural feature, or, in other words, the effect of the divisive normalization is to choose which coefficients are most effective for describing a given image structure, similar to sparsification in those models based on an overcomplete and non-orthogonal basis set (see Olshausen and Field 1997). Figure 4 shows conditional histograms of two adjacent samples in space, so that we can appreciate clearly the progressive statistical independence achieved by successive application of the linear and nonlinear transforms. The upper panel corresponds to adjacent pixels p_1 and p_2 in the original image. The slope of nearly 1 indicates the strong correlation between coefficients, whereas the progressive spreading of the histogram suggests the presence of higher-order statistical dependences. The wavelet transform eliminates this correlation quite efficiently. We can see in the middle row panel, where now c_1 and c_2 are Daubechies wavelet coefficients, that now the slope is close to zero. However, this linear transform cannot remove higher-order statistical dependences, as suggested by the 'bowtie' shape of the histogram. The bottom row shows two conditional histograms after divisive normalization. The bottom left panel shows the statistical dependence between two output responses (r_1 and r_2) when we directly use the parameters of the Gaussian statistical model as the divisive normalization parameters, as in the Schwartz

Network Downloaded from informahealthcare.com by Staats- und Universitaetsbibliothek Hamburg on 12/27/14
For personal use only.

RIGHTSLINK

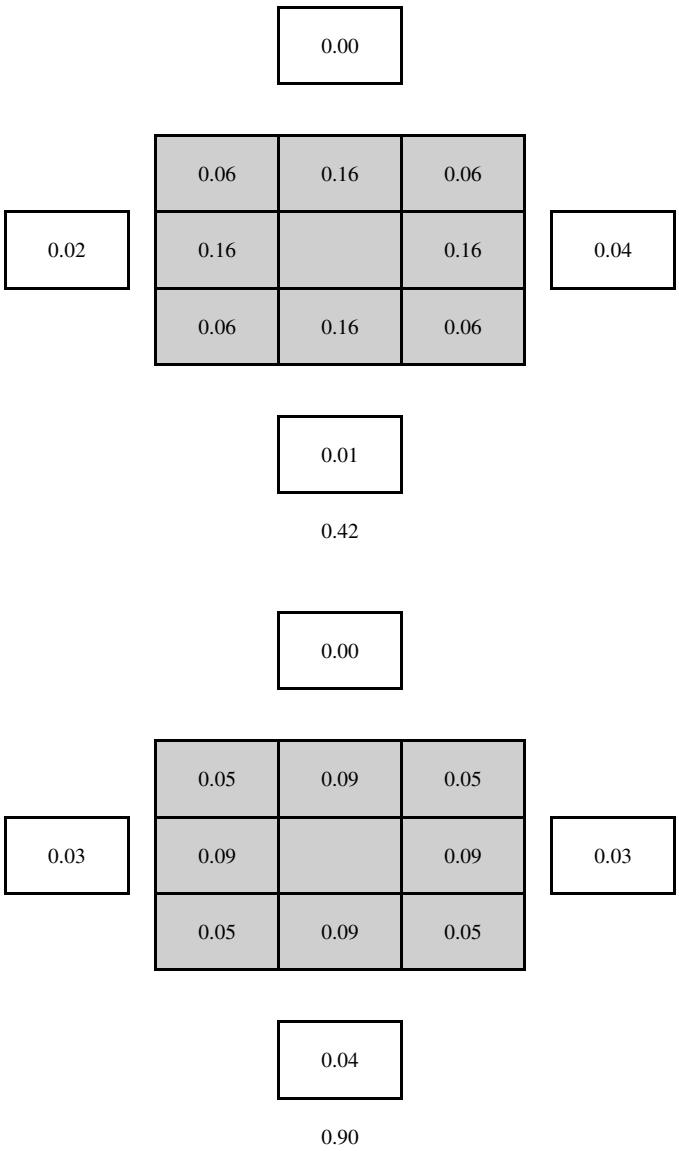

Figure 2. Parameter values of the Gaussian statistical model (upper panel) and the divisive normalization (bottom panel) for the lowest scale vertical subband. The shaded values correspond to the 8 spatial parameters. The two scale parameters are arranged vertically and the two orientation parameters horizontally. The bottom row contains the value of a_i^2 and d_i^2, respectively.

and Simoncelli (2001) approximate solution. The bottom right panel shows the statistical dependence between r_1 and r_2 when we use our approximation to the optimal solution with $e_{ii} = 0$. Both conditional histograms look quite similar, and in both cases, after normalization, pairwise dependences are removed in practice since the resulting conditional histograms are basically independent of the value of the abscissa. Nevertheless, we will show below that the mutual information is generally better when applying our optimal approximation.

The effect of divisive normalization on marginal statistics is illustrated in figure 5. As we can see, the marginal densities of the nonlinear responses, $p(r_i)$, are much more kurtotic

Network Downloaded from informahealthcare.com by Staats- und Universitaetsbibliothek Hamburg on 12/27/14
For personal use only.

RIGHTSLINK

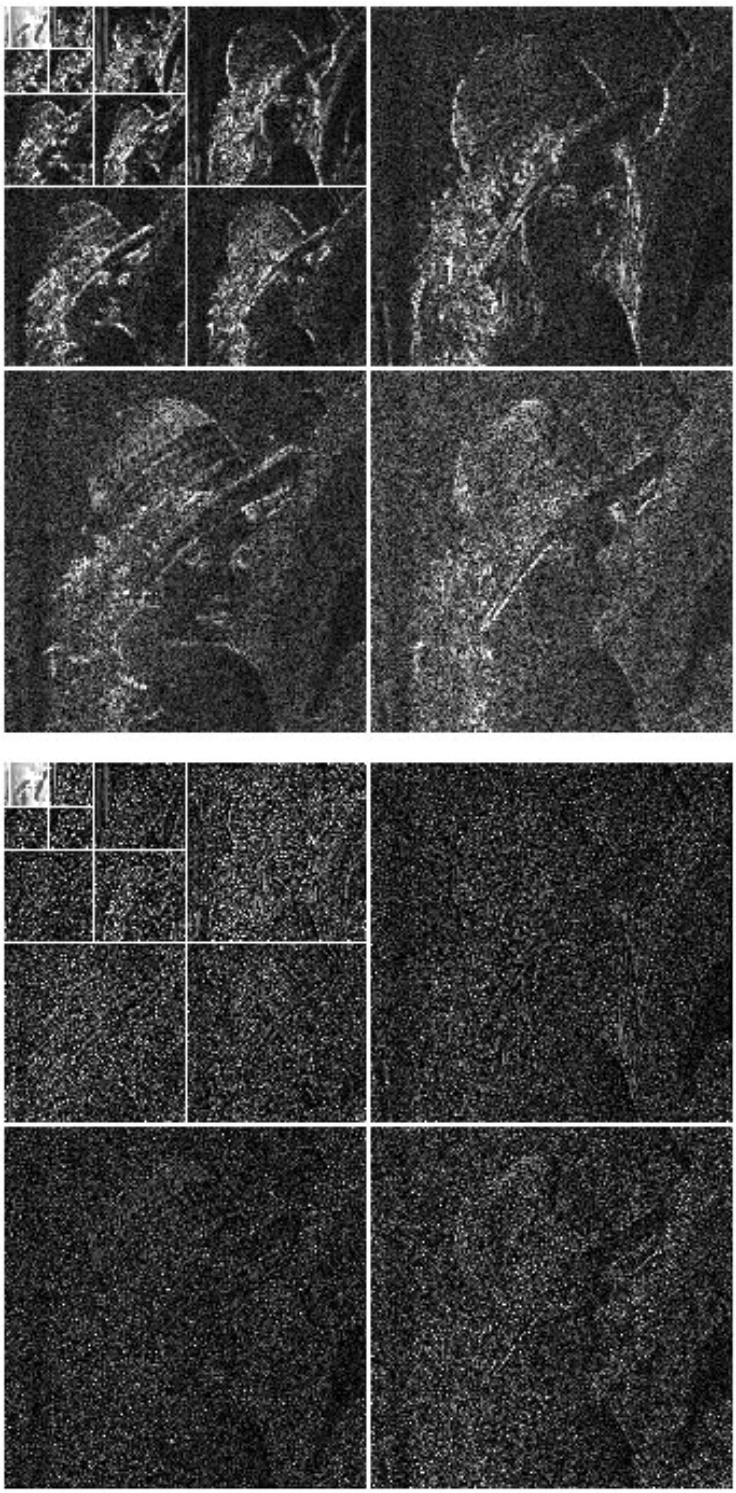

Figure 3. Daubechies and nonlinear divisive normalization decomposition of the 'Lena' image.

Network Downloaded from informahealthcare.com by Staats- und Universitaetsbibliothek Hamburg on 12/27/14
For personal use only.

RIGHTSLINK

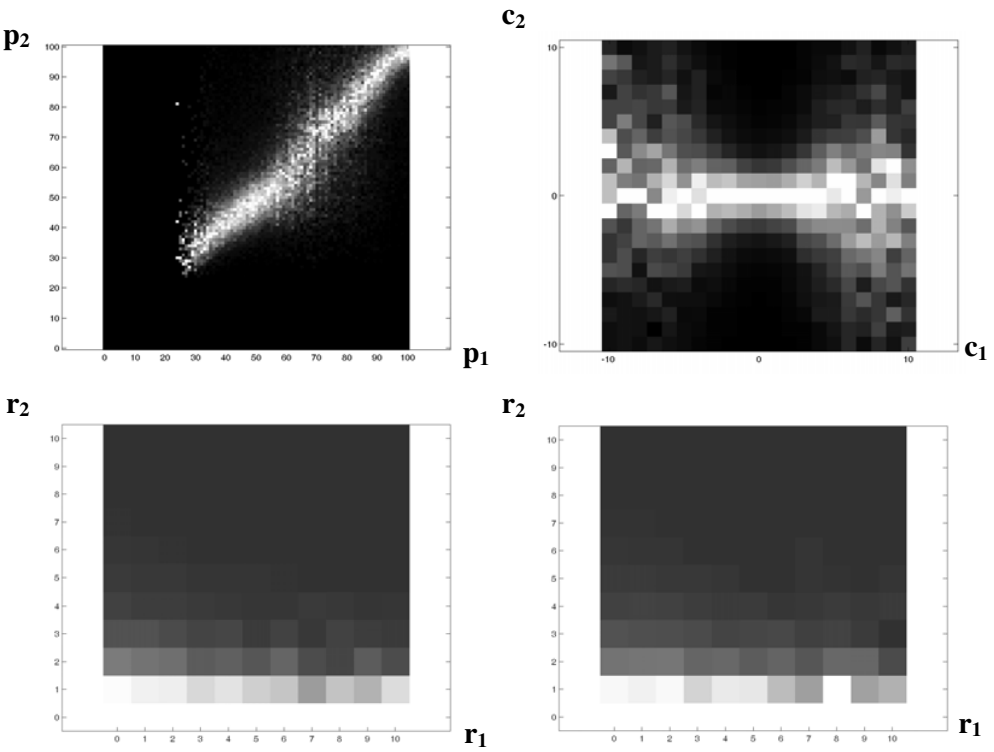

Figure 4. Conditional histograms of two neighbouring pixels (p_2, p_1) where p_1 is the right down neighbour of p_2, wavelet coefficients (c_2, c_1) and nonlinear responses (r_2, r_1) in the approximate (lower left panel) and the quasi-optimal case with $e_{ii} = 0$ (lower right panel) of the 'Lena' image. The considered subband is the lowest scale vertical one.

than the corresponding marginal densities of the linear inputs, $p(c_i)$. In addition, we want to remark that the resulting marginal densities of the nonlinear responses $p(r_i)$ closely fit the approximate expression (equation (14)) used in our method, which is an empirical proof of the validity of that approximation.

Table 1 shows some numerical measures of statistical dependence in terms of mutual information (Kullback–Leibler divergence) for images of the 'training set'. Consistent with figure 4, mutual information is high in the image domain (pixels p). Mutual information is much lower in the wavelet domain (linear coefficients c) after removing basically linear correlations. Finally, divisive normalization (normalized responses r) further decreases the mutual information, reaching much lower values close to zero. The two right columns of the table compare the mutual information obtained by the approximate solution and our solution with $e_{ii} = 0$. Our solution generally yields better results (lower mutual information) although it is obvious that the approximate solution also provides satisfactory results. In fact, the differences between the two cases are small.

In addition to the feature of giving almost statistically independent coefficients, the optimal divisive normalization has another important property, namely it is invertible (except for the signs that obviously need to be stored). Note that we can reconstruct the squared input linear coefficients c_i^2 from the normalized responses r_i and the parameters of the divisive

Network Downloaded from informahealthcare.com by Staats- und Universitaetsbibliothek Hamburg on 12/27/14
For personal use only.

RIGHTSLINK

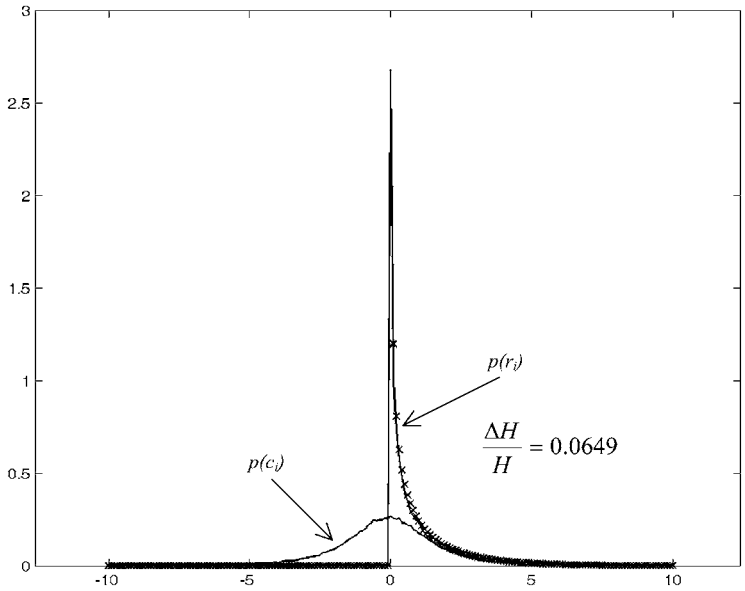

Figure 5. Marginal density functions of the wavelet coefficients (c_i) and the nonlinear responses (r_i) in the lowest scale vertical subband of the 'Lena' image. Crosses represent the considered approximation of the marginal density of the nonlinear responses: $p(r_i) \approx \frac{1}{\sqrt{2\pi r_i}} \exp\{-\frac{r_i}{2}\}$ and $\frac{\Delta H}{H}$ is the relative entropy ΔH (Kullback–Leibler divergence) between the histogram of the nonlinear responses and the approximation as a fraction of the histogram entropy H.

Table 1. Mutual information between adjacent pixels (p_2, p_1) where p_1 is the right down neighbour of p_2, between the resulting wavelet coefficients (c_2, c_1) and between the corresponding normalized responses (r_2, r_1) in the approximate case (A) and in the quasi-optimal case (QO) with $e_{ii} = 0$. The last column also shows the relative difference of the mutual information in the quasi-optimal case with respect to that in the approximate case. The considered subband is always the lowest scale vertical one. Mutual information has been calculated from 200 bin joint histograms in the interval $(-100,100)$ of the corresponding random variables after fixing their standard deviation to 5 in order to compare the results (note that to multiply one or two variables by a factor does not modify their mutual information).

	p_2, p_1	c_2, c_1	r_2, r_1 (A)	r_2, r_1 (QO)
'Boats'	1.1075	0.1736	0.0121	0.0090 (−26%)
'Elaine'	1.4450	0.0480	0.0106	0.0098 (−8%)
'Goldhill'	1.2340	0.0980	0.0119	0.0121 (+2%)
'Lena'	1.4364	0.1278	0.0113	0.0101 (−11%)
'Peppers'	1.5772	0.0861	0.0103	0.0097 (−6%)
'Sailboat'	1.2340	0.1164	0.0099	0.0100 (+1%)

normalization d_i^2 and $\{e_{ij}\}$ by simply operating in equation (1) and using matrix formulation:

$$\begin{pmatrix} c_1^2 \\ c_2^2 \\ \cdots \\ c_n^2 \end{pmatrix} = (\mathbf{Id} - \mathbf{R} \cdot \mathbf{E})^{-1} \begin{pmatrix} r_1 d_1^2 \\ r_2 d_2^2 \\ \cdots \\ r_n d_n^2 \end{pmatrix}. \tag{18}$$

Network Downloaded from informahealthcare.com by Staats- und Universitaetsbibliothek Hamburg on 12/27/14
For personal use only.

RIGHTSLINK

4. Summary and discussion

This work has focused on analysing the efficiency of statistically motivated nonlinear divisive normalization models of V1 neurons (Simoncelli and Schwartz 1999) in terms of statistical independence of the resulting responses, according to the efficient coding hypothesis. For this purpose we have formulated the problem in terms of the mutual information, as a direct measure of statistical independence. In this way, we arrived at the general expression for the global minimum of mutual information. The next step was to make use of the empirical knowledge of the statistical properties of natural images, namely that conditional histograms of neighbouring linear coefficients fit a Gaussian model with a quite reasonable accuracy (Schwartz and Simoncelli 2001). This is consistent with the hypothesis that divisive normalization in V1 neurons is adapted to the signals to which neurons are exposed. Under this theoretical framework, we have first analysed the divisive normalization proposed by Schwartz and Simoncelli (2001), which corresponds to directly using the parameters of the Gaussian conditional density as the parameters of divisive normalization: $d_i^2 = a_i^2$, $e_{ij} = b_{ij}$ $(i \neq j)$ and $e_{ii} = 0$. As we have shown here, there is no mathematical reason to fix $e_{ii} = 0$, which is, on the other hand, a somewhat controversial issue since with this choice the divisive normalization model loses biological plausibility. Moreover, the choice of parameters of Schwartz and Simoncelli (2001) does not correspond to the minimum of mutual information, i.e. is not optimal, but it is a first approximation, possibly not too far from that minimum. Nevertheless, this approximate solution is very convenient in terms of numerical implementation, so we decided to use it as the input guess in the optimization procedure to find the actual minimum of mutual information.

On the other hand, we have found an exact and an approximate expression of the minimum of the mutual information. This approximate expression is numerically tractable with a reasonable computational cost, yielding in general a better result than the initial approximation.

Numerical results constitute the empirical proof of both the theoretical formulation and the numerical procedures. Furthermore, the combination of theoretical and numerical results suggests that we have arrived very close to the actual minimum of the mutual information, that is, the optimal divisive normalization adapted to find the maximum statistical independence of the responses corresponding to the 'training set'. On the other hand, we could not demonstrate that the actual minimum is zero. Taking all these considerations and results together, we believe that this cannot be demonstrated simply because the minimum will be greater than zero in general. In other words, divisive normalization would strongly reduce statistical dependence, but could not eliminate it completely in general (perhaps residual very high-order dependences remain).

To really improve the results, one way could be to use a more accurate model to capture statistical properties of linear coefficients of natural images. Here, we have used a Gaussian model, but other models could provide a better fit. In addition to use of a different model, it is possible that by using a somewhat more sophisticated expression for the divisive normalization one could improve the results significantly. However, using more sophisticated models and independization mechanisms has the drawback of losing the elegance and relative simplicity associated with the current formulation of divisive normalization. Taking into account the high efficiency reached, there is room only for a little improvement, probably at the cost of a much higher complexity. In fact, even departing from this relative simple model and formulation, we had to make approximations to arrive at a numerically tractable optimization problem.

We want to finish by remarking that a complete image representation having the nice feature of statistical independence could find multiple applications in image analysis and processing (restoration, synthesis, fusion, coding and compression, registration, etc). The fact

Network Downloaded from informahealthcare.com by Staats- und Universitaetsbibliothek Hamburg on 12/27/14
For personal use only.

RIGHTSLINK

that the divisive normalization transform can be inverted (if we keep the signs of the linear coefficients) opens interesting possibilities in this sense. Similar schemes (Simoncelli 1997, Malo *et al* 2000) have already been used very successfully in image analysis and processing applications.

Acknowledgments

This research was supported by the Spanish Commission for Research and Technology (CICYT) under grant DPI2002-04370-C02-02. RV was supported by a Madrid Education Council and Social European Fund Scholarship for Training of Research Personnel, and by a City Hall of Madrid Scholarship for Researchers and Artists in the Residencia de Estudiantes.

References

Abramowitz M and Stegun I 1972 *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (Washington DC: US Government Printing Office)

Atick J J 1992 Could information theory provide an ecological theory of sensory processing? *Network: Comput. Neural Syst.* **3** 213–51

Attneave F 1954 Some informational aspects of visual perception *Psych. Rev.* **61** 183–93

Barlow H B 1961 Possible principles underlying the transformation of sensory messages *Sensory Communication* vol 217, ed W A Rosenblith (Cambridge, MA: MIT Press) pp 217–34

Bell A J and Sejnowski T J 1997 The 'independent components' of natural scenes are edge filters *Vis. Res.* **37** 3327–38

Bonds A B 1989 Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex *Vis. Neurosci.* **2** 41–55

Carandini M, Heeger D J and Movshon J A 1997 Linearity and normalization in simple cells of the macaque primary visual cortex *J. Neurosci.* **17** 8621–44

Daubechies I 1992 Ten lectures on wavelets *CBMS-NSF Lecture Notes* no 61 (Philadelphia, PA: SIAM)

Field D J 1994 What is the goal of sensory coding? *Neural Comput.* **6** 559–601

Geisler W S and Albrecht D G 1992 Cortical neurons: isolation of contrast gain control *Vis. Res.* **8** 1409–10

Heeger D J 1992 Normalization of cell responses in cat striate cortex *Vis. Neurosci.* **9** 181–98

Kullback S 1959 *Information Theory and Statistics* (New York: Wiley)

Kullback S and Leibler R A 1951 On information and sufficiency *Ann. Math. Stat.* **22** 79–86

Laughlin S B 1981 A simple coding procedure enhances a neuron's information capacity *Z. Naturf.* C **36** 910–12

Lewicki M S and Olshausen B A 1999 Probabilistic framework for the adaptation and comparison of image codes *J. Opt. Soc. Am.* A **16** 1587–601

Malo J, Ferri F, Navarro R and Valerio R 2000 Perceptually and statistically decorrelated features for image representation: application to transform coding *Proc. 15th Int. Conf. on Pattern Recognition* vol 3 (Los Alamitos, CA: IEEE Computer Society Press) pp 242–5

Marcus M and Minc H 1992 *A Survey of Matrix Theory and Matrix Inequalities* (New York: Dover)

Martin R J 1993 Approximations to the determinant term in Gaussian maximum likelihood estimation of some spatial models *Commun. Stat.-Theory Methods* **22** 189–205

Olshausen B A 2002 Sparse codes and spikes *Statistical Theories of the Brain* ed R Rao, B Olshausen and M Lewicki (Cambridge, MA: MIT Press) pp 257–72

Olshausen B A and Field D J 1996 Emergence of simple-cell receptive field properties by learning a sparse code for natural images *Nature* **381** 607–9

Olshausen B A and Field D J 1997 Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vis. Res.* **37** 3311–25

Papoulis A 1991 *Probability, Random Variables and Stochastic Processes* 3rd edn (Singapore: McGraw-Hill)

Rieke F, Bodnar D A and Bialek W 1995 Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents *Proc. R. Soc.* B **262** 259–65

Schwartz O and Simoncelli E P 2001 Natural signal statistics and sensory gain control *Nat. Neurosci.* **4** 819–25

Simoncelli E P 1997 Statistical models for images: compression, restoration and synthesis *Asilomar Conf. Signals, Systems, Comput.* (Los Alamitos, CA: IEEE Computer Society Press) pp 673–9

Simoncelli E P 1999 Modeling the joint statistics of images in the wavelet domain *Wavelet Applications in Signal and Image Processing VII* ed M A Unser, A Aldroubi and A F Laine *Proc. SPIE* **3813** 188–95

Network Downloaded from informahealthcare.com by Staats- und Universitaetsbibliothek Hamburg on 12/27/14
For personal use only.

RIGHTSLINK

Simoncelli E P and Olshausen B A 2001 Natural image statistics and neural representation *Ann. Rev. Neurosci.* **24** 1193–216

Simoncelli E P and Schwartz O 1999 Modeling surround suppression in V1 neurons with a statistically-derived normalization model *Adv. Neural Inform. Process. Syst.* **11** 153–9

Teo P C and Heeger D J 1994 Perceptual image distortion *Human Vision, Visual Processing, and Digital Display V* ed B E Rogowitz and J P Allebach *Proc. SPIE* **2179** 127–39

Valerio R and Navarro R 2002 Input–output statistical independence in divisive normalization models of V1 neurons *Network: Comput. Neural Syst.* submitted

Van Hateren J H 1992 A theory of maximizing sensory information *Biol. Cybern.* **68** 23–9

Van Hateren J H and van der Schaaf A 1998 Independent component filters of natural images compared with simple cells in primary visual cortex *Proc. R. Soc.* B **265** 359–66

Wainwright M J, Schwartz O and Simoncelli E P 2001a Natural image statistics and divisive normalization: modeling nonlinearities and adaptation in cortical neurons *Statistical Theories of the Brain* ed R Rao, B Olshausen and M Lewicki (Cambridge, MA: MIT Press) pp 203–22

Wainwright M J and Simoncelli E P 2000 Scale mixtures of Gaussians and the statistics of natural images *Adv. Neural Inform. Process. Syst.* **12** 855–61

Wainwright M J, Simoncelli E P and Willsky A S 2001b Random cascades on wavelet trees and their use in modeling and analyzing natural imagery *Appl. Comput. Harmon. Anal.* **11** 89–123

Watson A B and Solomon J A 1997 Model of visual contrast gain control and pattern masking *J. Opt. Soc. Am.* A **14** 2379–91

Wegman B and Zetzsche C 1990 Statistical dependence between orientation filter outputs used in an human vision based image code *Visual Communications and Image Processing '90,* ed M Kunt, *Proc. SPIE* **1360** 909–22

Network Downloaded from informahealthcare.com by Staats- und Universitaetsbibliothek Hamburg on 12/27/14
For personal use only.

RIGHTSLINK