A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images

Aapo Hyvärinen[1,2] and Patrik O. Hoyer[1]

1) Neural Networks Research Centre, Helsinki University of Technology
P.O. Box 5400, FIN-02015 HUT, Finland

2) Department of Psychology, General Psychology Division, University of Helsinki
P.O Box 13, FIN-00014 Univ. of Helsinki, Finland

submitted to Vision Research

January 18, 2001

Abstract

The classical receptive fields of simple cells in the visual cortex have been shown to emerge from the statistical properties of natural images by forcing the cell responses to be maximally sparse, i.e. significantly activated only rarely. Here, we show that this single principle of sparseness can also lead to emergence of topography (columnar organization) and complex cell properties as well. These are obtained by maximizing the sparsenesses of locally pooled energies, which correspond to complex cell outputs. Thus we obtain a highly parsimonious model of how these properties of the visual cortex are adapted to the characteristics of the natural input.

Keywords: natural images, neural networks, spatial vision, cortex, independent component analysis.

1 Introduction

The spatial classical receptive fields (CRF's) of neurons in the primary visual cortex (V1) of primates are selective for location, orientation and frequency (Hubel and Wiesel, 1968; DeValois et al., 1982). The neurons are topographically organized by these same parameters (Hubel and Wiesel, 1977; Blasdel, 1992; Tootell et al., 1988), which means that the preferred values for these parameters tend to change smoothly when moving tangentially to the cortical surface. Further, most cells can be divided into the categories of simple (essentially linear) vs. complex (phase-insensitive) cells (Hubel and Wiesel, 1968; Pollen and Ronner, 1983).

A fundamental problem in vision research is to determine why the selectivities and the organization of the cells are as described above. Recently, several authors have considered possible connections between the properties of the early visual system and the statistical properties of natural images. It is reasonable to assume that the visual system is adapted to process the particular kind of input it receives. Such an adaptation would be produced by the combined forces of evolution and neural development.

It is clear that the visual input has certain statistical characteristics that distinguishes it from any arbitrary input. First, the visual input is not white noise: The Fourier amplitudes fall off approximately proportional to the inverse of the frequency (see e.g. (Ruderman and Bialek, 1994)). Second, visual input is not Gaussian: The outputs of linear filters are usually strongly non-Gaussian, which is especially true of Gabor filters mimicking simple cells (Field, 1994). Third, on a higher level of description, visual input contains structure such as edges, bars, and different textures.

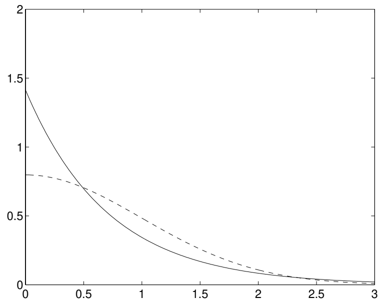

Figure 1: Illustration of a sparse probability density. Vertical axis: probability density. Horizontal axis, (absolute) value of random variable s_i. The exponential density function $p(s_i) = \sqrt{2}\exp(-\sqrt{2}s_i)$ is given by the solid curve; this models the distribution of the absolute values of a sparse variable. For comparison, the density of the absolute value of a Gaussian random variable is given by the dash-dotted curve. Both are normalized so that $E\{s_i^2\} = 1$, as explained in the text.

Thus, it could been argued that the properties of V1 reflect the statistical properties of the input (Barlow, 1961; Barlow, 1972; Field, 1994). No statistical signal processing system can be optimal for analyzing *any* kind of input; for an input set with given statistical properties, one can find a system that is optimal in a given sense. For example, the CRF's of simple cells are not very useful for analyzing data that consists of Gaussian noise, since such noise could be better analyzed by Fourier methods (Field, 1994). The reason why the CRF's have Gabor-like shapes might thus be that this kind of CRF's are optimal for analyzing the input that the visual system typically receives.

On a very low level, one can model the statistical structure of natural images by a linear model. The basic model that we consider here expresses the static monochrome image $I(x, y)$ as a linear superposition of some features or basis vectors $a_i(x, y)$:

$$I(x, y) = \sum_{i=1}^{n} a_i(x, y)s_i. \tag{1}$$

The s_i are stochastic coefficients, different for each image $I(x, y)$. In a cortical interpretation, the s_i model the responses of (signed) simple cells, and the a_i are closely related to their CRF's (Olshausen and Field, 1996; Olshausen and Field, 1997). Note that we are here considering the contrast only, i.e. the local mean or DC component has been removed from the image, so we can assume that the s_i have zero mean.

A fundamental assumption in this model is that the s_i are nongaussian in a particular way, called sparseness, or alternatively supergaussianity or leptokurtosis (Barlow, 1972; Field, 1994). Sparseness means that the random variable takes very small (absolute) values or very large values more often than a gaussian random variable would; to compensate, it takes values in between relatively more rarely. Thus the random variable is activated, i.e. significantly non-zero, only rarely. The probability density of the absolute value of a sparse random variable is often modelled as an exponential density, which has a higher peak at zero than a Gaussian density, see Fig. 1.

Sparseness is not dependent on the variance (scale) of the random variable. To measure the sparseness of the random variable s_i, let us first normalize its scale so that the expectation $E\{s_i^2\}$ equals some given constant. Then the sparseness can be measured as the expectation $E\{G(s_i^2)\}$ of a suitable nonlinear function of the square. Typically, G is chosen to be convex, i.e. its second derivative is positive, e.g. $G(s_i^2) = (s_i^2)^2$. Convexity implies that this expectation is large when s_i^2 typically takes values that are either very close to 0 or very large, i.e. when s_i is sparse.

Further, the s_i in the basic model are assumed to be statistically independent, i.e. the value of s_j cannot be used to predict s_i for $i \neq j$. The resulting model is called sparse coding (Olshausen and Field, 1996), or independent component analysis (ICA) (Jutten and Herault, 1991; Comon, 1994; Hyvärinen and Oja, 2000), and it can be considered a nongaussian version of factor analysis.

Estimation of the ICA model consists of determining the values of both s_i and $a_i(x, y)$ for all i and (x, y), given a sufficient number of observations of images, or in practice, image patches $I(x, y)$. Estimation can be performed by maximizing the likelihood (Pham et al., 1992). We restrict ourselves here to the basic case where the $a_i(x, y)$ form an invertible linear system, in which case we can compute the s_i as dot-products

$$s_i = \langle w_i, I \rangle = \sum_{x,y} w_i(x, y) I(x, y). \tag{2}$$

where the w_i denote the inverse filters which are closely related, if not necessarily identical, to the CRF's (Olshausen and Field, 1996; Olshausen and Field, 1997). To further simplify the estimation, one often preprocesses the input data $I(x, y)$ by whitening, that is, by removing second-order correlations. Then the w_i can be constrained to be orthogonal and to have unit norm, that is, $\langle w_i, w_j \rangle$ equals 0 for $i \neq j$ and 1 for $i = j$ (Comon, 1994; Hyvärinen and Oja, 2000). This procedure guarantees that the basis is complete and the s_i are uncorrelated with $E\{s_i^2\} = 1$, and it also somewhat simplifies the likelihood. Then the likelihood L of observed image patches $I_t, t = 1, ..., T$ can be formulated as:

$$\log L(I_1, \ldots, I_T;\ w_1, \ldots, w_n) = \sum_{t=1}^{T} \sum_{i=1}^{n} G(\langle w_i, I_t \rangle^2) \tag{3}$$

where $G(s_i^2) = \log p_i(s_i)$ with p_i being the probability density of s_i, here assumed to be identical for all i. Due to the sparsity of the s_i, the function G is typically convex; for example, it is essentially the negative square root for the exponential density in Fig. 1. Thus, maximization of likelihood can be seen as maximization of sparsity (Olshausen and Field, 1996). When the model is estimated with input data consisting of patches of natural scenes, the obtained basis vectors $a_i(x, y)$ have the principal properties of simple cell CRF's (Olshausen and Field, 1996; Bell and Sejnowski, 1997; van Hateren and van der Schaaf, 1998).

In this paper, we extend the sparse coding principle to model complex cell properties and topography. By topography, we mean the columnar or clustering organization of the cells. Several neural network models have been proposed for learning these properties (von der Malsburg, 1973; Kohonen, 1982; Linsker, 1988; Obermayer et al., 1990; Erwin et al., 1995; Miller, 1995; Kohonen, 1996; Swindale, 1996; Hyvärinen and Hoyer, 2000), but none has succesfully demonstrated emergence of all of them. Here we show that these properties emerge from a two-layer sparse coding model that is fed natural image data as its input.

2 The model

The extension of the basic ICA or sparse coding model so that it models complex cell properties and topography is possible simply by considering the sparsenesses of *local activations* instead of simple cell responses. The starting point is that the components given by ICA are actually not independent, and the remaining dependencies can be further analyzed.

Each simple cell is modelled as a linear filter with adaptable weights, given equivalently by the basis vectors a_i or the filters w_i. The output of the simple cell with index i, when input with an image patch I_t, is thus given by the dot-product or convolution as in equation (2). This is like in basic ICA.

Here we introduce a *complex-cell layer* to analyze the dependencies not considered by basic ICA or linear sparse coding. Simple cell outputs are rectified by taking squares (energies), and these are fed to the complex cells. To model the organization of the cells, it is further assumed that the *simple cells are arranged in a two-dimensional grid* or lattice as is typical in topographic models. The topography is formally expressed by a neighbourhood function $h(i, j)$ that gives the proximity of the cells with indices i and j (note that these indices are two-dimensional). Typically, one defines that $h(i, j)$ is 1 if the cells are sufficiently close to each other, and 0 otherwise. There is a one-to one correspondence between the simple cells and complex cells: for each value of the index i, there is one simple cell and one complex cell, so the complex cells have an organization as well, but their organization is not essential here.

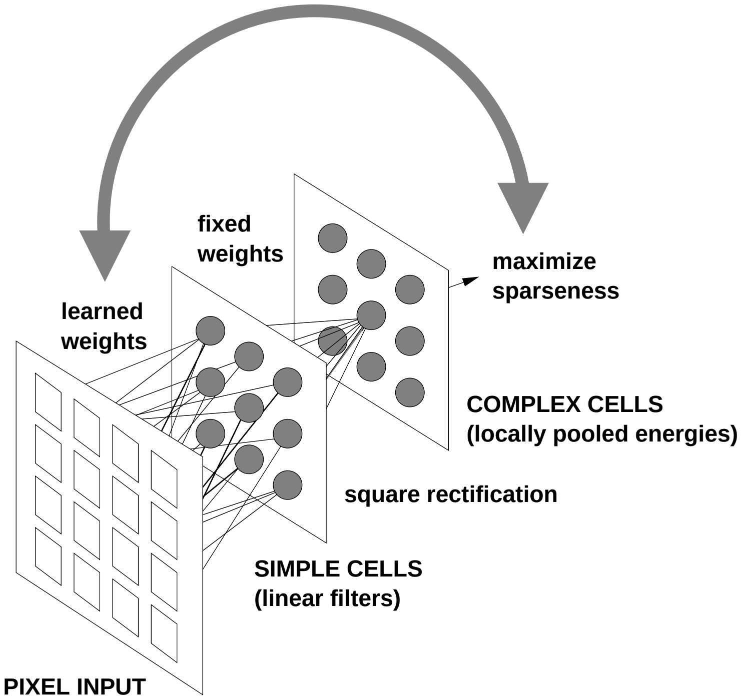

Figure 2: Illustration of the two-layer network. The first layer consists of simple cells or linear filters. The second layer is a complex cell layer where the energies of simple cell outputs are pooled locally. The filter weights in the first layer are estimated from natural image data. The estimation is performed by maximizing the sparsenesses of the outputs of complex cells (or local activations). A specific spatial organization of the cells emerges in the process. Note that while the complex cell pooling weights are fixed, the sparseness of complex cells and the topographic organization are directly attained by learning just the simple cell weights.

To fix the pooling weights from simple cells to complex cells, we make here the assumption that complex cells only pool outputs of simple cells that are near-by on the topographic grid. Thus, the complex cell outputs are given by the local activations. The local activation c_{it}, at a position i on the grid for stimulus I_t, means a weighted sum of the energies of near-by simple cells (see below). Such a spatial arrangement would be useful to minimize wiring length (Durbin and Mitchison, 1990) and neuroanatomic measurements indicate that the wiring of complex cells may indeed be so constrained (see discussions in (Blasdel, 1992; DeAngelis et al., 1999)).

The pooling process into local activations or complex cell responses can be expressed using the above-defined neighbourhood function $h(i, j)$. This function directly gives the pooling weights or the connections between the simple cell with index i and the complex cell with index j. Thus, we define the local activation as

$$c_{it} = \sum_{j=1}^{n} h(i, j) \langle w_j, I_t \rangle^2. \tag{4}$$

This two-layer network where the outputs of the simple cells are square-rectified and locally pooled in the complex cell layer is illustrated in Fig. 2.

After fixing the network structure, we define its learning process as the estimation of a generative model that is an extension of ICA. The connections between simple cells and complex cells, given by $h(i, j)$, are considered fixed and are *not* learned from the natural image input. What is learned are the basis vectors a_i, or equivalently, the inverse filters w_i. The learning of the basis vectors a_i is modulated by the complex cell pooling process, which is why the results are not the same with basic ICA.

Thus we define a topographic extension of the ICA model in which the likelihood (i.e. the probability of the data given the model parameters) is given by

$$\log L(I_1, \ldots, I_T;\ w_1, \ldots, w_n) = \sum_{t=1}^{T} \sum_{i=1}^{n} G(c_{it}). \tag{5}$$

The topography given by $h(i,j)$ is considered fixed, and only the first-layer weights w_j are estimated, so this likelihood is a function of the w_i only. The function G is again typically convex, and the expectation of its argument is constant due to whitening and orthogonalization.

The central feature of this model is that the responses s_i of near-by simple cells are statistically *not* independent in this model. The responses are still linearly uncorrelated,[1] but they have nonlinear dependencies. In fact, the rectified outputs s_i^2 are strongly positively correlated for neighbouring cells. This means *simultaneous activation* of neighbouring cells; such simultaneous activation is implicit in much of the work in cortical topography.

To illustrate the connection between sparseness of complex cell responses and simultaneous activation, let us consider the following two cases. Consider just two simple cells which have the same distributions for the output energies, and whose output energies are pooled to a complex cell response. If the outputs are statistically independent, the pooling to local activations reduces sparseness. This is because of the fundamental result given by the Central Limit Theorem in probability theory, and which forms the basis of much of the theory if ICA (Hyvärinen and Oja, 2000). Roughly, this result says that the sum of independent random variables is closer to Gaussian (and therefore, less sparse) than the original random variables themselves. Now, consider the contrasting extreme case where the cell outputs are perfectly dependent, that is, equal. This means that the distribution of the pooled energies is equal to the distribution of the original energies (up to a scaling constant), and therefore there is no reduction in sparseness. Thus, the sparseness of complex cell responses requires not only that the simple cell responses are sparse, but also that each complex cell pools responses that are not too independent.

Our model can be estimated by maximization of the likelihood. Maximization of the likelihood is here equivalent to maximizing the sparsenesses of the local activations or complex cell responses. This is because the likelihood now measures the sparsenesses of the complex cell responses (or, strictly speaking, the square roots of the responses), just like the likelihood in ICA measured the sparsenesses of the simple cell responses. Sparseness of local activations means that at any given time, simple cells that have significantly non-zero responses tend to be spatially clustered. Such sparse topographic activations are illustrated in Fig. 3. Sparse topographic activation is of course intimately related to simultaneous activation of neighbours.

To get some further insight to the learning process that ensues from the maximization of the likelihood of our model, one might consider an alternative two-step procedure. First, one might learn the basis vectors a_i by ordinary linear sparse coding, completely ignoring the second layer and the topography. After this, the cells (or basis vectors) could then be arranged on the 2-D topographic grid so that cells that are simultaneously active are as close to each other as possible. Essentially, estimation of our model combines these two hypothetical steps into a single process. There is no need to explicitly shift the cells afterwards (as in the two-step procedure) because the objective of sparseness of local activations automatically considers the spatial arrangement of the cells. Thus, the learning of the a_i is modulated by feedback so that the organization is automatically created without any need to learn the second-layer weights.

3 Data and methods

Using patches of natural images as the input data, we estimated the representation given by the model. For simplicity, we consider here only images that are static, monocular, and monochromatic.

[1]This is because in the model the probability only depends on the absolute value of s_i. This means that the linear correlation must be zero, because this symmetry implies that $E\{s_i s_j\} = E\{(-s_i)s_j\}$, and therefore necessarily $E\{(-s_i)s_j\} = 0$.

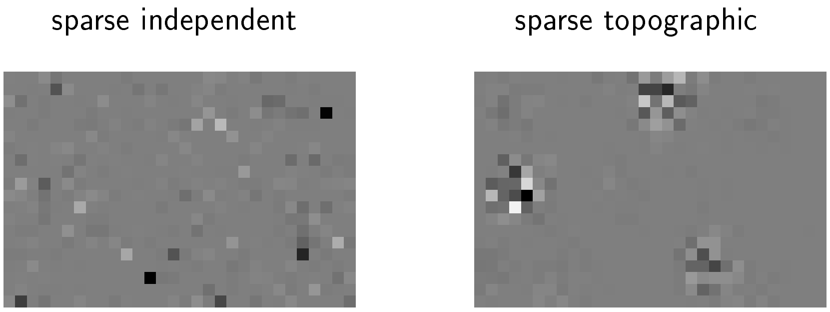

Figure 3: Illustration of sparse topographic (local) activations. Each pixel in the two plots gives the (signed) activity of one simple cell, grey means no activation and black and white are activated. Left: The original ICA or sparse coding model. The activations are sparse, but have no spatial structure. Right: Our model. The activations are sparse, and are in addition spatially clustered, or topographic.

A sample of 50,000 image patches of 32×32 pixels were sampled from natural images available on the WWW (`http://www.cis.hut.fi/projects/ica/data/images`). The choice of this image set was somewhat arbitrary; see Discussion for a critique. To reduce noise and aliasing artifacts, the dimension of the data was reduced to 625 components by principal component analysis: We rejected the components with low variances, as well as the 4 components with the largest variances. The total effect was a bandpass filtering. At the same time, the data was whitened. The basis vectors $a_i(x, y)$ are shown below in the original space, i.e. after inverting these preprocessing steps.

The topography was chosen as a 25×25 torus, i.e. a square whose opposite edges are considered connected together, to avoid border effects. The neighbourhood function was chosen so that for a given i, $h(i, j)$ is 1 if the cell j is in a 5×5 square centered on cell i; otherwise $h(i, j)$ is zero (see dashed square in Fig. 4). Thus, the pooling area for complex cells consisted of 25 simple cells. This neighbourhood function was chosen after some experimenting with different neighbourhood sizes. Smaller neighbourhood sizes lead to a weaker topographic organization, and very large ones did not give good (Gabor-like) simple cell CRF's.

We used three different convex functions $G(c) = -\alpha_1\sqrt{\epsilon + c} + \beta_1$, $G(c) = -\alpha_2 \tanh(c/2) + \beta_2$ and $G(c) = -\alpha_3 \log(1 + c) + \beta_3$ where $\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3 > 0$ are irrelevant normalization constants only needed for the likelihood interpretation, and $\epsilon = .00001$. These nonlinearities yielded essentially identical results; the results are shown here for the first one.

The likelihood in Eq. (5) was maximized by an ordinary gradient method; the iteration was started from random initial values for the w_i. This took about one week on a single RISC processor.

4 Results

The basis functions $a_i(x, y)$ are shown in Fig. 4, with their topographic ordering. The topographic map has basis vectors that are tuned for the three principal parameters: orientation, frequency and location. Visual inspection of the map shows that orientation and location mostly change smoothly as a function of position on the topographic grid. A striking feature of the map is a "blob" grouping low-frequency basis vectors. Thus the topography is determined by the same set of parameters for which the model cells are selectively tuned.

For further analysis, we fitted Gabor functions to the estimated basis vectors using the criterion of least squares. Thus we described each simple cell in terms of phase, location, orientation, and frequency. First we investigated local parameter correlations (DeAngelis et al., 1999). In Fig. 5 (a), we show the scatterplots for the values of a single parameter for every pair of two adjacent cells. Location, orientation,

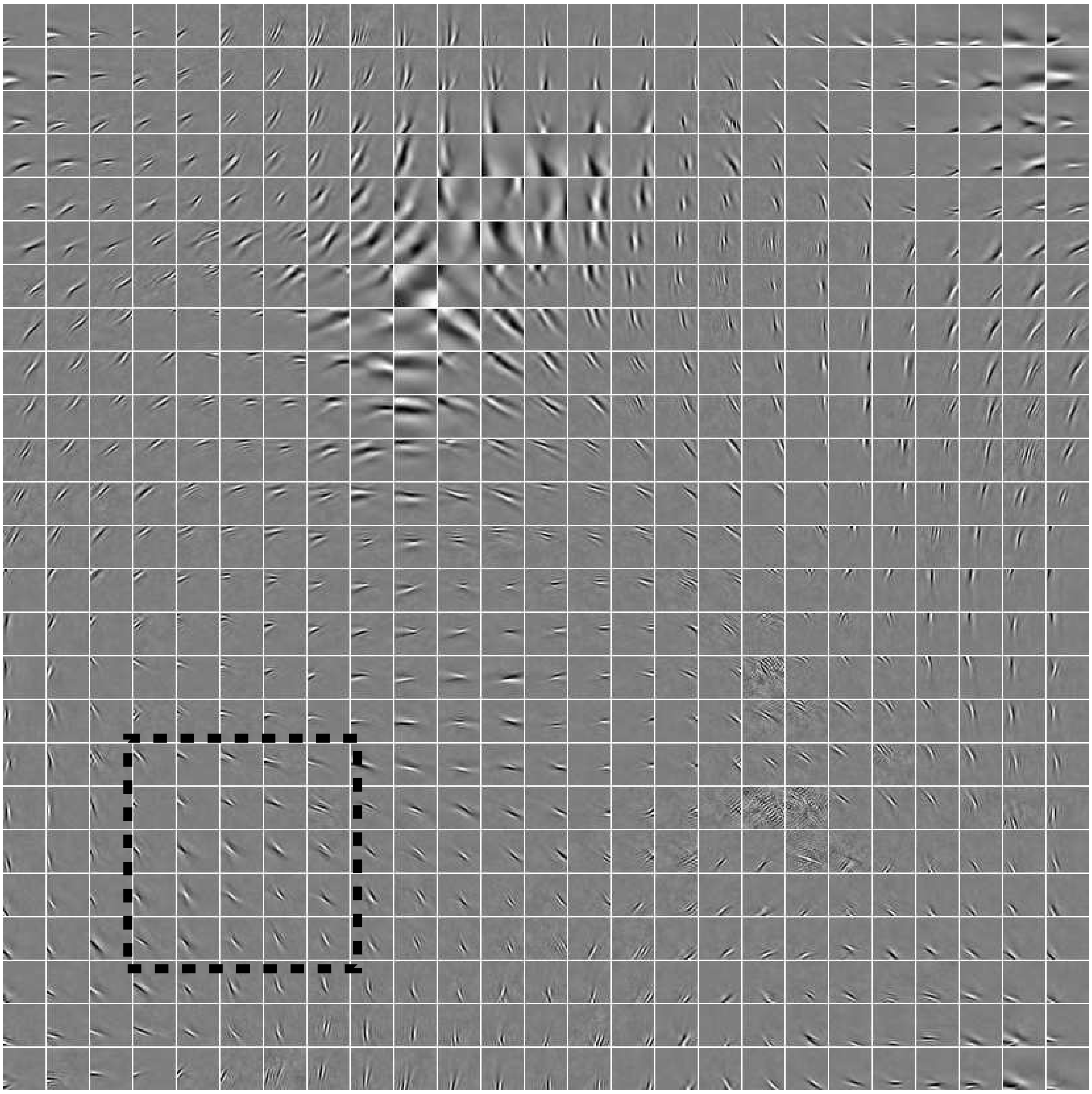

Figure 4: The topographic basis vectors a_i estimated from natural image data. The basis vectors are similar to those obtained by basic ICA or sparse coding (Olshausen and Field, 1996), but in our model they have a particular order. The extent of the neighbourhood (pooling area) is shown by the dashed square.

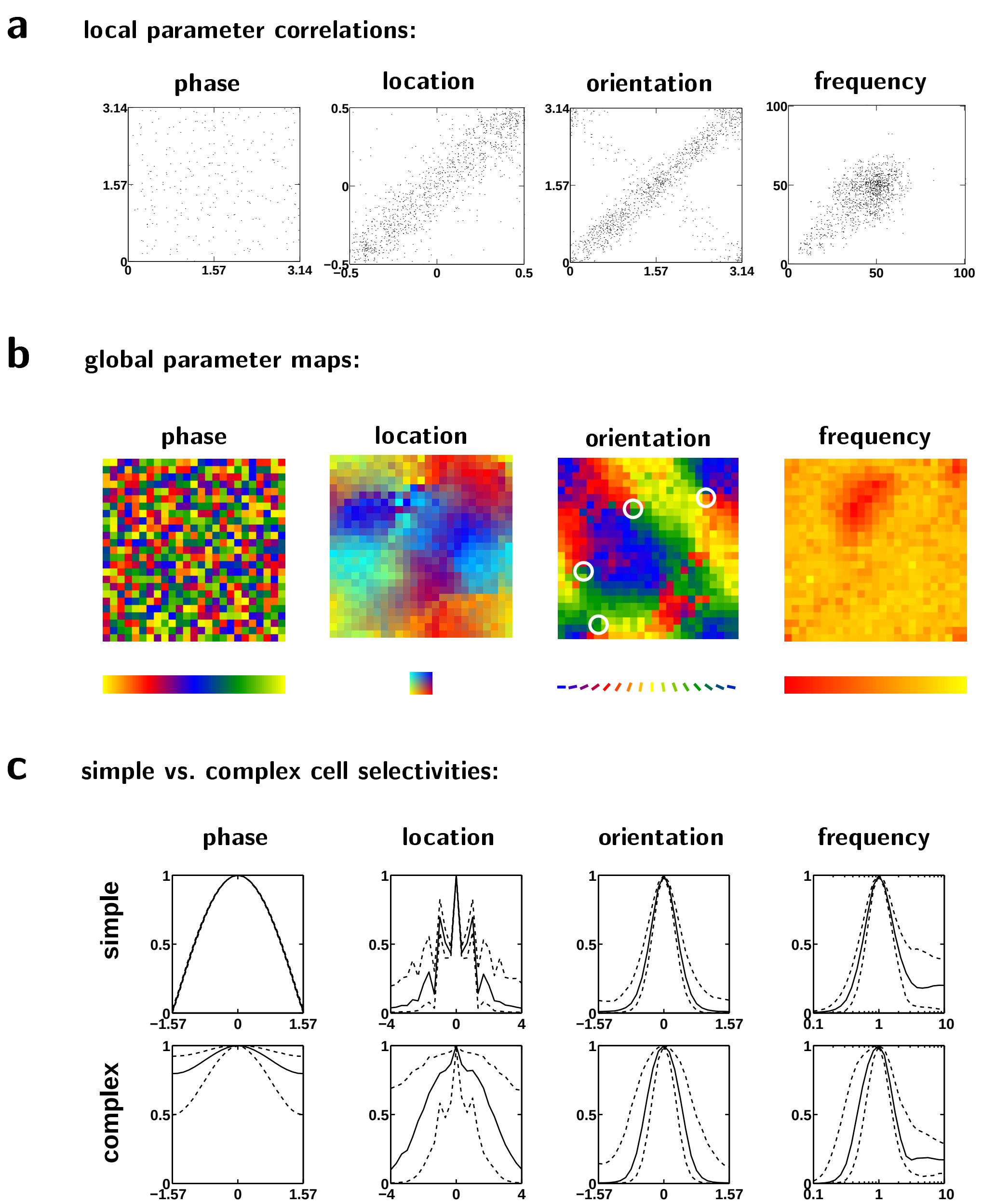

Figure 5: Analysis of topographic organization by fitted Gabor functions. In (a), scatterplots showing the local parameter correlations. Each pair of two adjacent cells is one dot on each plot, and the axes give the values of the parameter for the fitted Gabors. Location is along the x axis; the y axis gives virtually identical results. In (b), the global map of each parameter. Each cell corresponds to one pixel, whose color codes the value of the parameter of the fitted Gabor. In (c), analysis of complex cell properties of neighbourhoods. One parameter of the fitted Gabor was changed at a time, and the response (normalized so that the maximum is one) was plotted as a function of the parameter change. (Location was changed perpendicular to the preferred orientation.) Upper row: simple cells (absolute values), lower row: complex cells. Solid line: median of responses of all the 625 cells in the population, dotted lines: 90% and 10% quantiles.

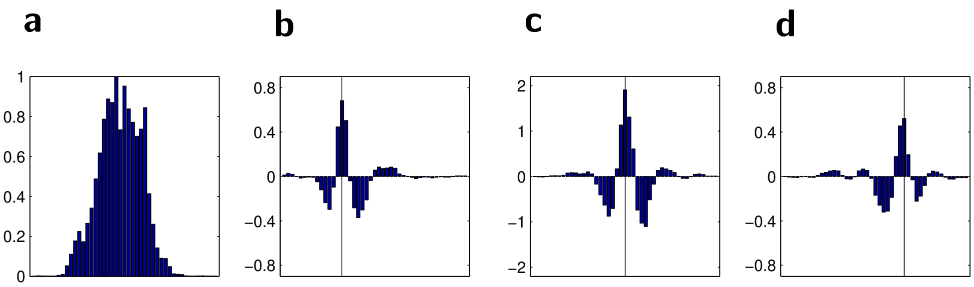

Figure 6: The two-bar experiment investigating the nonlinearity of complex cell responses. (a) The response of a complex cell to a single optimally oriented bar at various parts of the CRF. Horizontal axis: position of bar along the axis orthogonal to preferred orientation. (b)-(d) Nonlinearity of responses to two bars of same polarity. One was fixed at some given spot in the CRF, shown by the vertical line, and the other was moved as in (a). The plots show the difference of the actual response and the sum of responses for the two bars alone. The experiments are shown for the neighbourhood boxed in Figure 4, most neighbourhoods behaved very similarly.

and frequency of two adjacent cells are strongly correlated. [2] In contrast, the phases are *not* correlated.

Furthermore, we plotted the global maps for the same parameters, in Fig. 5 (b). In the orientation map, one can observe fractures, and especially pinwheels, which were determined manually and marked with white circles. The frequency map shows clearly the "blob" of low-frequency basis vectors that was already visible in Fig. 4. The phase map further confirms that the phases have no spatial structure. The map of locations shows that the model does have a local retinotopy, but the global retinotopy is less clear. Note that the neighbourhood structure was toroidal, so the edges of the map exist only in this 2-D plot, and they could be moved by rolling the map horizontally or vertically.

Finally, we analyzed the properties of the complex cells given by the model. The neighbourhood size (pooling area) for the complex cell outputs was the same as in the estimation (see Fig. 4). We varied the parameters of the optimal Gabor stimulus, computed separately for both simple and complex cells, and plotted the responses in Fig. 5 (c). Most complex cells are insensitive (invariant) to phase; moreover, their responses are somewhat insensitive to location. These two properties are in stark contrast to those of simple cells. On the other hand, cells in both categories are selective to orientation and frequency. (Most of the neighbourhoods not showing these properties are located in fractures and pinwheel centres.) A further experiment was made by simulating an experiment that measures the nonlinearity of the complex cell by presenting two bars (Movshon et al., 1978). We computed regions where the two bars gave stronger or weaker responses than predicted by linear superposition (Szulborski and Palmer, 1990). The results (Fig. 6) show clear excitatory and inhibitory regions.

5 Discussion

5.1 Comparison with V1 properties

The principal characteristics of the spatial CRF's of primate simple cells seem to be selective tuning for location, orientation, and frequency (Hubel and Wiesel, 1968; DeValois et al., 1982). These are the properties that emerged in the basic ICA or sparse coding model (Olshausen and Field, 1996). A systematic comparison (van Hateren and van der Schaaf, 1998) showed that the filters w_i (as well as the a_i, see below)

[2] In the orientation plot, we see a faint wrong diagonal (4% of points), which is due to anisotropies in the image data (van Hateren and van der Schaaf, 1998). To confirm that this was not an aliasing artifact, we first computed the results for strongly low-pass filtered data, which did not change the diagonal, and second, we rotated the raw data by 45 degrees, which simply changed the position of the diagonal.

are quite similar to the CRF's of simple cells; the main difference is that ICA gives basis vectors that are predominantly of high frequency whereas CRF's in V1 seem to be more evenly distributed. The properties of simple cells in our model are essentially similar to those in the basic ICA model.

The topography in V1 seems to be mainly arranged according to the same parameters. The fundamental topographic arrangement is by retinotopy (spatial location) (Hubel and Wiesel, 1977). Orientation-preference seems to change smoothly as well, except in pinwheels and other singularities (Blasdel, 1992). Frequency selectivity seems to be arranged topographically with respect to the cytochrome oxidase (CO) blobs, so that the blobs (or at least their centers) contain predominantly cells that prefer low-frequency cells and the interblob cells prefer higher frequencies (Tootell et al., 1988; Silverman et al., 1989; Edwards et al., 1996). In contrast, phase seems to be random, and does not seem to have any spatial structure (DeAngelis et al., 1999).

The results of our model are consistent with these data. In our results, as shown in Fig. 5 (a) and (b), orientation changes smoothly in most parts of the map, and one can find pinwheels. Phase changes randomly and does not determine the topography; this is an emergent property of the model, dictated by the statistical structure of the input. Low-frequency components are arranged in a single "blob". The fact that we have a single blob is presumably because the spatial extent of the patch is so small. With larger patches, the emergence of several blobs seems likely, especially if a more strongly high-pass filtering is performed in the preprocessing. The model shows only a local retinotopy, whereas globally retinal location is not well ordered. As with blobs, global retinotopy should be analyzed by much larger patches which would probably lead to a clear global retinotopy.

The properties of the model complex cells also give a qualitative match with V1 complex cells: The model complex cells are insensitive to phase and have larger receptive fields, while still tuned to a specific orientation and frequency. The two-bar experiment shows excitatory and inhibitory regions that are qualitatively very similar to kernels measured from complex cells (Movshon et al., 1978; Szulborski and Palmer, 1990), which may in fact be a rather general property of energy models (Sakai and Tanaka, 2000).

A further issue is how the spatial maps of the different parameters are related to each other. There are contradictory results on what the situation is like in primate V1. It has been reported (DeValois and DeValois, 1988; Blasdel, 1992) that the CO blobs tend to contain the centers of the pinwheels, but such an arrangement was not found by others (Bartfeld and Grinvald, 1992). (In the cat, contradictory results have been reported as well (Kim et al., 1999; Hübener et al., 1997)). In the results given by our model, visual inspection of Fig. 5 (b) seems to show that the maps for different parameters are essentially independent, thus conforming to the results in (Bartfeld and Grinvald, 1992).

5.2 Comparison with other models

When compared with other models on V1 topography, we see three important new features in our model.

First, our model shows emergence of a topographic organization using the above-mentioned three principal parameters: location, frequency and orientation. The use of these particular three parameters is not predetermined by the model, but determined by the statistics of the input. This is in contrast to most models that only model topography with respect one or two parameters (usually orientation possibly combined with binocularity) that are chosen in advance. The results that are closest to our model in this respect were obtained by the ASSOM model (Kohonen, 1996), but even in that model the nature of the topography was strongly influenced by an artificial manipulation of the input (a sampling window that moves smoothly in time).

Second, no other model has shown, to our knowledge, the emergence of a low-frequency blob. In fact, most models have exclusively concentrated on orientation-preference and ocular-dominance columns.

Third, our model may be the first one to explicitly show a connection between topography and complex cells. The topographic, columnar organization of the simple cells is such that complex cell properties are automatically created when considering local activations, which is related to the randomness of phases.

It is likely that the two latter properties (blobs and complex cells) can only emerge in a model that uses, firstly, natural images as input, and, secondly, is based on simultaneous activation instead of similarity of CRF's as measured by Euclidean distances or CRF correlations. This is because Euclidean distances or correlations between basis vectors of different frequencies, or of different phases, are quite arbitrary: they can obtain either large or small values depending on the other parameters. Thus, they do not offer enough information to qualitatively distinguish the effects of phase vs. frequency, so that phase can be random and frequency can produce a blob.

There are also models that do not consider V1 topography but are related on the level of theory. Our main contribution was to combine the basic framework of sparse coding with the ideas of simultaneous activation and local pooling. Simultaneous activation can also be found in some recent models of natural image statistics, where it has been considered to be related to complex cells (Hyvärinen and Hoyer, 2000; Zetzsche and Krieger, 1999) or normalization models (Simoncelli and Schwartz, 1999). A more general signal processing framework with this kind of dependencies was proposed in (Hyvärinen et al., 2001). Some relation may also be found with the models of positive factor analysis or non-negative matrix factorization (Lee and Seung, 1999; Paatero, 1997).

5.3 Extensions and critique

Extending this modelling approach to non-spatial properties of complex cells and topography may be possible by adding these properties to the input data. The basic ICA model has had some success in modelling properties of simple cells related to motion (van Hateren and Ruderman, 1998), color (Hoyer and Hyvärinen, 2000), and stereopsis (Hoyer and Hyvärinen, 2000). Such extensions would provide further tests on the validity of our model.

We have used here the basic hierarchical energy model for complex cells. There is evidence against this very simple model, and alternatives have been proposed, see (Mel et al., 1998; Sakai and Tanaka, 2000). It remains to be seen if our model could be modified to work with such alternative models.

One argument against our model could be made by referring to results that show that even in the absence of visual input, some orientation-preference can be found in simple cells (Hubel and Wiesel, 1963), and rearing in a visually restricted environment has only limited influence on the response properties of the neurons (Sengpiel et al., 1999). This is in fact not in contradiction with our modelling approach since we are modelling the combined effect of evolution and prenatal as well as postnatal development. Quite probably, part of the "estimation" of the model is accomplished by genetic instructions. Presumably, these instructions in their turn have been influenced by the natural environment. Therefore, our estimation procedure is not meant as a concrete developmental learning rule, although part of the estimation could be accomplished by such a rule.

Another point to note is the relation between the basis vectors a_i and the filters w_i. We have shown the basis vectors in Fig. 4, following (Olshausen and Field, 1996), although it is the w_i that more closely correspond to the CRF's of the model cells. It should be noted here that the a_i are basically low-pass filtered versions of the w_i. In fact, simple calculations show [3] that the a_i can be obtained by filtering the w_i (considered as image patches as in Fig. 4) by a filter whose coefficients are given by the autocovariance function of the data. This filter is a symmetric, approximately isotropic low-pass filter (Ruderman and Bialek, 1994). Thus, a_i and w_i have essentially the same orientation, location and frequency tuning properties. On the other hand, the a_i are better to visualize because because they actually correspond to parts of the image data; especially with data that is not purely spatial, visualization of the filters would not be straightforward (Hoyer and Hyvärinen, 2000).

Finally, the choice of the images that we used may be criticized as too limited. We used 13 images of natural scenes, with a large emphasis on forests, mountains, and fields, as well as animals. This may not be a very representative set of the visual input that has shaped the visual system. On the other hand, the very definition of a representative set of visual input is not clear-cut. In particular, it is not clear to what extent the direction of eye gaze should be taken into account. This may greatly affect the input (Reinagel and Zador, 1999), especially the fovea probably receives more input from "meaningful" parts of the visual scene. In future research, more emphasis may need to be laid on data acquisition. On the positive side,

[3] The autocovariance $c(x,y;x',y')$ of the data in this model equals $\sum_{ij} a_i(x,y)a_j(x',y')E\{s_i s_j\} = \sum_i a_i(x,y)a_i(x',y')$, because the s_i are uncorrelated and have unit variance. Thus we have $\sum_{x',y'} c(x,y;x',y')w_i(x',y') = a_i(x,y)$ by definition of the w_i.

the results of basic ICA seem to be quite robust regarding the choice of the data set, since qualitatively similar results have been obtained with different data sets collected by rather different means (Olshausen and Field, 1996; van Hateren and van der Schaaf, 1998; van Hateren and Ruderman, 1998; Hoyer and Hyvärinen, 2000), and one may assume that this robustness should also be found in the results of the present model.

5.4 Conclusion

We extended the ICA or sparse coding approach to a model that learns a two-layer representation for natural image data. The model exploits the fact that the components given by ICA are not actually independent. The dependencies that are not cancelled by ICA are utilized to determine a topographic (that is, columnar) organization for the simple cells.

The learning is based on finding a representation where the local activations in the first layer are maximally sparse. In other words, complex cells pool inputs from their immediate neighbourhood only, and the sparsenesses of their outputs are maximized. In contrast to most models, the topography is thus determined by the simultaneous activities of neighbouring cells, and only indirectly by the similarity of their CRF's.

The model shows emergence of a topographic organization, in addition to the simple cell properties that were already seen in the original ICA model. The topography that emerges is similar to that found in V1 in that it is based on location, frequency and orientation, and is independent of phase. Moreover, the topography has the property that the local activations are similar to complex cell responses.

Thus, the model shows that natural image statistics have a clear connection to the columnar organization of V1, in addition to the tuning properties of the individual cells. This lends further support to the hypothesis that the structure of the early visual system is strongly influenced by the input that it receives in a natural environment.

References

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In Rosenblith, W. A., editor, *Sensory Communication*, pages 217–234. MIT Press.

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, 1:371–394.

Bartfeld, E. and Grinvald, A. (1992). Relationships between orientation-preference pinwheels, cytochromase oxidase blobs, and ocular-dominance columns in primate striate cortex. *Proc. National Academy of Sciences (USA)*, 89:11906–11909.

Bell, A. and Sejnowski, T. (1997). The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338.

Blasdel, G. G. (1992). Orientation selectivity, preference, and continuity in monkey striate cortex. *Journal of Neuroscience*, 12(8):3139–3161.

Comon, P. (1994). Independent component analysis – a new concept? *Signal Processing*, 36:287–314.

DeAngelis, G. C., Ghose, G. M., Ohzawa, I., and Freeman, R. D. (1999). Functional micro-organization of primary visual cortex: Receptive field analysis of nearby neurons. *Journal of Neuroscience*, 19(10):4046–4064.

DeValois, R. L., Albrecht, D. G., and Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research*, 22:545–559.

DeValois, R. L. and DeValois, K. K. (1988). *Spatial Vision*. Oxford University Press, New York.

Durbin, R. and Mitchison, G. (1990). A dimension reduction framework for understanding cortical maps. *Nature*, 343:644–647.

Edwards, D. P., Purpura, H. P., and Kaplan, E. (1996). Contrast sensitivity and spatial frequency response of primate cortical neurons in and around the cytochromase oxidase blobs. *Vision Research*, 35(11):1501–23.

Erwin, E., Obermayer, K., and Schulten, K. (1995). Models of orientation and ocular dominance columns in the visual cortex: a critical comparison. *Neural Computation*, 7:425–468.

Field, D. (1994). What is the goal of sensory coding? *Neural Computation*, 6:559–601.

Hoyer, P. O. and Hyvärinen, A. (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, 11(3):191–210.

Hubel, D. H. and Wiesel, T. N. (1963). Receptive fields of cells in striate cortex of very young, visually inexperienced kittens. *Journal of Neurophysiology*, 26:994–1002.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. of Physiology (London)*, 195:215–243.

Hubel, D. H. and Wiesel, T. N. (1977). Functional architecture of macaque monkey visual cortex (Ferrier Lecture). *Proc. Royal Soc. London Ser. B*, 198:1–59.

Hübener, M., Shoham, D., Grinvald, A., and Bonhoeffer, T. (1997). Spatial relationships among three columnar systems in cat area 17. *J. of Neuroscience*, 17(23):9270–9284.

Hyvärinen, A. and Hoyer, P. O. (2000). Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720.

Hyvärinen, A., Hoyer, P. O., and Inki, M. (2001). Topographic independent component analysis. *Neural Computation*. in press.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430.

Jutten, C. and Herault, J. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10.

Kim, D.-S., Matsuda, Y., Ohki, K., Ajima, A., and Tanaka, S. (1999). Geometrical and topological relationships between multiple functional maps in cat primary cortex. *Neuroreport*, 10(12):2515–2522.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):56–69.

Kohonen, T. (1996). Emergence of invariant-feature detectors in the adaptive-subspace self-organizing map. *Biological Cybernetics*, 75:281–291.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, 21:105–117.

Mel, B. W., Ruderman, D. L., and Archie, K. A. (1998). Translation-invariant orientation tuning in visual "complex" cells could derive from intradendritic computations. *Journal of Neuroscience*, 18:4325–4334.

Miller, K. D. (1995). Receptive fields and maps in the visual cortex: Models of ocular dominance and orientation columns. In Domany, E., van Hemmen, J. L., and Schulten, K., editors, *Models of Neural Networks III*, pages 55–78. Springer-Verlag, New York.

Movshon, J. A., Thompson, I. D., and Tolhurst, D. J. (1978). Receptive field organization of complex cells in the cat's striate cortex. *Journal of Physiology*, 283:79–99.

Obermayer, K., Ritter, H., and Schulten, K. (1990). A principle for the formation of the spatial structure of cortical feature maps. *Proceedings of the National Academy of Science, USA*, 87:8345–8349.

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

Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325.

Paatero, P. (1997). Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 37:23–35.

Pham, D.-T., Garrat, P., and Jutten, C. (1992). Separation of a mixture of independent sources through a maximum likelihood approach. In *Proc. EUSIPCO*, pages 771–774.

Pollen, D. and Ronner, S. (1983). Visual cortical neurons as localized spatial frequency filters. *IEEE Trans. on Systems, Man, and Cybernetics*, 13:907–916.

Reinagel, P. and Zador, A. (1999). Natural scenes at the center of gaze. *Network: Computation in Neural Systems*, 10:341–350.

Ruderman, D. L. and Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Physics Review Letters*, 73(6):814–817.

Sakai, K. and Tanaka, S. (2000). Spatial pooling in the second-order spatial structure of cortical complex cells. *Vision Research*, 40:855–871.

Sengpiel, F., Stawinski, P., and Bonhoeffer, T. (1999). Influence of experience on orientation maps in cat visual cortex. *Nature Neuroscience*, 2(8):727–732.

Silverman, M. S., Grosof, D. H., DeValois, R. L., and Elfar, S. D. (1989). Spatial-frequency organization in primate striate cortex. *Proc. National Academy of Sciences (USA)*, 86(2):711–715.

Simoncelli, E. P. and Schwartz, O. (1999). Modeling surround suppression in V1 neurons with a statistically-derived normalization model. In *Advances in Neural Information Processing Systems 11*, pages 153–159. MIT Press.

Swindale, N. V. (1996). The development of topography in the visual cortex: a review of models. *Network*, 7(2):161–247.

Szulborski, R. G. and Palmer, L. A. (1990). The two-dimensional spatial structure of nonlinear subunits in the receptive fields of complex cells. *Vision Research*, 30(2):249–254.

Tootell, R. B. H., Silverman, M. S., Hamilton, S. L., Switkes, E., and Valois, R. L. D. (1988). Functional anatomy of macaque striate cortex. V. Spatial frequency. *Journal of Neuroscience*, 8:1610–1624.

van Hateren, J. H. and Ruderman, D. L. (1998). Independent component analysis of natural image sequences yields spatiotemporal filters similar to simple cells in primary visual cortex. *Proc. Royal Society ser. B*, 265:2315–2320.

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. Royal Society ser. B*, 265:359–366.

von der Malsburg, C. (1973). Self-organization of orientation-sensitive cells in the striate cortex. *Kybernetik*, 14:85–100.

Zetzsche, C. and Krieger, G. (1999). Nonlinear neurons and high-order statistics: new approaches to human vision and electronic image processing. In Rogowitz, B. and Pappas, T., editors, *Human Vision and Electronic Imaging IV (Proc. SPIE vol. 3644)*, pages 2–33. SPIE, Bellingham, WA, USA.

Acknowledgements

The authors would like to thank Pentti Laurinen, Bruno Olshausen, Tarja Peromaa, and Martin Stetter for interesting discussions; Mika Inki for performing additional experiments; and Erkki Oja for a stimulating work environment. The first author was funded by the Academy of Finland, project #44886 "New information processing principles", Finnish Centre of Excellence Programme (2000-2005).