STATISTICAL MODELS OF IMAGES AND EARLY VISION

Aapo Hyvärinen, Patrik O. Hoyer, Jarmo Hurri, and Michael Gutmann

Dept of Computer Science
Helsinki Institute of Information Technology
University of Helsinki, Finland

ABSTRACT

A fundamental question in visual neuroscience is: Why are the receptive fields and response properties of visual neurons as they are? A modern approach to this problem emphasizes the importance of adaptation to ecologically valid input. In this paper, we will review work on modelling statistical regularities in ecologically valid visual input ("natural images") and the obtained functional explanation of the properties of visual neurons. A seminal statistical model for natural images was linear sparse coding which is equivalent to the model called independent component analysis (ICA). Linear features estimated by ICA resemble wavelets or Gabor functions, and provide a very good description of the properties of simple cells in the primary visual cortex. We have introduced extensions of ICA that are based on modelling dependencies of the "independent" components estimated by basic ICA. The dependencies of the components are used to define either a grouping or a topographic order between the components. With natural image data, these models lead to emergence of further properties of visual neurons: the topographic organization and complex cell receptive fields. We have also modelled the temporal structure of natural image sequences, which provides an alternative approach to the sparseness used in most models. These models can be combined in a unifying framework that we call bubble coding. Finally, we will discuss a promising new direction of research: predictive visual neuroscience. There, the goal is to try to predict response properties of neurons in areas that are poorly understood, still based on statistical modelling of natural input.

1. INTRODUCTION

Recently, modelling images or image patches (windows) using statistical generative models has emerged as a new area of research, for reviews see (Hyvärinen et al., 2001b; Olshausen, 2003; Simoncelli and Olshausen, 2001). Such an approach has applications both in image processing and visual neuroscience.

In image processing, using statistical generative models enables principled derivation of methods for denoising, compression, and other operations (Simoncelli and Adelson, 1996; Hyvärinen, 1999b; Portilla et al., 2003). In particular, a generative model gives a prior that can be used in Bayesian methods. In this paper, we will concentrate on applications in biological modelling, although the same models could be used rather directly in image processing.

A widely-spread assumption is that biological visual systems are adapted to process the particular kind of information they receive (Field, 1994). The visual system is important for survival and reproduction, and evolutionary forces thus drive the visual system towards signal processing that is optimal for the natural stimuli. This does not imply that genetic instructions completely determine the properties of the visual system: a large part of the adaptation to the natural stimuli could be accomplished during individual development.

Natural images have important statistical regularities that distinguish them from other kinds of input. For example, the gray-scale values or luminances at different pixels have robust and non-trivial statistical dependencies. Models of the statistical structure show what a statistically adapted representation of visual input should be like. Such models thus indicate what the visual system should be like if it followed the assumption of optimal adaptation to the visual input.

Statistical models of natural images thus enable us to provide (an) answer to the fundamental question: *Why are the response properties of visual neurons as they are?* Previous theories, such as edge detection and space-frequency analysis, are unsatisfactory because they only give vague qualitative predictions on how the visual neurons should respond to visual stimulation. Statistical models offer exact quantitative prediction that often turn out to be very much in line with measurements from the visual cortex.

In the following, we first review very briefly the structure of the human visual system, see, e.g., (Palmer, 1999) for a more detailed account. Then we go on to discuss different models that we and others have developed to model the statistics of natural images and the visual system.

2. EARLY VISUAL PROCESSING IN THE HUMAN CORTEX

2.1. From the eye to the cortex

Figure 1 illustrates the earliest stages of the main visual pathway. Light is detected by the photoreceptors in the retinas, and the final output of the retinas is sent by the retinal ganglion cells through the optic nerve. The signal goes through the lateral geniculate nucleus (LGN) of

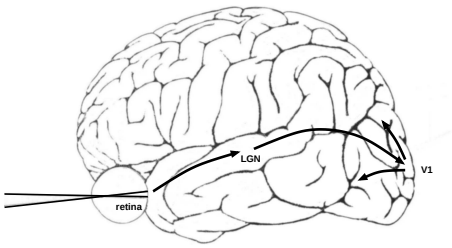

Figure 1. The main visual pathway in the human brain.

the thalamus to the visual cortex at the back of the head, where most of the visual processing is performed.

The main information processing workload of the brain is carried by nerve cells, or neurons. The majority of neurons communicate by action potentials (also called spikes), which are electrical impulses traveling down the axons (something like output wires) of neurons. Most research has concentrated on the neurons' *firing rates*, i.e. the number of spikes "fired" by a neuron per second (or some other time interval).

Thus, much of visual neuroscience has been concerned with measuring the firing rates of cells as a function of some properties of a visual input. For example, an experiment might run as follows: An image is suddenly projected onto a (previously blank) screen that an animal is watching, and the number of spikes fired by some recorded cell in the next second are counted. By systematically changing some properties of the stimulus and monitoring the elicited response, one can make a quantitative model of the response of the neuron. Such a model mathematically describes the response (firing rate) r_j of a neuron as a function of the stimulus $I(x, y)$.

In the early visual system, the response of a typical neuron depends only on the intensity pattern of a very small part of the visual field. This area, where light increments or decrements can elicit increased firing rates, is called the (classical) *receptive field* of the neuron. More generally, the concept also refers to the particular light pattern that yields the maximum response.

So, what light patterns actually elicit the strongest responses? This of course varies from neuron to neuron. The retinal ganglion cells as well as cells in the lateral geniculate nucleus typically have circular center-surround receptive field structure: Some neurons are excited by light in a small circular area of the visual field, but inhibited by light in a surrounding annulus. Other cells show the opposite effect, responding maximally to light that fills the surround but not the center. This is depicted in Figure 2a.

2.2. V1 response properties

The cells that we are modelling are mainly in the primary visual cortex (V1). Cells in V1 have more interesting receptive fields than those in the retina or LGN. The so-called *simple cells* typically have adjacent elongated (instead of concentric circular) regions of excitation and inhibition. This means that these cells respond maximally to *oriented* image structure. This is illustrated in figure 2b.

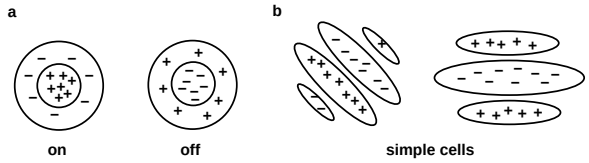

Figure 2. Typical classical receptive fields of neurons early in the visual pathway. Plus signs denote regions of the visual field where light causes excitation, minuses regions where light inhibits responses. (a) Retinal ganglion and LGN neurons typically exhibit center-surround receptive field organization, in one of two arrangements. (b) The majority of simple cells in V1, on the other hand, have oriented receptive fields.

Typically, classical receptive fields are modeled by a linear model: the response of a neuron can be predicted reasonably well by a weighted sum of the image intensities, as in

$$r_j = \sum_{x,y} W_j(x, y)I(x, y) + c, \qquad (1)$$

where $W_j(x, y)$ contains the pattern of excitation and inhibition for light for the neuron j in question, and c is the spontaneous firing rate which the neuron has with no stimulation.

Although these linear models are useful in modelling many cells, there are also neurons in V1 called *complex cells* for which these models are completely inadequate. These cells do not show any clear spatial zones of excitation or inhibition. Complex cells respond, just like simple cells, selectively to bars and edges at a particular location and of a particular orientation; they are, however, relatively invariant to the spatial phase of the stimulus. An example of this is that reversing the contrast polarity (e.g. from white bar to black bar) of the stimulus does not markedly alter the response of a typical complex cell. The responses of complex cells have often been modeled by the classical 'energy model'. (The term 'energy' simply denotes the squaring operation.) In such a model (see Figure 3) we have

$$r_j = \left(\sum_{x,y} W_{j_1}(x, y)I(x, y)\right)^2 + \left(\sum_{x,y} W_{j_2}(x, y)I(x, y)\right)^2$$

where $W_{j_1}(x, y)$ and $W_{j_2}(x, y)$ are quadrature-phase Gabor functions, i.e., they have a phase-shift of 90 degrees, one being odd-symmetric and the other being even-symmetric. It is often assumed that V1 complex cells pool the responses of simple cells, in which case the linear responses in the above equation are outputs of simple cells.

2.3. Topographic organization

A further interesting point is how the receptive fields of neighboring cells are related. In the retina, the receptive fields of retinal ganglion cells are necessarily linked to the physical position of the cells. This is due to the fact

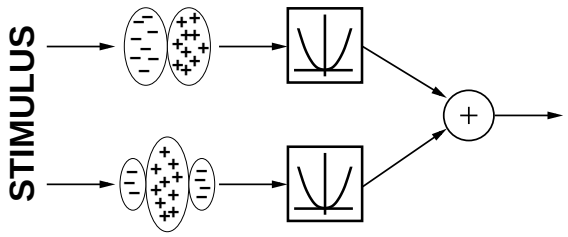

Figure 3. The classic energy model for complex cells. The response of a complex cell is modeled by linearly filtering with quadrature-phase Gabor filters (Gabor functions whose sinusoidal components have a 90 degrees phase difference), taking squares, and summing. Note that this is purely a mathematical description of the response and should not be directly interpreted as a hierarchical model summing simple cell responses.

that the visual field is mapped in an orderly fashion to the retina. Thus, neighboring retinal ganglion cells respond to neighboring areas of the visual field. However, there is nothing to guarantee the existence of a similar organization further up the visual pathway.

But the fact of the matter is that, just like in the retina, neighboring neurons in the LGN and in V1 tend to have receptive fields covering neighboring areas of the visual field. Yet this is only one of several types of organization. In V1, the orientation of receptive fields also tends to shift gradually along the surface of the cortex. In fact, neurons are often approximately organized according to several functional parameters simultaneously. This kind of *topographic organization* also exists in many other visual areas.

Topographical representations are not restricted to cortical areas devoted to vision, but are present in various forms throughout the brain (for review, see (Mountcastle, 1997)). Examples include the tonotopic map (frequency-based organization) in the primary auditory cortex and the complete body map for the sense of touch. In fact, one might be pressed to find a brain area that would not exhibit any sort of topography.

2.4. After V1

From V1, the visual signals are sent to other areas, such as V2, V4, and V5. The function of some of these areas (mainly V5) is relatively well understood (Simoncelli and Heeger, 1998), but the function of most of them is not really understood at all. For example, it is assumed that V2 is the next stage in the visual processing, but the differences in the features computed in V1 and V2 are not really known (Boynton and Hedgé, 2004).

3. LINEAR MODELS OF NATURAL IMAGES

Now we can start addressing the question of why the reponse properties of visual neurons in the cortex are as they are. The starting point is generative models of natural images, i.e. ecologically valid input.

The statistical generative models in visual modelling are typically linear, or at least they incorporate a linear part. Let us denote by $I(x, y)$ the pixel gray-scale values (point luminances) in an image, or in practice, a small image patch. The models that we consider here express each image patch as a linear superposition of some features or basis vectors A_i:

$$I(x, y) = \sum_{i=1}^{n} A_i(x, y)s_i \tag{2}$$

for all x and y. The s_i are stochastic coefficients, different from patch to patch.

In a neuroscientific interpretation, the latent variables s_i model the responses of simple cells, and the A_i are closely related to their receptive fields (see below). Thus, in the following, we will use the expressions "simple-cell outputs" or "latent variables" interchangably.

For simplicity, we assume that the number of pixels equals the number of basis vectors, in which case the linear system in Eq. (2) can be inverted. Then, each latent variable or simple-cell response is obtained by applying a linear transformation to the data; the linear transformation gives the receptive field. Denoting by W_i the coefficients of the transformation, the output of the simple cell with index i, when the input is an image patch I, is given by

$$s_i = \langle W_i, I \rangle = \sum_{x,y} W_i(x, y)I(x, y). \tag{3}$$

It can be shown (Hyvärinen and Hoyer, 2001) that the A_i are basically low-pass filtered versions of the receptive fields W_i. Therefore, the properties of the W_i and A_i are for most purposes identical.

Estimation of the model consists of determining the values of A_i, observing a sufficient number of patches I without observing the latent variables s_i. This is equivalent to determining the values of W_i, or the values of s_i for each image patch. This is a case of *unsupervised learning* since there is no "teacher" that would give the right output values s_i.

Now, the question is: How to describe the statistical properties of natural images with the linear generative model? In other words, what are the statistical properties of linear transformations of the data?

4. SPARSENESS

A considerable proportion of the models on natural image statistics is based on one particular statistical property, sparseness, which is closely related to the properties of supergaussianity or leptokurtosis (Field, 1994; Hyvärinen et al., 2001b; Olshausen and Field, 1996), and to ICA estimation methods. The outputs of linear filters that model simple-cell receptive fields are very sparse; in fact, they maximize a suitable defined measure of sparseness.

Sparseness is a property of a random variable. Sparseness means that the random variable takes very small (absolute) values and very large values more often than a gaussian random variable; to compensate, it takes values in between relatively more rarely. Thus, the random variable is activated, i.e. significantly non-zero, only rarely.

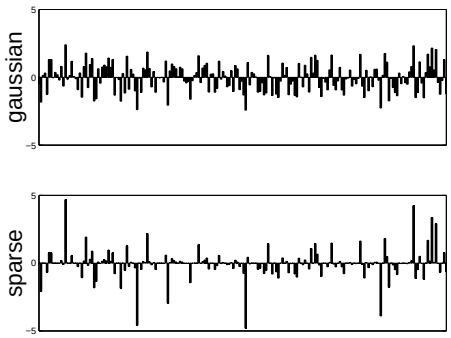

Figure 4. Illustration of sparseness. Random samples of a gaussian variable (top) and a sparse variable (bottom). The sparse variable is practically zero most of the time, occasionally taking very large values. Note that the variables have the same variance, and that the time structure is irrelevant in the definition of sparseness.

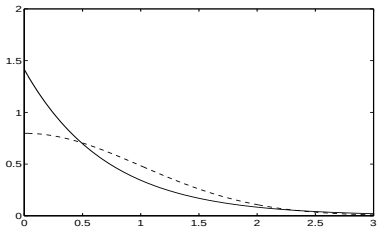

Figure 5. Illustration of a sparse probability density. Vertical axis: probability density. Horizontal axis, (absolute) value of random variable s. The sparse exponential density function is given by the solid curve. For comparison, the density of the absolute value of a gaussian random variable of the same variance is given by the dash-dotted curve.

This is illustrated in Fig. 4. We assume here and in what follows that the variable has zero mean.

The probability density function p of a sparse variable, say s, is characterized by a large value ("peak") at zero, and relatively large values far from zero ("heavy tails"). Here, "relatively" means compared to a gaussian distribution of the same variance. For example, the absolute value of a sparse random variable is often modelled as an exponential density. The exponential density is compared with the density of the absolute value of a gaussian variable in Fig. 5. If the absolute value of a symmetric random variable has an exponential distribution, the distribution is called Laplacian. Scaling the distribution to have variance equal to one, the density function is then given by

$$p(s) = \frac{1}{\sqrt{2}} \exp(-\sqrt{2}|s|) \tag{4}$$

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

For example, if G is the square function, sparseness is measured by the fourth moment $E\{s^4\}$. This is closely related to using the classical fourth-order cumulant called kurtosis, defined as $\text{kurt}(s) = E\{s^4\} - 3(E\{s^2\})^2$. If the variance is normalized to 1, kurtosis is in fact the same as the fourth moment minus a constant (three). This constant is chosen so that kurtosis is zero for a gaussian random variable. If kurtosis is positive, the variable is called leptokurtic, which is a simple operational definition of sparseness.

However, kurtosis suffers from some adverse statistical properties (Hyvärinen et al., 2001b), which is why in practice other functions may have to be used. Both information-theoretic and estimation-theoretic considerations show that in some ways the ideal functions would be such that $G(s^2)$ is equal to the logarithm of a sparse probability density function, optimally of s itself, as will be discussed below.

For example, taking the logarithm of the Laplacian density, one obtains

$$G(s^2) = -\alpha\sqrt{s^2} + \beta = -\sqrt{2}|s| - \log\sqrt{2} \tag{5}$$

In practice, a smoother version of the absolute value may be useful because the peak of absolute value at zero may lead to technical problems in optimization algorithms. A widely-used smoother version is given by $G(s^2) = \log\cosh\sqrt{s^2} = \log\cosh s$.

5. INDEPENDENT COMPONENT ANALYSIS

5.1. Independence

Maximization of sparseness with these sparseness measures is, in fact, very closely related to estimation of the model called independent component analysis (ICA).

The key concept in ICA is statistical independence. Let us consider two random variables, say y_1 and y_2. Basically, the variables y_1 and y_2 are independent if information on the value of y_1 does not give any information on the value of y_2, and vice versa. It is important to understand the difference between independence and uncorrelatedness. If the two random variables are independent, they are necessarily uncorrelated as well. However, it is quite possible to have random variables that are uncorrelated, yet strongly dependent.

Thus, correlatedness is a special kind of dependence. In fact, if two random variables y_i and y_j were independent, any nonlinear transformation of the outputs would be uncorrelated as well (Hyvärinen et al., 2001b):

$$\begin{aligned}&\text{cov}(g_1(y_i), g_2(y_j)) \\ &= E\{g_1(y_i)g_2(y_j)\} - E\{g_1(y_i)\}E\{g_2(y_j)\} = 0\end{aligned} \tag{6}$$

for any two functions g_1 and g_2. If g_1 and g_2 are identity functions $g_i(u) = u$, then the covariance in (6) is just

the ordinary covariance and the equation simply expresses that y_1 and y_2 are uncorrelated. This shows how independence is a much more general property than mere uncorrelatedness. When probing the dependence of y_i and y_j, a simple approach would thus be to consider the correlations of some nonlinear functions.

Technically, independence can be defined by the probability densities. Let us denote by $p(y_1, y_2)$ the joint probability density function (pdf) of y_1 and y_2. Let us further denote by $p_1(y_1)$ the marginal pdf of y_1, i.e. the pdf of y_1 when it is considered alone:

$$p_1(y_1) = \int p(y_1, y_2) dy_2, \tag{7}$$

and similarly for y_2. Then we define that y_1 and y_2 are independent if and only if the joint pdf is *factorizable* in the following way:

$$p(y_1, y_2) = p_1(y_1) p_2(y_2). \tag{8}$$

This definition extends naturally for any number n of random variables, in which case the joint density must be a product of n terms.

5.2. Correlations in natural images and preprocessing

As an example of dependencies in natural images, let us consider the gray-scale values of two neighbouring pixels. Going through many different locations in an image in random order, we get two random variables, each of them giving the gray-scale value in one of the pixels. These random variables are *not* independent. One of the basic statistical properties of natural images is that two neighbouring pixels are correlated.

To make estimation of ICA easier, it is useful to try to reduce dependencies in the data as a preprocessing step. This usually means transforming the data to a space where the variables are uncorrelated, a process called decorrelation. We can then do ICA in this transformed space and after ICA, invert the transformation to get back to the original space.

Such a preprocessing is usually done in two different steps in the context of image analysis. First, we can consider only the local changes in gray-scale values (called contrast), and remove the local mean (called the DC component) from the image. This also implies that the s_i in the linear model have zero mean. Second, we whiten the data in the spatial domain. This means that the data is linearly transformed to an image space where for any two spatial points (x, y) and (x', y') the value of $I(x, y)$ and $I(x', y')$ are uncorrelated, and all points are normalized to unit variance.

The main utility of the whitening step is due to the fact that simple cell outputs are assumed uncorrelated and normalized to unit variance in all the relevant models, and these properties are equivalent to orthonormality of the A_i in the whitened space (Hyvärinen et al., 2001b). Thus, in the whitened space, we can then consider orthonormal transformations only, i.e. $\sum_{x,y} A_i(x, y) A_j(x, y) = 0$ if $i \neq j$ and 1 if $i = j$ (the same applies for the W_i). This reduces the number of free parameters in the model. The nonlinear correlations that remain in the data should then be used toetimate the vectors A_i.

5.3. Definition of ICA as generative model

The central point in ICA is the interpretation of the linear mixing in (2) as a statistical generative model, i.e. we assume that the image data I is really generated according to such a linear superposition. The vectors A_i are considered parameters in this statistical model, and the s_i are latent random variables. Then, once we define what the joint probability distribution of the s_i is, we have defined a statistical model of the observed data.

As the name says, the key idea is to define the distribution of the s_i by assuming that the s_i are mutually statistically independent. We do not need to define their distribution any more than this. In principle, the marginal distributions can be considered extra parameters in the model and they can be estimated from the data just as the matrix A_i. However, most ICA methods do not explicitly perform estimation of the distributions of the s_i but either do it implicitly or assume that we know reasonably well what the distributions are.

A fundamental theorem in the theory of ICA says that if the components s_i are not only independent, but also *nongaussian*, then the model can actually be estimated. In other words, we can recover the vectors A_i if the data were actually generated by the model. Thus, the model can only be meaningfully applied on nongaussian data.

Let us assume that we have observed K image patches $I_k(x, y), k = 1, \ldots, K$ that are extracted (sampled) at random locations in some natural images. Because they are randomly collected, we can assume that the patches are independent from each other. Thus, the probability of observing all these patches is the product of the probabilities of each patch. This gives the likelihood L of the observed data:

$$L(W_i, i = 1, \ldots, n) = \prod_{k=1}^{K} |\det(\mathbf{W})| \prod_{i=1}^{n} p_i(\langle W_i, I \rangle) \tag{9}$$

where $\mathbf{W}$ is a matrix that contains the vectors W_i as its rows. The term $|\det \mathbf{W}|$ is the determinant of the Jacobian that is needed when a probability density of a transformation is calculated. We can now apply classic maximum likelihood estimation to estimate the ICA model.

It is much simpler to look at the logarithm of the likelihood. Obviously, maximization of the likelihood is the same as maximization of this log-likelihood. Maximum likelihood estimation now means that we maximize this probability with respect to the parameters, that is, the weights W_i. By simple rearrangemnt we see that the

$$\log L(W_i, i = 1, \ldots, n) = K \log |\det(\mathbf{W})| + \sum_{i=1}^{n} \sum_{k=1}^{K} \log p_i(\langle w_i, I_k \rangle) \tag{10}$$

As discussed above, if the data is whitened as a preprocessing step, we can constrain $\mathbf{W}$ to be orthogonal. Then, the determinant of $\mathbf{W}$ is equal to one. So the first term on the right-hand-side is zero and can be omitted. The second term on the right-hand-side is the expectation of a nonlinear function $\log p_i$ of the output of the feature detector (more precisely, an estimate of that expectation, since this is computed over the sample). Thus, what the likelihood really boils down to is measuring the expectations of the form $E\{G(s_i)\}$ for a function G which is here given by the logarithm of p_i.

The connection to maximization of sparseness is now evident. Since the outputs are constrained to have unit variance (by prewhitening the data and constraining $\mathbf{W}$ to be orthogonal), maximization of the likelihood is equivalent to maximization of the sparsenesses of the outputs, *if* the functions $G(s) = \log p_i(s)$ are of the form required for sparseness measurements, i.e. if we we can express them as

$$\log p_i(s) = h_i(s^2) \tag{11}$$

where the functions h_i are *convex*. It turns out that this is usually the case in natural images.

The results obtained when an ICA or sparse coding model is estimated for image patches (Olshausen and Field, 1996; Bell and Sejnowski, 1997) are shown in Fig. 8. A comparison with simple-cell measurements shows quite a good match with respect to almost all parameters (van Hateren and van der Schaaf, 1998; van Hateren and Ruderman, 1998).

6. TEMPORAL COHERENCE AND BURST CODING

An alternative to sparseness is given by temporal coherence or stability (Földiák, 1991; Kayser et al., 2001; Hurri and Hyvärinen, 2003a; Stone, 1996; Wiskott and Sejnowski, 2002). This means that when the input consists of natural image *sequences*, i.e. video data, the outputs of simple cells in subsequent time points should be "coherent" or "stable", i.e. change as little as possible. The change can be defined in many ways, and therefore temporal coherence can give rise to quite different definitions and measures.

First, it must be noted that using ordinary *linear* (auto)correlation or covariance is *not* enough to produce well-defined receptive fields. That is, if we measure the temporal coherence of a cell output $s(t)$, centered to have zero mean, as

$$\text{corr}(s(t), s(t-\tau)) = E\{s(t)s(t-\tau)\}, \tag{12}$$

where τ is a time lag (delay), maximization of this measure does not characterize most simple-cell receptive fields. In fact, this measure is maximized by low-pass filters, such as the DC component of image patches (Hurri and Hyvärinen, 2003a).

This failure of linear measures can be partly explained by basic results in the literature of blind source separation (Hyvärinen et al., 2001b). The autocovariance (for a given time lag) of the sum $a_i s_i + a_j s_j$ of two independent signals is given by $a_i^2 \text{cov}(s_i(t), s_i(t-\tau)) + a_j^2 \text{cov}(s_j(t), s_j(t-\tau))$. Consider a case where s_i and s_j have equal variances and autocovariances. Then, if the mixing coefficients fulfill $a_i^2 + a_j^2 = 1$, the mixture has the same variance *and* the same autocovariance as the original signals. There is an infinite number of such sums, and thus we cannot tell them apart if we just look at the autocovariance (and variance). This shows that maximization of autocorrelation does not properly define linear filters, and we have to use nonlinear autocorrelations.

Thus, we must use some kind of *nonlinear temporal correlations*. We have proposed (Hurri and Hyvärinen, 2003a) that temporal coherence could be measured by the correlation of squares (energies):

$$\text{corr}(s(t), s(t-\tau)) = E\{s(t)^2 s(t-\tau)^2\} \tag{13}$$

It was found that the typical simple-cell receptive fields maximize this criterion, just like sparseness. This measure was inspired by recent advances in the theory of blind source separation, where it has been shown that the correlation of squares is a valid measure for blind source separation (Hyvärinen, 2001). In fact, this method can be seen as a variant of the class of blind separation methods using *nonstationary variance* (Matsuoka et al., 1995; Pham and Cardoso, 2000).

Thus, when properly defined and measured, temporal coherence does provide an alternative to sparseness, leading to the emergence of principal simple-cell receptive field properties from natural images. The result of applying temporal coherence on natural image sequences is shown in Figure 9.

Note that the outputs the obtained receptive fields are also sparse – that is, practically zero most of the time, occasionally taking large values. The combination of sparseness and temporal coherence of activity points to a *burst code*, infrequent periods of high activity in the outputs of simple cells, see also (Reinagel, 2001).

7. DEPENDENCIES BETWEEN COMPONENTS

7.1. Definition and models

The third important statistical property used in these models considers the relationships between the different latent components (outputs of simple cells) s_i in (2). When using sparseness / ICA or temporal coherence, the outputs of simple cells s_i are usually assumed independent, i.e. the value of s_j cannot be used to predict s_i for $i \neq j$.

However, when ICA or sparse coding models are estimated for natural images, the obtained components are *not* independent. Basically, there are not enough parameters in the model to render the estimated linear components completely independent, since independence is a very complex phenomenon (cf. Eq.6). What ICA is able to do is to find the linear transformation that makes the components as independent as possible by a linear transformation, but some dependencies still remain.

Thus, we need to model the statistical dependencies of the linear filters, assuming that their joint distribution is

dictated by the natural image input (Simoncelli and Schwartz, 1999; Hyvärinen and Hoyer, 2000; Hyvärinen et al., 2001a). Remaining dependencies actually offer a great opportunity because it means that we can hope to model further properties of visual neurons, such as complex cell receptive fields and topography, by building more sophisticated statistical models of natural images.

Note that we must consider *nonlinear* correlations. Linear correlations are not interesting in this respect because they can easily be set to zero by standard whitening procedures. In fact, in ICA estimation, the components are often constrained to be exactly uncorrelated, as discussed above.

When probing the dependence of s_i and s_j, a simple approach would be to consider the correlations of some nonlinear functions, just as in the case of temporal coherence. In image data, the principal form of dependency between two simple-cell outputs seems to be captured by the correlation of their energies, or squares s_i^2. This means that

$$\text{cov}(s_i^2, s_j^2) = E\{s_i^2 s_j^2\} - E\{s_i^2\}E\{s_j^2\} \neq 0. \quad (14)$$

Here, we assume that this covariance is positive, which is the usual case.

Intuitively, correlation of energies means that the cells tend to be active, i.e. have non-zero outputs, at the same time, but the actual values of s_i and s_j are not easily predictable from each other. For example, if the variables are defined as products of two independent components o_i, o_j and a common "variance" variable v (Hyvärinen et al., 2001a; Simoncelli and Olshausen, 2001):

$$s_i = o_i v \quad (15)$$

$$s_j = o_j v \quad (16)$$

then s_i and s_j are uncorrelated, but their energies are not.

While the formulation above makes energy correlation easy to understand by using a separate variance variable v, it is not very suitable for practical computations, in which we need a simple expression for the joint probability density function of s_i and s_j. A simple density that incorporates both energy correlation and sparseness is given by (Hyvärinen and Hoyer, 2000; Hyvärinen et al., 2001a)

$$p(s_i, s_j) = \frac{3}{2\pi} \exp(-\sqrt{3}\sqrt{s_i^2 + s_j^2}) \quad (17)$$

This could be considered as a two-dimensional generalization of the Laplacian distribution. (This density has been standardized to that its mean is zero and the variances are equal to one.) The correlation of energies in this probability distribution is illustrated in Fig. 6. A generalization of the probability density to more than two dimensions is straightforward by just taking the sum of the squares of more than two components inside the square root in the exponential; the scaling and additive constants are then difficult to calculate but they are rarely needed.

Just as in the case of sparseness measures, the density in Eq. (17) gives us a measure of the combination of

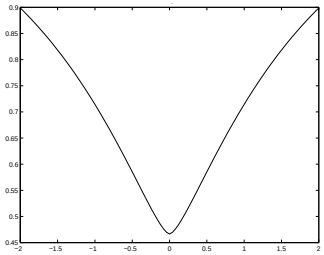

Figure 6. Illustration of the energy correlation in the probability density in Equation (17). The conditional variance of s_j (vertical axis) for a given s_i (horizontal axis). Here we see that conditional variance grows with the square (or absolute value) of s_i.

energy correlation and sparseness by considering the expectation of the log-density. We can take the logarithm of the density to obtain a function of the form

$$E\{G(s_i^2 + s_j^2)\} \quad (18)$$

where $G(b) = -\sqrt{b}$, up to irrelevant constants. This is a measure that is simple to compute. To get insight to this measure, consider what happens when G is the square function. Then the measure gives $E\{s_i^4 + s_j^4 + 2s_i^2 s_j^2\}$. The expectations of the first two terms measure sparseness just as kurtosis, while the expectation of the third is just the first term in the covariance of the squares. In practice, however, we prefer the logarithm of the density function to the square function because of the same statistical reasons (discussed above) that we prefer the log-density to kurtosis as a measure of sparseness.

7.2. Subspaces based on dependencies

The correlation of energies could be embedded in a model of natural image statistics in many ways. A very simple way would be to *divide the latent variables into groups* (Cardoso, 1998), so that the s_i in the same group have correlation of energies, whereas s_i in different groups are independent. In such a model (Hyvärinen and Hoyer, 2000), it was found that the groups (called "independent subspaces") show emergence of complex cells properties, see Figs 10 and 7. Thus, after estimation of the model, simple cells that pool to the same complex cell have energy correlations, whereas simple cells that are not pooled together are independent.

The sum of squares inside a group (which could be considered an estimate of the variance variable associated with that group) has the principal invariance properties of complex cells. That is, the sum of squares is largely invariant to changes in the phase of the input stimulus, while still being very selective to orientation and frequency. This can be understood by noting that the basis vectors in the same subspace have very similar orientations and frequencies (and rather similar locations), whereas their phases are quite different from each other. An invariant feature (complex cell output) is thus computed by summing up responses of lower-order features

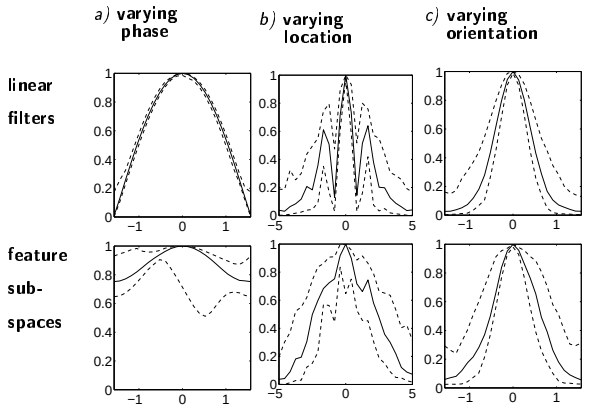

Figure 7. Invariances of independent subspaces (Hyvärinen and Hoyer, 2000). Gabor stimuli were input to the independent subspaces and the responses were compared with the outputs of the linear filters. First, we determined the Gabor stimulus that elicited the maximum response. Then, we varied the stimulus according to one of the parameters (phase, location, or orientation). In all plots, the solid line gives the median response in the population of all cells (filters or subspaces), and the dotted lines give the 90% and 10% percentiles of the responses. Top row: responses (in absolute values) of linear filters (simple cells). Bottom row: responses of feature subspaces (complex cells). a) Effect of varying phase. b) Effect of varying location (shift). c) Effect of varying orientation. We see that the subspace (complex cell response) is invariant to phase of the input, while selective to the other parameters; linear filter (simple cell) reponses are not invariant with respect to any of the parameters.

(simple cell or linear filter outputs) over the invariant dimension. This is not unlike classical models of complex cell responses as in Fig. 3.

7.3. Topography based on dependencies

Instead of a simple grouping, we have also proposed a more sophisticated way of modelling the correlations of squares of simple-cell outputs, based on topography or spatial organization of the cells (Hyvärinen et al., 2001a; Hyvärinen and Hoyer, 2001). The concept of cortical topography was reviewed earlier in Section 2.

Let us assume that the components s_i are arranged on a two-dimensional grid or lattice as is typical in topographic models (Kohonen, 1995). The restriction to 2D is motivated by cortical anatomy, but is not essential. The topography is formally expressed by a neighbourhood function $h(i, j)$ that gives the proximity of the components (cells) with indices i and j. Here the index i of a cell is defined as the cell's location on the two-dimensional lattice; therefore the indices are also two-dimensional. Typically, one defines that $h(i, j)$ is 1 if the cells are sufficiently close to each other, and 0 otherwise.

Our purpose was to define a statistical model in which the *topographic organization reflects the statistical dependencies* between the components. The components (simple cells) are arranged on the grid so that any two cells that are close to each other have dependent outputs, whereas cells that are far from each other have independent outputs. Since we are using the correlation of energies as the measure of dependency, the energies are strongly positively correlated for neighbouring cells.

We have defined such a statistical model, topographic ICA (Hyvärinen et al., 2001a; Hyvärinen and Hoyer, 2001), which incorporates just this kind of dependencies and can be estimated for natural images. When the model is applied on natural image data (see Fig. 11), the organization of simple cells is qualitatively very similar to the one found in the visual cortex: there is orderly arrangement with respect to such parameters as location, orientation, and spatial frequency – and no order with respect to phase. This is the first model that shows emergence of all these principal properties of cortical topography (Hyvärinen and Hoyer, 2001).

An interesting point is that the topography defined by dependencies is closely related to complex cells: The topographic matrix $h(i, j)$ can be interpreted as the pooling weights from simple cells to complex cells. The pooling weights have now been set by making the assumption that complex cells only pool outputs of simple cells that are near-by on the topographic grid. Thus, we see how modelling the dependencies by topography is a generalization of a simple division of the cells into groups.

This model of topography defined by energy correlation is very different from those typically used in models of (cortical) topography. Usually, the similarity of simple cells is defined by Euclidean distances or related measures, but correlation of energies is a strongly non-Euclidean measure.

8. BUBBLES: A UNIFYING FRAMEWORK

Now we discuss a unifying theoretical framework for the statistical properties discussed above: sparseness, temporal coherence, and topography. This is based on the concept of a spatiotemporal bubble (Hyvärinen et al., 2003).

The key to the unifying framework is the observation that in the models above, we used the same kind of dependence through variances (which expresses itself in the correlation of squares or energies) to model two different things: dependencies between the latent variables, and temporal dependencies of a single latent variable.

Combination of sparseness and topography means that each input activates a limited number of spatially limited "blobs" on the topographic grid, as in topographic ICA. If these regions are also temporally coherent, they resemble activity bubbles as found in many earlier neural network models. A spatiotemporal activity bubble thus means the *activation of a spatially and temporally limited region* of cells, or in general, representational units. This is illustrated in Fig. 12 for a one-dimensional topography.

What could such bubbles represent in practice? Since we are about to define a general-purpose unsupervised learning procedure, the meaning of bubbles depends

on the data on which they are applied. In the case of natural image sequences, we can assume that the topographic grid is rather similar to the one obtained by topographic ICA. Then, a bubble would mean activation of Gabor-like basis vectors with similar orientation and spatial frequency, in near-by points on the image. This would correspond to a *short contour element* of given orientation and spatial frequency. In contrast to an "independent component" of an image, this contour element can move a bit, and its phase can change, during the temporal extent of the bubble.

Now, we formulate a generative model based on activity bubbles. We postulate a higher-order random process u that determines the variance at each point. This non-negative, highly sparse random process obtains independent values at each point in time and space (space referring to the topographic grid). For simplicity, let us denote the location on the topography by a single index i. Then, the variances v of the observed variables are obtained by a spatiotemporal convolution

$$v_i(t) = \sum_j h(i,j)[\varphi(t) * u_j(t)] \tag{19}$$

where $h(i,j)$ is the neighbourhood function that defines the spatial topography, and φ is a temporal smoothing kernel. The simple-cell outputs are now obtained by multiplying simple gaussian white noise $o_i(t)$ by this variance signal:

$$s_i(t) = v_i(t) o_i(t) \tag{20}$$

Finally, the latent signals $s_i(t)$ are mixed linearly to give the image. Denote by $I(x,y,t)$ an image sequence. Then the mixing can be expressed as

$$I(x,y,t) = \sum_{i=1}^{n} a_i(x,y) s_i(t). \tag{21}$$

The three Eqs. (19–21) define a statistical generative model for natural image sequences.

The higher-order process $u_i(t)$ could be called the bubble process. When this process obtains a value that is different from zero, which is a rare event by definition, a bubble is created: The non-zero value spreads to neighbouring temporal and spatial locations due to the smoothing by φ and h. The spread of activation means that simple cells have large variances inside that spatiotemporal window.

For experiments and estimation methods regarding the bubble model, see (Hyvärinen et al., 2003).

9. A TWO-LAYER MODEL WHERE BOTH LAYERS ARE ESTIMATED

We have also developed a two-layer model of natural image sequences that has the interesting property that both layers can be estimated (Hurri and Hyvärinen, 2003b). This is in stark contrast to the models discussed above that fix the second layer (pooling of simple-cell responses) beforehand, and only estimate the basis vectors (linear mixing matrix).

Technically, the estimation of two-layer models is quite difficult. In the models introduced above, estimation of the pooling weights is possible, in principle, by considering them as parameters just as the basis vectors. However, this introduces a normalization constant in the likelihood, because the integral of the probability density must equal one for any values of the pooling weights. Evaluation of this constant is most difficult (see, however, recent theoretical developments in (Hyvärinen, 2005)).

We have been able to circumvent this problem by using a multivariate autoregressive model on the activity levels of simple cells. The activity levels correspond to the variances used in earlier sections, but for technical reasons, they are here defined simply as the absolute values. Let us denote by $\mathbf{abs}(\mathbf{s}(t))$ a vector that contains the absolute values of the elements of $\mathbf{s}(t)$. Further, let $\mathbf{v}(t)$ denote a driving noise signal in the autoregressive process. Let us denote by $\mathbf{M}$ a $K \times K$ matrix that gives the autoregressive coefficients, and let τ denote a time lag. Our model for the activities is a *constrained multidimensional first-order autoregressive process*, defined by

$$\mathbf{abs}(\mathbf{s}(t)) = \mathbf{M}\,\mathbf{abs}(\mathbf{s}(t-\tau)) + \mathbf{v}(t). \tag{22}$$

Just as in ICA, the scale of the latent variables is not well defined, so we define that the variances of $s_i(t)$ are equal to unity.

The model is complicated by the fact that the absolute values must be non-negative. Thus, there are dependencies between the driving noise $\mathbf{v}(t)$ and the $\mathbf{s}(t-\tau)$. To define a generative model for the driving noise $\mathbf{v}(t)$ so that the non-negativity of the absolute values holds, we proceed as follows. Let $\mathbf{u}(t)$ denote a zero-mean random vector with components which are statistically independent of each other. We define $\mathbf{v}(t) = \max(-\mathbf{M}\mathbf{abs}(\mathbf{s}(t-\tau)), \mathbf{u}(t))$ where the maximum is computed componentwise.

To make the generative model complete, a mechanism for generating the signs of components $\mathbf{s}(t)$ must be included. We specify that the signs are generated randomly with equal probability for plus or minus after the strengths of the responses have been generated. All the signs are mutually independent, both over time and the cell population, and also independent of the activity levels. Note that one consequence of this random generation of signs is that that filter outputs are uncorrelated (Hurri and Hyvärinen, 2003b).

We have developed a method for estimating both the autoregressive matrix $\mathbf{M}$ and the basis vectors simultaneously. This is important because in ICA and related methods, the set of basis vectors is not well-defined because of multiple local minima. Furthermore, there is little justification to assume that the maximally independent basis vectors given by ICA would be the optimal ones to use in a multi-layer model, since the structure of the higher layer affects the likelihood. For a description of the estimation method, and an interesting graphical representation of the resulting basis vectors and $\mathbf{M}$, see (Hurri and Hyvärinen, 2003b).See also (Karklin and Lewicki, 2003) for related

work.

In fact, after developing this model for natural image sequences, we realized that it can be generalized to a model where the quantitative values of the dependencies (correlations of squares) are arbitrary. This leads to a separation method that is *double-blind* in the sense that no apriori assumptions are made either on the mixing matrix or on the higher-order correlations. See (Hyvärinen and Hurri, 2004) for a detailed description of this concept.

10. PREDICTIVE MODELLING BEYOND V1

It would be most useful if we could use this modelling endeavour, based on statistical models of ecologically valid stimuli, in a *predictive* manner. This means we would be able to predict properties of cells in the visual cortex, in cases where the properties have not yet been demonstrated experimentally. This would give testable, quantitative hypotheses that might lead to great advances, especially in the research in extrastriate areas such as V2, whose function is not well understood at this point. Here, we describe our recent attempts to accomplish such predictive modelling. In particular, we attempt to predict properties of a third processing step that follows the two described above in Figure 3, i.e. linear filtering and summation of squares that are assumed to correspond to simple and complex cells in V1.

The basic idea is to fix these two first processing stages, to compute the output of such a two-layer model when the input consists of natural images, and then model these outputs by a suitable statistical model. This was done in (Hyvärinen et al., 2005), where we input a large number of natural image patches into model complex cells that computed the sum of squares of outputs of two simple cells, one odd-symmetric and the other even-symmetric. Then, we performed independent component analysis of the complex cell outputs using the FastICA algorithm (Hyvärinen, 1999a).

ICA estimates here higher-order features that correspond to typical patterns of complex cell activity. A random selection of such higher-order features learned from natural images is shown Figure 13. What we can see is emergence of collinear features. That is, the higher-order features code for the simultaneous activation of complex cells that together form something similar to a straight line segment. (An earlier model that showed this kind of emergence was given in (Hoyer and Hyvärinen, 2002)).

What is remarkable in these results is that many cells pool (combine) responses over different frequencies. The activity of a single higher-order feature codes for the simultaneous activity of complex cells that are in different frequency bands but have similar orientations and locations.

What is the functional meaning of the pooling we have found? We propose that this spatially coherent pooling of multiple frequencies leads to representation of an edge that is more realistic than the band-pass edges given by typical Gabor filters (Griffin et al., 2004). Presumably, this is largely due to the fact that natural images contain many sharp, step-like edges that are not contained in a single frequency band. Thus, representation of such edges is difficult unless information from different frequency bands is combined. In terms of frequency channels, the model predicts that frequency channels should be pooled together after complex cell processing.

The results in (Hyvärinen et al., 2005) are an instance of predictive modelling, where we attempt to predict properties of cells and cell assemblies that have not yet been observed in experiments. To be precise, the prediction is that in V2 (or some related area) there should be cells whose optimal stimulus is a broad-band edge that has no sidelobes while being relatively sharp, i.e. the optimal stimulus is closer to a step-edge than the band-pass edges that tend to be optimal for V1 simple and complex cells. The optimal stimulus should also be more elongated (Polat and Tyler, 1999) than what is usually observed in V1, while being highly selective for orientation.

Statistical models of natural images offer a framework that lends itself to predictive modelling of the visual cortex. First, they offer a framework where we often see emergence of new kinds of feature detectors — sometimes very different from what was expected when the model was formulated. Second, the framework is highly constrained and data-driven. The rigorous theory of statistical estimation makes it rather difficult to insert the theorist's subjective expectations in the model, and therefore the results are strongly determined by the data. Third, the framework is very constructive. From just a couple of simple theoretical specifications, e.g. non-Gaussianity, natural images lead to the emergence of complex phenomena.

11. CONCLUSION

Modelling the statistical structure of natural images is useful in vision research as well as in image processing. Possibly the most fundamental model is ICA, although it was originally motivated by sparse coding. The obtained components are not really independent, which shows, in fact, an opportunity to model further aspects of the visual system.

We have developed models on the dependencies of the "independent" components. The most important kind of dependency seems to be the *correlation of squares* (energies), in other words, dependence through variances or activity levels. These dependencies are modelled by 1) independent subspaces and 2) a topographic organization of the components based on their dependency structure.

We have also modelled the temporal structure of natural image sequences using the very same kind of (temporal) dependencies through variances. This eventually lead to the unifying framework of spatiotemporal activity *bubbles*. Finally, we have developed a method of *double-blind* source separation, which is blind to the particular higher-order correlations of the components as well.

A most interesting line of research is where we can use this framework to predict response properties of cells

in areas (such as V2) which are poorly understood. First attempts in this direction are based on modelling the outputs of fixed complex cell models by statistical models.

References

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

Boynton, G. and Hedg´e, J. (2004). Visual cortex: The continuing puzzle of area V2. *Current Biology*, 14:R523–R524.

Cardoso, J.-F. (1998). Multidimensional independent component analysis. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'98)*, Seattle, WA.

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

F¨oldi´ak, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3:194–200.

Griffin, L., Lillholm, M., and Nielsen, M. (2004). Natural image profiles are most likely to be step edges. *Vision Research*, 44(4):407–421.

Hoyer, P. O. and Hyv¨arinen, A. (2002). A multi-layer sparse coding network learns contour coding from natural images. *Vision Research*, 42(12):1593–1605.

Hurri, J. and Hyv¨arinen, A. (2003a). Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation*, 15(3):663–691.

Hurri, J. and Hyv¨arinen, A. (2003b). Temporal and spatiotemporal coherence in simple-cell responses: A generative model of natural image sequences. *Network: Computation in Neural Systems*, 14(3):527–551.

Hyv¨arinen, A. (1999a). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634.

Hyv¨arinen, A. (1999b). Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. *Neural Computation*, 11(7):1739–1768.

Hyv¨arinen, A. (2001). Blind source separation by nonstationarity of variance: A cumulant-based approach. *IEEE Transactions on Neural Networks*, 12(6):1471–1474.

Hyv¨arinen, A. (2005). Estimation of non-normalized statistical models using score matching. *J. of Machine Learning Research*, 6:695–709.

Hyv¨arinen, A., Gutmann, M., and Hoyer, P. (2005). Statistical model of natural stimuli predicts edge-like pooling of spatial frequency channels in V2. *BMC Neuroscience*, 6(12).

Hyv¨arinen, 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.

Hyv¨arinen, A. and Hoyer, P. O. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, 41(18):2413–2423.

Hyv¨arinen, A., Hoyer, P. O., and Inki, M. (2001a). Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558.

Hyv¨arinen, A. and Hurri, J. (2004). Blind separation of sources that have spatiotemporal variance dependencies. *Signal Processing*, 84(2):247–254.

Hyv¨arinen, A., Hurri, J., and V¨ayrynen, J. (2003). Bubbles: A unifying framework for low-level statistical properties of natural image sequences. *J. of the Optical Society of America A*, 20(7):1237–1252.

Hyv¨arinen, A., Karhunen, J., and Oja, E. (2001b). *Independent Component Analysis*. Wiley Interscience.

Karklin, Y. and Lewicki, M. S. (2003). Learning higher-order structures in natural images. *Network: Computation in Neural Systems*, 14(3):483–500.

Kayser, C., Einh¨auser, W., D¨ummer, O., K¨onig, P., and K¨ording, K. (2001). Extracting slow subspaces from natural videos leads to complex cells. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN2001)*, pages 1075–1080, Vienna, Austria.

Kohonen, T. (1995). *Self-Organizing Maps*. Springer.

Matsuoka, K., Ohya, M., and Kawamoto, M. (1995). A neural net for blind separation of nonstationary signals. *Neural Networks*, 8(3):411–419.

Mountcastle, V. B. (1997). The columnar organization of the neocortex. *Brain*, 120:701–722.

Olshausen, B. A. (2003). Principles of image representation in visual cortex. In Chalupa, L. and Werner, J., editors, *The Visual Neurosciences*. MIT Press.

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.

Palmer, S. E. (1999). *Vision Science – Photons to Phenomenology*. The MIT Press.

Pham, D.-T. and Cardoso, J.-F. (2000). Blind separation of instantaneous mixtures of non-stationary sources. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 187–193, Helsinki, Finland.

Polat, U. and Tyler, C. (1999). What pattern the eye sees best. *Vision Research*, 39(5):887–895.

Portilla, J., Strela, V., Wainwright, M., and Simoncelli, E. (2003). Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Transactions in Image Processing*, 12(11):1338–1351.

Reinagel, P. (2001). How do visual neurons respond in the real world? *Current Opinion in Neurobiology*, 11(4):437–442.

Simoncelli, E. and Heeger, D. (1998). A model of neuronal responses in visual area mt. *Vision Research*, 38(5):743–761.

Simoncelli, E. P. and Adelson, E. H. (1996). Noise removal via bayesian wavelet coring. In *Proc. Third IEEE Int. Conf. on Image Processing*, pages 379–382, Lausanne, Switzerland.

Simoncelli, E. P. and Olshausen, B. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–216.

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.

Stone, J. (1996). Learning visual parameters using spatiotemporal smoothness constraints. *Neural Computation*, 8(7):1463–1492.

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.

Wiskott, L. and Sejnowski, T. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770.

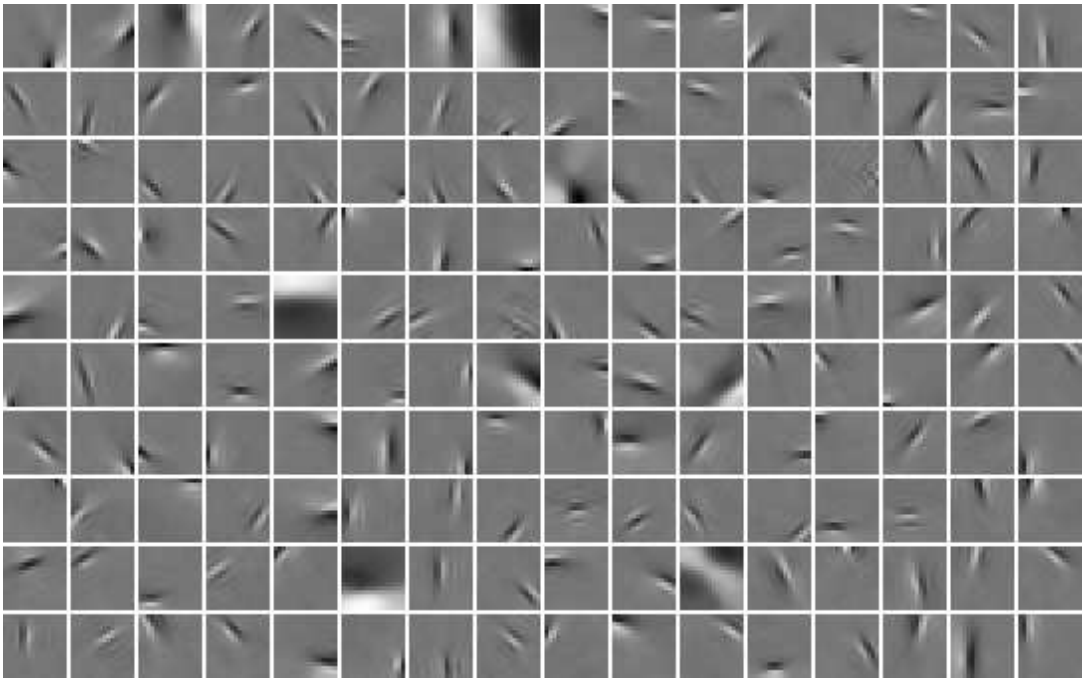

Figure 8. Basis vectors estimated by ICA / sparse coding. A set of 10,000 image patches of 16×16 pixels were randomly sampled from natural images, and input to the FastICA algorithm (Hyvärinen, 1999a)

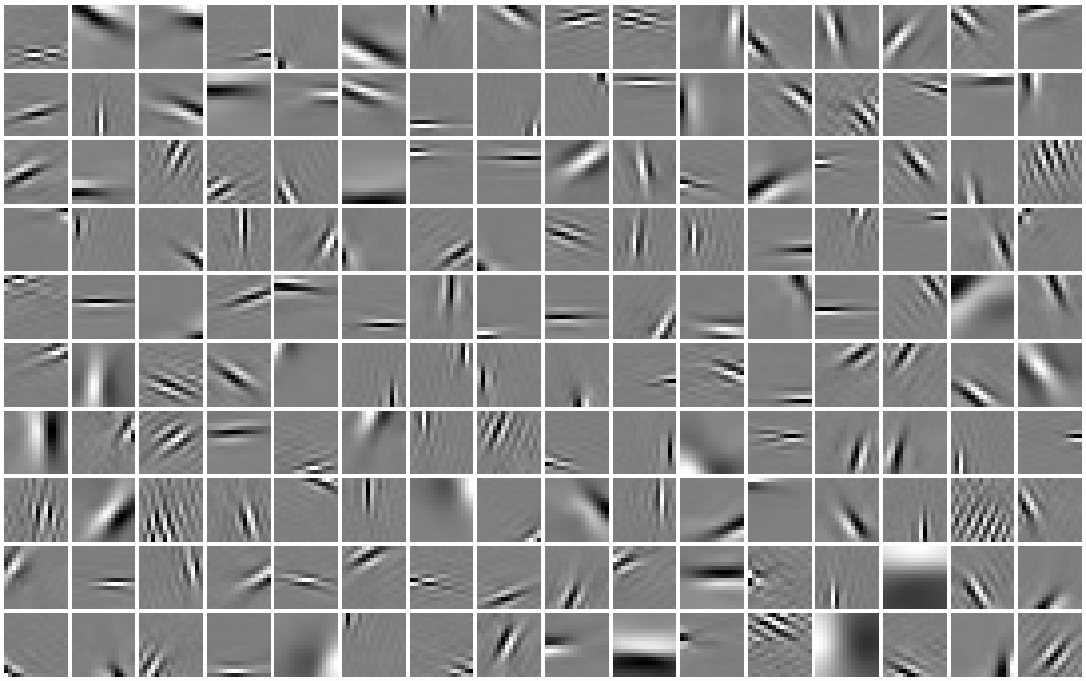

Figure 9. Basis vectors estimated by temporal coherence.

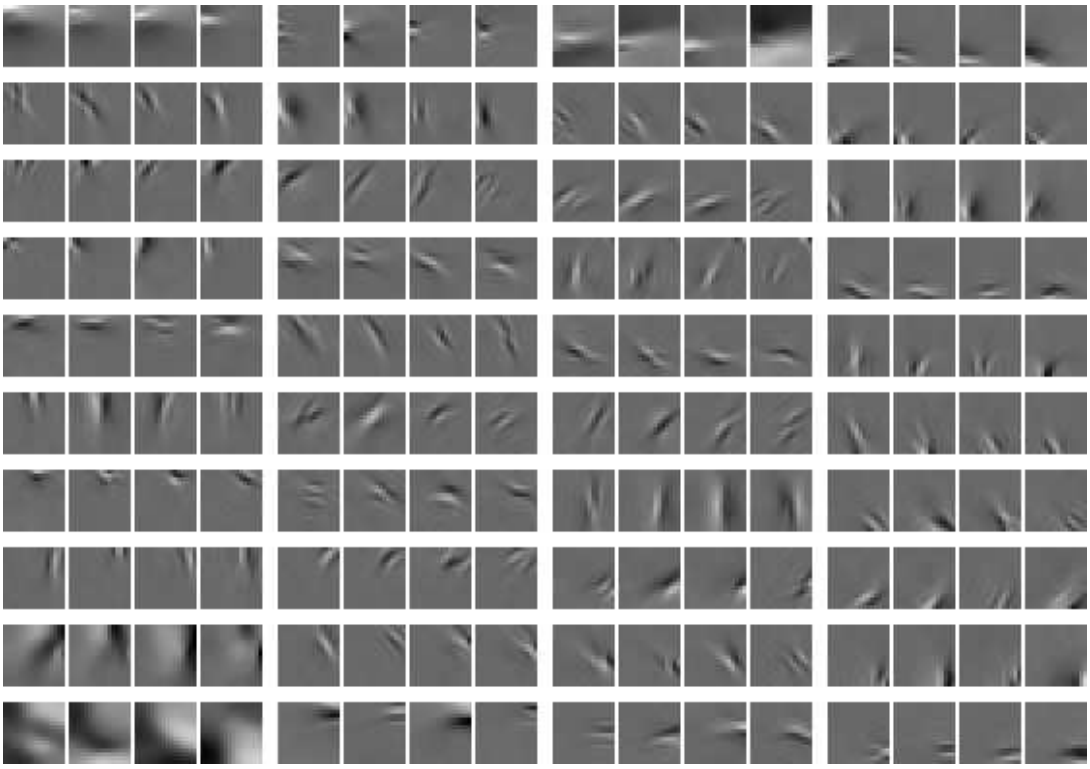

Figure 10. Basis vectors, and their grouping into 4-D subspaces, estimated by independent subspace analysis.

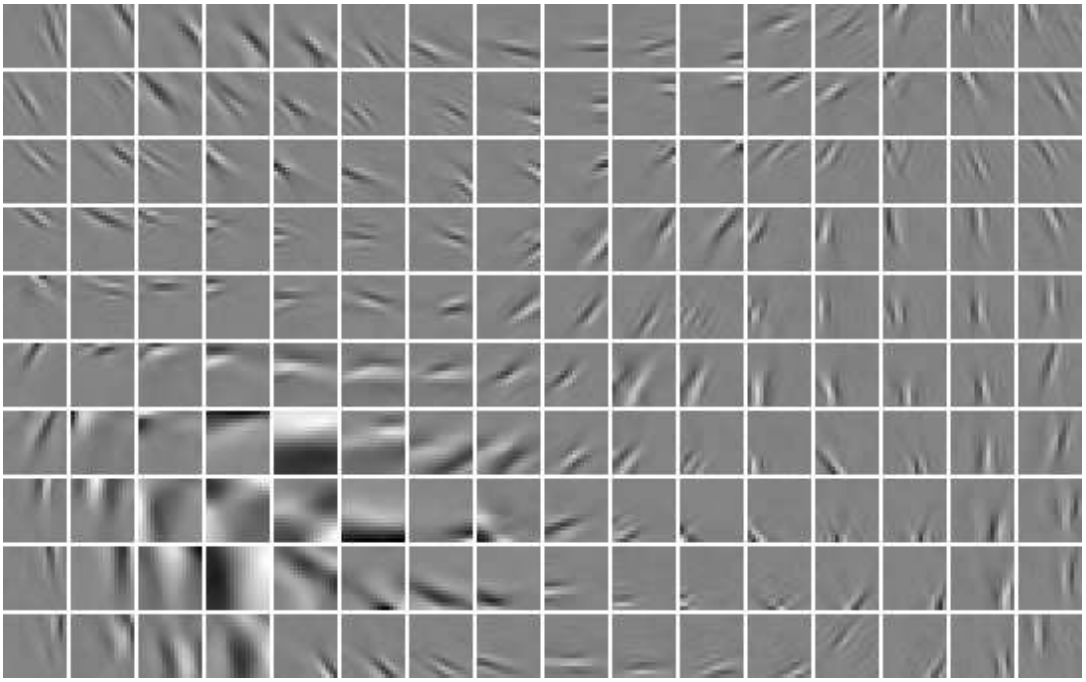

Figure 11. Basis vectors, and their topographic organization, estimated by topographic ICA.

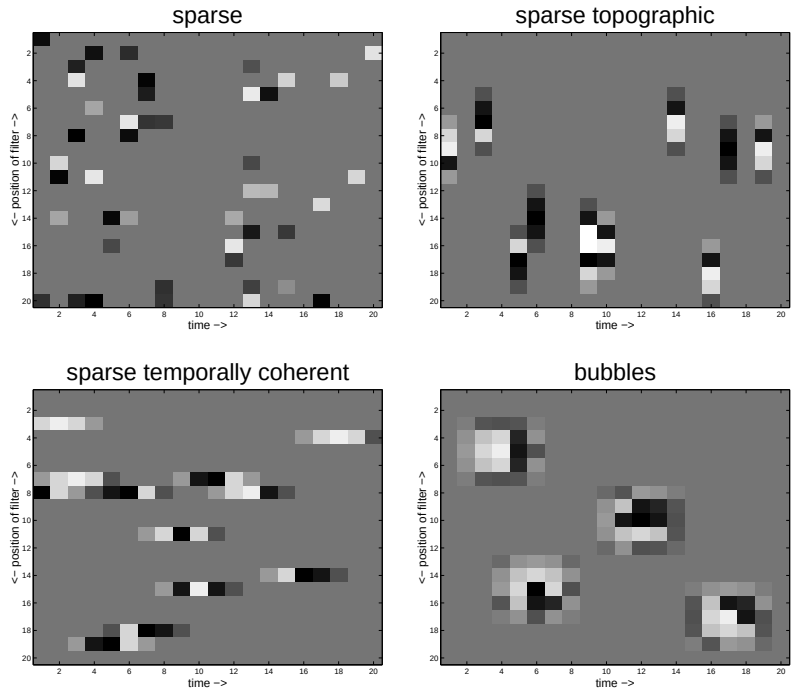

Figure 12. The four types of representation. The plots show the outputs of filters as a function of time and the position of the filter on the topographic grid. Each pixel is the activity of one unit at a give time point, gray being zero, white and black meaning positive and negative outputs. For simplicity, the topography is here one-dimensional. In the basic sparse (ICA) representation, the filters are independent. In the topographic representation, the activations of the filters are also spatially grouped. In the representation that has temporal coherence, they are temporally grouped. The bubble representation combines all these aspects, leading to spatiotemporal activity bubbles. Note that the two latter types of representation require that the data has a temporal structure, unlike basic sparse coding or ICA.

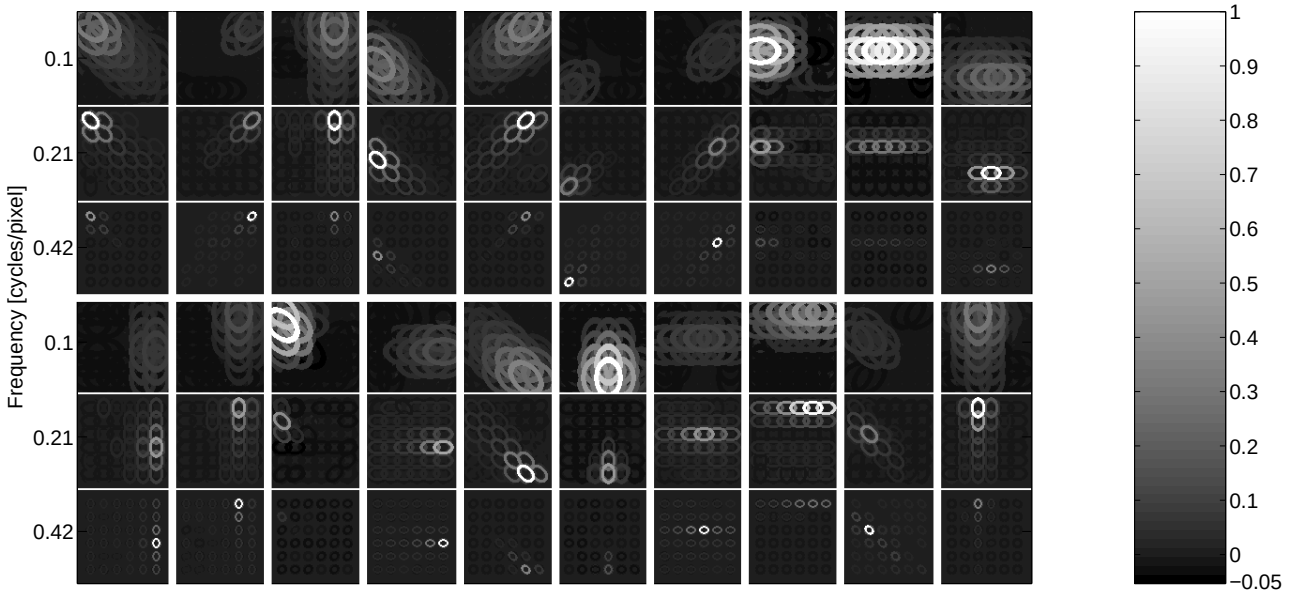

Figure 13. A random selection of higher-order basis vectors estimated from natural images in (Hyvärinen et al., 2005). Each display of three patches gives the coefficients of one higher-order feature. Each patch gives the coefficients of one higher-order feature in one frequency band. Each ellipse means that the complex cell in the corresponding location, and of the corresponding orientation and frequency is present in the higher-order feature, brightness of ellipse is proportional to coefficient.