Bubbles: a unifying framework for low-level statistical properties of natural image sequences

Aapo Hyvärinen, Jarmo Hurri, and Jaakko Väyrynen

Neural Networks Research Centre, Helsinki University of Technology, P.O. Box 9800, FIN-02015 HUT, Finland

Received September 26, 2002; revised manuscript received February 10, 2003; accepted February 21, 2003.

Recently, different models of the statistical structure of natural images have been proposed. These models predict properties of biological visual systems and can be used as priors in Bayesian inference. The fundamental model is independent component analysis, which can be estimated by maximization of the sparsenesses of linear filter outputs. This leads to the emergence of principal simple cell properties. Alternatively, simple cell properties are obtained by maximizing the temporal coherence in natural image sequences. Taking account of the basic dependencies of linear filter outputs permit modeling of complex cells and topographic organization as well. We propose a unifying framework for these statistical properties, based on the concept of spatiotemporal activity "bubbles." A bubble means here an activation of simple cells (linear filters) that is contiguous both in space (the cortical surface) and in time.

OCIS codes: 330.3790, 330.4060, 330.4270.

1. INTRODUCTION

A widespread assumption is that the visual cortex is adapted to process the particular kind of information it receives.[1,2] The visual cortex is important for survival and reproduction, and evolutionary forces thus drive the visual system toward 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.

One property that distinguishes natural images from other kinds of input is statistical structure. The gray-scale values of luminances at different retinal points, for example, have robust and nontrivial statistical regularities. Previous research has built statistical models of natural images and utilized them to model the receptive fields, the spatial organization, and the interaction of neurons in the visual cortex.[3–5] Such models can also be used as priors in Bayesian inference.[6–9]

This paper proposes a unifying framework for several models of the statistical structure of natural image sequences. The framework combines three properties: sparseness, temporal coherence, and energy correlations; these will be reviewed below. It leads to models where the joint activation of the linear filters (simple cells) takes the form of "bubbles," which are regions of activity that are localized both in time and in space, space meaning the cortical surface or a grid on which the filters are arranged.

The paper is organized as follows. First, we discuss the principal statistical properties of natural images investigated so far, and we examine how these can be used in the estimation of a linear image model (Section 2). Then we show how sparseness and temporal coherence can be combined in a single model, which is based on the concept of temporal bubbles, and attempt to demonstrate that this gives a better model of the outputs of Gabor-like linear filters than either of the criteria alone (Section 3). We extend the model to include topography as well, leading to the intuitive notion of spatiotemporal bubbles (Section 4). We also discuss the extensions of the framework to spatiotemporal receptive fields (Section 5). Finally, we discuss the utility of our model and its relation to other models (Section 6).

2. BASIC STATISTICAL PROPERTIES OF NATURAL IMAGES

Here we review the research on the basic statistical properties to be included in our model. These are sparseness, temporal coherence, and correlation of energies.

A. Sparseness

Sparseness is a property of a random variable, such as the output of a linear filter when the input consists of natural images. Sparseness means that the random variable takes very small (absolute) values or 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 nonzero, only rarely. 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 ("heavy tails") far from zero. Here "relatively" means compared with a Gaussian distribution of the same variance. For example, the absolute value of a sparse random variable is often modeled as an exponential density. The exponential density is compared with the density of the absolute value of a Gaussian variable in Fig. 1. If the absolute value of a symmetric random variable has an exponential distribution, the distribution is called Laplacian. If we scale the distribution to have variance equal to 1, the density function is then given by

1084-7529/2003/070001-16$15.00

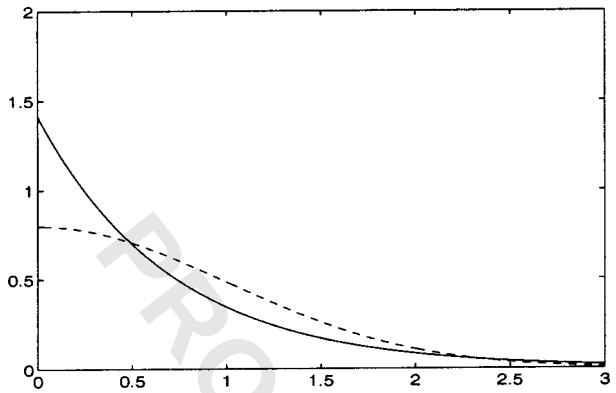

Fig. 1. Illustration of a sparse probability density. The vertical axis is the probability density, and the horizontal axis is the (absolute) value of random variable s. The sparse exponential density function is given by the solid curve. For comparison, the density of the absolute value of a Gaussian random variable of the same variance is given by the dashed curve.

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

Sparseness is not dependent on the variance (scale) of the random variable. To measure the sparseness of a random variable s, let us first normalize its scale so that the variance $E\{s^2\}$ equals 1. Sparseness can then 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. For example, if G is the square function, sparseness is measured by the fourth moment $E\{s^4\}$. This is closely related to using kurtosis,[2,5] defined as $\mathrm{kurt}(s) = E\{s^4\} - 3(E\{s^2\})^2$. 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,[10] which is why in practice other functions G may have to be used.

Both information-theoretic and estimation-theoretic considerations show that in some ways the ideal function would be such that $G(s^2)$ is equal to the logarithm of a sparse probability density function, optimally of s itself.[5] For example, taking the logarithm of the Laplacian density, one obtains

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

The constants $\alpha = \sqrt{2}$ and $\beta = -\log\sqrt{2}$ are needed in the probability density to make its integral equal to 1 and to standardize it to unit variance, but they are irrelevant when considering a sparseness measure, so one could just as well take $\alpha = 1$ and $\beta = 0$ in any practical measurement of sparseness.

The importance of sparseness lies in its ability to model the principal properties of simple cell receptive fields. Given natural image input, the outputs of linear filters that model simple cells are very sparse; in fact, they maximize typical measures of sparseness.[2,11]

B. Temporal Coherence

An alternative to sparseness is given by temporal coherence or stability.[12–16] This means that when the input consists of natural image sequences, i.e., video data, the outputs of simple cells (linear filters) 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 ordinary linear (auto)correlation or (auto)covariance is *not* able to produce plausible receptive fields. That is, if we measure the temporal coherence of a cell output $s(t)$, centered to have zero mean, as

$$\mathrm{cov}[s(t),\, s(t-\tau)] = E\{s(t)s(t-\tau)\}, \tag{3}$$

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 and nonlocalized low-frequency Fourier components.[16] Note that maximizing this measure is equivalent to minimizing the mean change $E\{[s(t) - s(t-\tau)]^2\}$ if the variance of $s(t)$ is kept constant.

We proposed previously[16] that temporal coherence could be measured by the correlation of squares (energies):

$$\mathrm{cov}\{[s(t)]^2, [s(t-\tau)]^2\} = E\{[s(t)]^2[s(t-\tau)]^2\} - E\{[s(t)]^2\}E\{[s(t-\tau)]^2\}. \tag{4}$$

This measure was inspired by recent advances in the theory of blind source separation, in which it was shown that the correlation of squares is a valid measure for blind source separation.[17] It was found that typical simple cell receptive fields maximize this criterion, just like sparseness.[16] Thus, when properly defined and measured, temporal coherence provides an interesting alternative to sparseness.

C. Correlation of Energies

1. Definition and Models

The third statistical property considers the relationships between the outputs of different linear filters (simple cells), which will be denoted by s_i, $i = 1,\ldots,n$. When sparseness or temporal coherence is used, 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 $i \neq j$. To go beyond this basic framework, we need to model the statistical dependencies of the linear filters, assuming that their joint distribution is dictated by the natural image input.[18–21]

Note that, again, 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 the estimation of simple cell receptive fields, their outputs are often constrained to be exactly uncorrelated[10,21] (see also Subsection 2.D). In image data, the principal form of dependency between two model simple cell outputs seems to be captured[18–22] by the correlation of their energies, or squares s_i^2. This means that

$$\mathrm{cov}(s_i^2,\, s_j^2) = E\{s_i^2 s_j^2\} - E\{s_i^2\}E\{s_j^2\} \neq 0. \tag{5}$$

This covariance is usually positive. Thus we see a close connection to the temporal coherence framework.

Intuitively, positive correlation of energies means that the cells tend to be active, i.e., have nonzero 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 z_i, z_j and a common "variance" variable v,[21,23] given by

$$s_i = z_i v, \tag{6}$$

$$s_j = z_j v, \tag{7}$$

then s_i and s_j are uncorrelated but their energies are not.[21] In fact, assuming that z_i and z_j have zero mean and unit variance, the covariance of their energies can be calculated to equal $E\{v^4\} - (E\{v^2\})^2$, which is positive because it equals the variance of v^2. Further, if z_i and z_j are chosen Gaussian, the resulting variables s_i and s_j can be shown to be sparse (leptokurtic).[21]

A simple density that incorporates both energy correlation and sparseness is given by[20,21]

$$p(s_i, s_j) = \frac{2}{3\pi} \exp(-\sqrt{3}\sqrt{s_i^2 + s_j^2}). \tag{8}$$

This could be considered a two-dimensional generalization of the Laplacian distribution, standardized so that its variance equals unity. The correlation of energies in this probability distribution is illustrated in Fig. 2. A generalization to more than two dimensions is straightforward by just taking the sum of the squares 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. (8) gives us a measure of the combination of 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

$$E\{G(s_i^2 + s_j^2)\}, \tag{9}$$

where $G(b) = -\sqrt{b}$, up to irrelevant constants.

2. Topographic Structure of Dependencies

The correlation of energies could be embedded in a model of natural image statistics in many ways. A simple way would be to divide the s_i into groups, so that the s_i in the *same* group have correlation of energies whereas the s_i in *different* groups are independent. In such a model,[20] it was found that the groups (called "independent subspaces") show emergence of complex cell properties. 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. Thus simple cells that pool to the same complex cell have energy correlations, whereas simple cells that are not pooled together are independent.

Here we concentrate on a more general framework of modeling the energy correlation of modeled simple cell outputs, based on topography or spatial organization of the cells. By topography, we mean here the existence of ordered maps, in which the spatial location of a cell on the cortical surface is related in a systematic way to its functional properties. In the visual cortex, the location of the receptive field, as well as selectivity for orientation, spatial frequency, and many other parameters, forms such cortical maps.[24–27] Let us assume that the s_i are arranged on a two-dimensional grid or lattice, as is typical in topographic models.[28–30] The topography is formally expressed by a neighborhood 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.

Looking at the statistical structure of natural images, we see that the pairwise dependencies can be used to define a topography.[31] This means that model simple cells can be arranged on a 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 neighboring cells. This means simultaneous activation of neighboring cells; such simultaneous activation is implicit in much of the work in cortical topography.

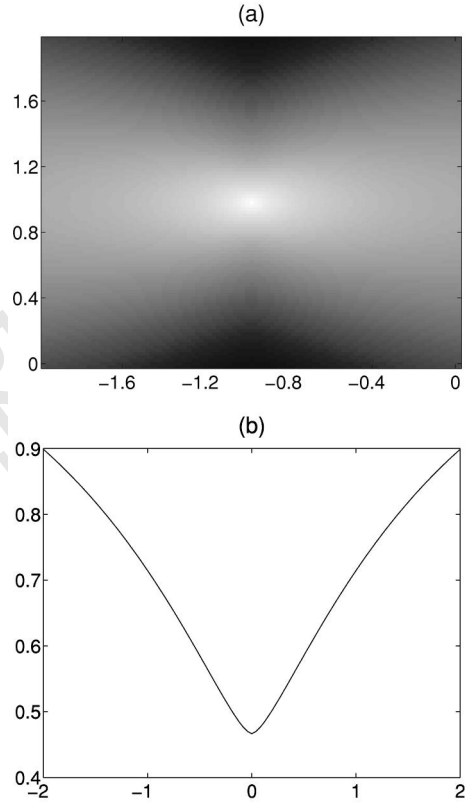

Fig. 2. Illustration of the energy correlation in the probability density in Eq. (8). (a) Two-dimensional conditional density of s_j (vertical axis) given s_i (horizontal axis). The conditional density is obtained by taking vertical slices of the density function, and then normalizing each slice so that it integrates to 1, and thus defines a proper probability density function. Black means low probability density, and white means high probability density. We see that the conditional distribution gets broader as s_i goes further from zero in either direction.[19] This leads to correlation of energies, since the expectation of the square is nothing but the variance. (b) Conditional variance of s_j (vertical axis) for given s_i (horizontal axis).

D. Linear Generative Models of Natural Images

A powerful framework for utilizing the statistical properties discussed above is provided by generative models.[4,8] 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:

$$I(x, y) = \sum_{i=1}^{n} a_i(x, y)s_i \tag{10}$$

for all x and y. The s_i are stochastic coefficients, different from patch to patch. In a cortical interpretation, the s_i model the responses of (signed) simple cells, and the a_i are closely related to their classical receptive fields. For simplicity, we consider only spatial receptive fields in most of this paper; spatiotemporal receptive fields are considered in Section 5.

For simplicity, we assume that the number of pixels equals the number of basis vectors, in which case the linear system in Eq. (10) can be inverted. Then a simple cell with index i is modeled as a spatial linear filter with adaptable weights, say w_i. The output of the simple cell, when the input is an image patch I, is given by

$$s_i = \langle w_i, I \rangle = \sum_{x,y} w_i(x, y)I(x, y). \tag{11}$$

It can be shown[31] that the a_i are basically low-pass filtered versions of the receptive fields w_i.

Estimation of the model consists in determining the values of a_i, observing a sufficient number of patches I without knowledge of the modeled simple cell outputs s_i. In the most basic models, the s_i are assumed to be statistically independent. Then we can use either sparseness or temporal coherence to estimate the receptive fields. That is, we assume either that the s_i are sparse or that they have temporal coherence.

The estimation can be simplified by suitable preprocessing. First, we consider the contrast only, i.e., the local mean or dc component has been removed from the image, which also implies that the s_i have zero mean. Second, we whiten the data in the spatial domain: The data are transformed into an image so that for any two spatial points (x, y) and (x', y') the values of $I(x, y)$ and $I(x', y')$ are uncorrelated, and all points are normalized to unit variance. In the whitened space, we can then consider orthonormal transformations only, i.e., $\Sigma a_i(x, y)a_j(x, y) = 0$ if $i \neq j$ and 1 if $i = j$. This is because the 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.[5]

If sparseness is used,[11,32,33] the temporal structure of the data is ignored; indeed, the data do not need to have any temporal structure in the first place. The resulting model is called independent component analysis (ICA),[5,34,35] and it can be considered a non-Gaussian version of factor analysis. A deep result[5] in the theory of ICA says that if the data are actually generated according to a linear generative model, as in Eq. (10), the underlying basis vectors can be recovered by finding basis vectors (or, equivalently, receptive fields) such that the sparseness of the outputs is maximized under some conditions.

In the case of temporal coherence, it is not so well established that the estimation of a generative model could be accomplished by maximizing the temporal coherence of simple cell outputs. However, using a suitable definition of temporal coherence, such as the temporal correlation of squares, one can show such a connection.[17] In that case, the sparseness structure of the data is not utilized in the estimation.

When topography is used, the s_i are not assumed to be independent anymore. Instead, they have topographic energy correlations as defined in Subsection 2.C. This leads to the topographic ICA model,[21,31] which precisely combines the properties of sparse components and topographic dependencies in a single model. When the model is estimated from natural image data,[31] the emerging topography is qualitatively very similar to the one observed in V1: There are clear maps of orientation, frequency, and retinal location and no map for phase. Also, the 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 (energies of outputs of neighborhoods).

3. TEMPORAL BUBBLES: COMBINING SPARSENESS AND TEMPORAL COHERENCE

A. Definition of the Model

As discussed above, both maximization of the sparseness of linear filter outputs and maximization of their temporal coherence lead to receptive fields that have the principal properties of simple cells. How is it possible that two quite different criteria give quite similar receptive fields? What is the connection between the two criteria?

To answer these questions, we propose a model of the linear filter outputs that combines the two properties. The model explains why both criteria give similar estimation results from natural images and can be expected to give an improved model of the statistical structure of linear filter outputs. In this section, we consider only the estimation of simple cell receptive fields; dependencies and topography will be considered in Section 4.

The new model is based on the concept of a sparse temporal activity bubble. (This will be extended to a sparse *spatio*temporal activity bubble in Section 4.) We assume that the observed linear filter output $s(t)$ is the product of an underlying latent signal $z(t)$ and a variance signal $v(t)$, much as in Eqs. (6) and (7), but in the temporal domain. Thus we define

$$s(t) = v(t)z(t). \tag{12}$$

The underlying signal $z(t)$ does not need to have any special properties. In fact, we assume here, for simplicity, that $z(t)$ is Gaussian white noise with unit variance. The interesting statistical properties of $s(t)$ are thus due to $v(t)$ alone.

The crucial assumptions are that $v(t)$ is sparse and has temporal correlation. To model such a signal, we assume

that it is a low-pass filtered (smoothed) version of a very sparse signal possibly followed by a pointwise (scalar) function:

$$v(t) = f(\phi(t) * u(t)) = f\left(\sum_{\tau} \phi(\tau) u(t - \tau) \right), \quad (13)$$

where ϕ is a simple low-pass filter, such as the Gaussian kernel $\exp[-\tau^2/(2\sigma^2)]$. The random process $u(t)$ is obtained by sampling a very sparse nonnegative random variable independently at each time point, resulting in something similar to a point process with nonnegative values. The function f is a technical addition that has little influence on the basic principle, and in most cases we could just take a linear f. However, a suitable nonlinear f enables us to get a simple approximation of the probability densities involved, as will be seen below.

The signal generation is illustrated in Fig. 3. (Note that we ignore any border effects that will occur in the convolution of finite-length signals.) The resulting signal $s(t)$ has both sparseness and temporal coherence. The sparseness can be shown as follows[21]:

$$\begin{aligned} \text{kurt}[s(t)] &= E\{[s(t)]^4\} - 3(E\{[s(t)]^2\})^2 \\ &= E\{[v(t)]^4[z(t)]^4\} - 3(E\{[v(t)]^2[z(t)]^2\})^2 \\ &= 3[E\{[v(t)]^4\} - 3(E\{[v(t)]^2\})^2], \end{aligned} \quad (14)$$

which is always positive because it is the variance of $v(t)$ multiplied by 3. The correlation of squares follows from a proof similar to the one following Eq. (7).

Thus, if one mixes linearly independent signals of this kind, the original signals can be separated by using either of these two properties.[5,17] In particular, if we consider the image sequences to be linear sums of spatial basis vectors, i.e.,

$$I(x, y, t) = \sum_{i=1}^{n} a_i(x, y) s_i(t), \quad (15)$$

and assume that the signals $s_i(t)$ consist of temporal bubbles as defined above, it is natural that we obtain similar basis vectors with either criterion, since both are applicable on this data type.

For illustration, let us look at the output of a spatial ICA/sparse coding filter when the input consists of an image sequence. The output for a randomly sampled sequence of 1000 points is shown in Fig. 4. One can clearly see bubblelike behavior.

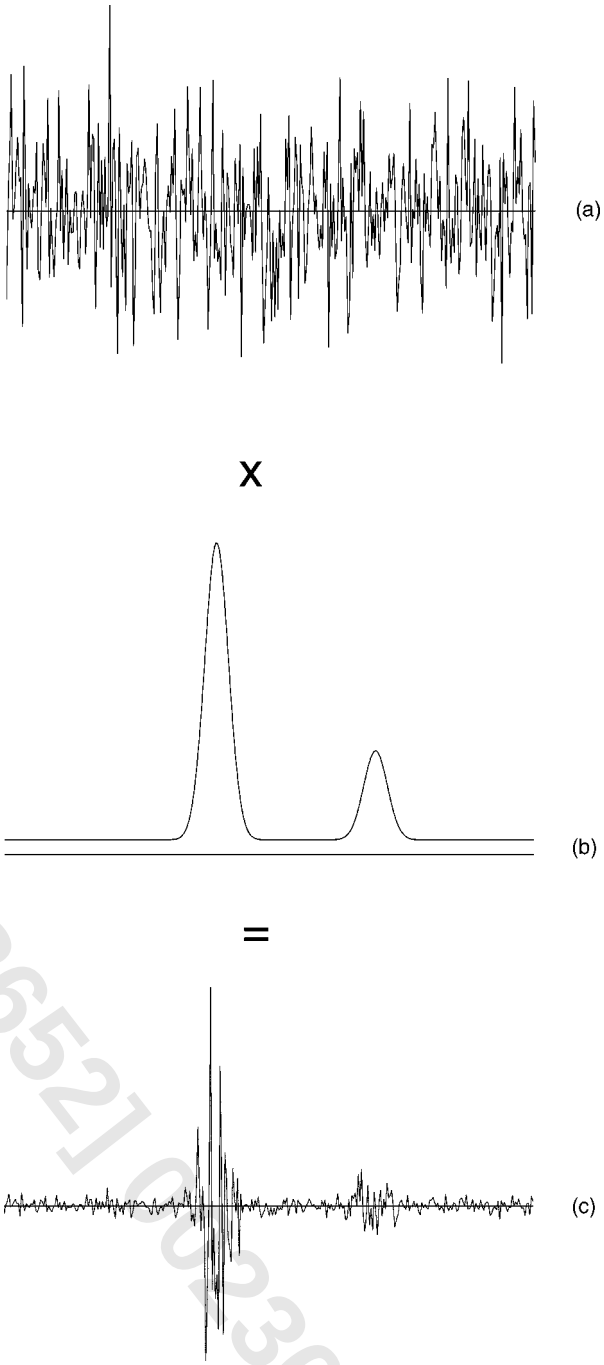

Fig. 3. Illustration of a temporal bubble. The original signal $z(t)$ (top) is multiplied by a variance (activity) signal $v(t)$ (middle) to obtain the observed signal $s(t)$ (bottom). The observed signal is both sparse and temporally coherent.

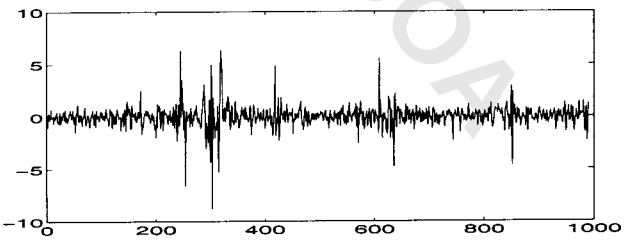

Fig. 4. Output of a filter estimated by ICA when the input consists of an image sequence. A temporal bubble structure is clearly visible. For details on the data, see Subsection 3.C.

B. Estimation of the Model

Next, we propose a computationally simple objective function for estimating optimal linear filters. Using the same derivation as that in topographic ICA[21] (see Appendix A), we can give a simple approximative formula for the probability density in the temporal bubble model:

$$\log p(s(1), \ldots, s(T)) \approx \sum_{t=0}^{T} G(b(t)), \quad (16)$$

where

$$b(t) = \phi(t) * [s(t)]^2 = \sum_{\tau} \phi(\tau)[s(t - \tau)]^2. \quad (17)$$

Here ϕ is the low-pass filter in Eq. (13). The function G depends on the probability densities of u and the nonlinearity f. For typical sparse densities, the function G is again a convex function, similar to those used above (see Appendix A for details).

Intuitively, $b(t)$ measures the sum of squares around the given time point. Since G is convex, this approximation can be interpreted as measuring the sparseness of the bubbles. This means that most of the data should have practically no activity, the variance signal $v(t)$ being almost zero. To compensate, the variance signal sometimes takes relatively large values.

Estimation of the filters and the basis vectors can be accomplished by maximizing the approximative likelihood of the linear generative model in Eq. (15), given observations of an image sequence $I(x, y, t)$ for all x and y and $t = 0,..., T$. This is given, based on the approximation in relation (16), by the sum

$$\log L(w_1,\ldots, w_n; I(x, y, t)) \approx \sum_{i=1}^{n} \sum_{t=0}^{T} G(b_i(t)), \tag{18}$$

where

$$b_i(t) = \sum_{\tau} \phi(\tau)\langle w_i, I_{t-\tau}\rangle^2 = \sum_{\tau} \phi(\tau) \times \left[\sum_{x,y} w_i(x, y) I(x, y, t - \tau)\right]^2 \tag{19}$$

and it is assumed that the filters w_i are constrained to be orthogonal.

A reasonable approximation of the sparse structure of image data is obtained[21] by defining G using the square root as in Eq. (8):

$$G(b) = -\alpha\sqrt{b} + \beta, \tag{20}$$

where $\alpha > 0$ and β are some unknown constants that are necessary for the approximation of relation (16) to be interpreted as a likelihood. The values of these constants have no effect on the maximal points of the function, however, so one can again take $\alpha = 1$, $\beta = 0$ in any optimization algorithm. To improve the stability of the optimization, it may be useful to avoid the singularity that the derivative of the square root has at zero by using a function of the form of $G(b) = -\sqrt{b + \epsilon}$, where ϵ is a small constant.

C. Experiment 1: Optimal Integration Time in Temporal Bubbles

1. Motivation

In an attempt to show that temporal bubbles characterize outputs of Gabor-like linear filters better than either sparseness or temporal coherence alone, we conducted separation experiments on natural image sequences.

In the experiments, we sought to find the optimal kernel ϕ for temporal integration to be used in the likelihood in relation (18) and Eq. (19). Note that the case of ordinary sparseness is obtained when the kernel does not integrate over time, being 1 for zero lag and 0 elsewhere. Thus, if the optimal kernel is longer than this, we have also proven that the temporal bubble model is better than plain sparseness. We considered only the space of kernels that are 1 inside a given integral and 0 outside the interval, thus reducing the problem to determining the optimal length of the integration interval.

2. Methods

We used the same data as those in previous work on natural image sequences.[16] The data consisted of natural image sequences[36] cleaned of some artifacts and less natural parts with only man-made objects.[16] A sampling rate of 25 samples per second was used.

In principle, the comparison might be accomplished by computing the values of the likelihoods for kernels ϕ of different widths. In practice, however, this is difficult, since different kernels need different normalization constants α and β, and these are most difficult to compute.

Thus we used a different approach based on signal separation. Given a kernel, we measured the modeling power of the bubble model by its efficiency in separating signals that have the relevant statistical structure.

At each trial, we computed four Gabor-like linear filters either by ICA or by maximization of temporal coherence. (The number 4 is arbitrary.) We took four image sequences of 1000 time points at random locations and computed the outputs of the four filters $s_i(t)$, $i = 1,..., 4$, with one input sequence fed to each filter (the location of the filter was fixed, so the point was to look at the image sequence "through a Gabor function"). Thus we obtained four source signals. Then we took a random orthogonal mixing matrix $\mathbf{A} = [a_{ji}]$, $i, j = 1,..., 4$, and mixed the four signals to obtain $x_j(t) = \Sigma_i a_{ji} s_i(t)$, $j = 1,..., 4$, just as in the basic linear mixing model in ICA. We then separated the signals, i.e., estimated the $s_i(t)$ from the mixtures, by finding linear combinations $s_i = \Sigma_j w_{ij} x_j$ that maximize the objective function in relation (18), where the mixtures x_j are substituted for the images sequence I by using $G(b) = \sqrt{b + \epsilon}$ with $\epsilon = 0.0001$. Maximization was performed by a gradient method using symmetric orthogonalization[5,10] of the four vectors $\mathbf{w}_i = (w_{i1}, w_{i2}, w_{i3}, w_{i4})^{\mathrm{T}}$, $i = 1,..., 4$.

After separation, we computed the error in the separation, which is a measure of how well we have modeled the signals $s_i(t)$, and compared the errors in the different cases.[5] Once we had obtained a separating matrix $\mathbf{W} = [w_{ij}]$, the error was computed as the sum of the squares of the matrix $\mathbf{WA}$, minus the four largest squares. If $\mathbf{W}$ really were the inverse of $\mathbf{A}$ (possibly up to permutation and arbitrary signs), this would equal zero.

We repeated this procedure 268 times (as much as our computer could take). At each trial, the data were sampled at different random points (half of the trials used temporally decorrelated data,[16] and half used the raw data), four different filters were randomly chosen from a set of 120 filters (half of the trials used ICA filters, and half used temporal coherence filters), and a new mixing matrix was randomly generated. Thus, for each kernel, we obtained 268 measured errors. We took the logarithms of the errors because the distribution of the log errors was close to normal, and we computed the averages. As mentioned above, we used the simplest possible kernels, ones that were 0 outside a given radius and 1 inside.

The radius ranged from 0 to 8, leading to intervals of size 1–17. Since the log errors were close to normally distributed, we computed the standard error of the mean in the usual way for each temporal kernel.

3. Results

The results are shown in Fig. 5. Temporal integration decreases the error, approximately by a factor of 2.8 (0.446 in $\log_{10}$ scale). The error for purely sparse coding (interval = 1) was very significantly larger ($p < 0.001$) than the others. Temporal integration with three time points was significantly worse ($p < 0.01$) than integration with at least seven points, although not significantly different from integration with five time points. Otherwise, the differences were not significant. Increasing the integration interval above seven time points does not yield significant improvement. Presumably, the error will start to increase when the interval becomes very long, but this was not visible in our results, probably because of the rather limited maximum length that we used.

We also tried the temporal coherence criterion[16] on the same problem. The errors (not shown) were approximately three times larger than those obtained with sparseness. Thus the bubble model also performs better than temporal coherence as a model of the statistical structure. This is not very surprising, since the temporal coherence criterion[16] was constructed so as to be blind to sparseness in order to demonstrate that temporal coherence alone is sufficient. Moreover, the temporal coherence criterion was based on simple nonlinear correlation and not on statistically optimal criteria such as likelihood. In fact, if kurtosis is used as the sparseness measure, the increase in estimation errors when compared with those of optimal maximum-likelihood estimation is quite comparable.[10]

Thus temporal integration in the temporal bubble model does reduce separation errors, indicating that the model is better than basic sparse coding or ICA. An interval of 5–7 time points seems to be sufficient. As the time interval between two samples was 40 ms, this corresponds to an optimal temporal integration interval of 200–300 ms.

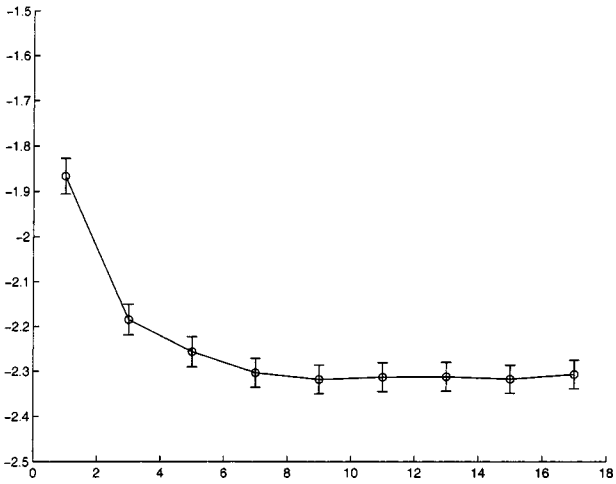

Fig. 5. Log error in separation of artificial signal mixtures as a function of the size of the interval of temporal integration. This reaches a minimum at approximately 7. Size 1 would correspond to ordinary sparseness, i.e., no temporal integration. Standard errors of the mean are shown as well.

4. SPATIOTEMPORAL BUBBLES: A UNIFYING FRAMEWORK

In *spatio*temporal bubbles, the idea is to combine all *three* properties discussed above: sparseness, temporal coherence, and topography. Combination of sparseness and temporal coherence was done in Section 3 and was shown to lead to temporal activity bubbles.

Combination of sparseness and topography means that each input activates a limited number of spatially limited "blobs" on the topographic grid.[31] If these regions are 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. This is illustrated in Fig. 6 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 a temporally persistent activation of Gabor-like basis vectors of similar orientation and spatial frequency in nearby spatial locations. This would correspond to a short contour element of given orientation and spatial frequency. In contrast to a spatial "independent component" of an image, this contour element can move a bit, and its phase can change, during the temporal extent of the bubble, because the bubble determines only the general activity level while the actual values of outputs randomly change from one time point to another. A bubble thus gives a more flexible, invariant feature than a coefficient in a linear representation.

A. Definition of the Model

Based on earlier work[31] and the developments in Section 3, we can formulate generative models based on activity bubbles. We postulate a higher-order random process u that determines the variance at each point. This nonnegative, highly sparse random process obtains independent values at each point in time and space (with 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 followed by a pointwise nonlinearity:

$$v_i(t) = f\left(\sum_j h(i,j)[\, \phi(t) * u_j(t)] \right), \tag{21}$$

where $h(i,j)$ is the neighborhood 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 $z_i(t)$ by this variance signal:

$$s_i(t) = v_i(t) z_i(t). \tag{22}$$

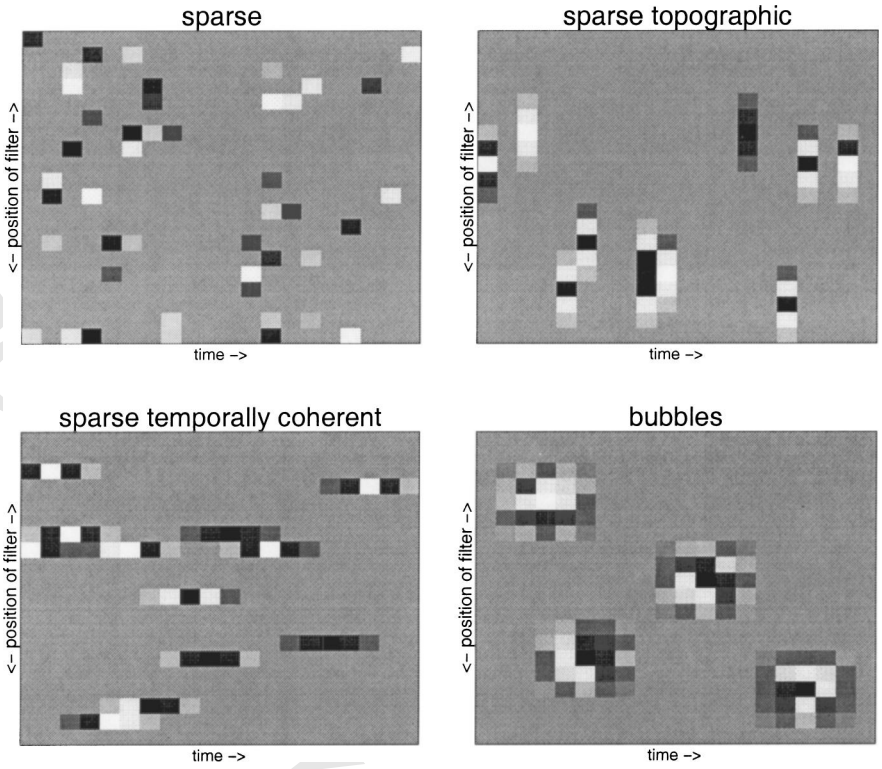

Fig. 6. Four types of representation. The plots show the outputs of filters as a function of time (horizontal axis) and the position of the filter on the topographic grid (vertical axis). Each pixel is the output of one unit at a given time point, gray being zero, white and black meaning positive and negative outputs. For simplicity, the topography is here one dimensional. In the basic sparse 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 have a temporal structure, unlike the two former ones.

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

$$I(x, y, t) = \sum_{i=1}^{n} a_i(x, y)s_i(t). \tag{23}$$

The three equations (21)–(23) define a statistical generative model for natural image sequences.

The combination of temporal and spatial energy correlation is illustrated in Fig. 7. The two signals in the figure are uncorrelated, and also have no temporal correlation, but the temporal dependence of activation is clear. Since the active intervals coincide, this is a prototype of what the dependency between two adjacent cells would look like.

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 nonzero value spreads to neighboring temporal and spatial locations because of the smoothing by ϕ and h. The spread of activation means that simple cells have large variances inside that spatiotemporal window.

The combination of temporal and spatial energy correlation also explains why sparse coding and temporal coherence give similar results when estimating complex cell receptive fields. For s_i and s_j close to each other in the topography, the joint distribution over time would be sparse and have strong energy correlation. If one estimates independent subspaces,[20] such components would fall in the same subspace. On the other hand, those same components together would form a subspace or an energy detector whose output has maximum correlation over time.[13,14] Thus independent subspaces and temporally stable subspaces coincide, which is why the two methods give similar results for natural image data. In

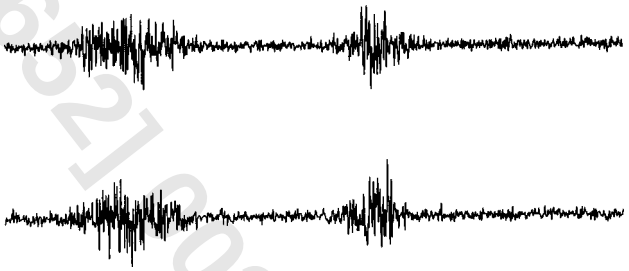

Fig. 7. Combination of temporal and spatial (i.e., topographic) energy correlation. The two signals are caricatures of what the outputs of two simple cells with strong energy correlation could look like. They are uncorrelated, both from each other and temporally. Nevertheless, we see temporal bubbles of activity in the outputs, and these bubbles are simultaneous, which eventually leads to spatiotemporal bubbles when there are many cells arranged topographically. Note that a very similar figure was used to illustrate basic energy correlation in topographic ICA.[21] In that context, the temporal energy correlation was added for the purposes of illustration only, whereas here it is an essential part of the model.

the same way, topographic ICA and its temporal coherence counterpart give similar results. This temporal coherence counterpart is a model[37] in which topography is defined solely by using temporal coherence of the local squared activations.

It should be emphasized that the model defined here is quite different from a basic ICA model using spatiotemporal basis vectors (see Section 5). The transformation from the bubble process $u_i(t)$ to the observed data is *not* linear even if the function f is linear; in fact, it is not even a deterministic transformation. The bubble process, after being convolved in time by ϕ and in space by h, gives only the variances of the linear components $s_i(t)$. Thus the basis vector a_i is added with random amplitudes and completely random signs in subsequent time points "inside" a bubble because of the interference of the variable $z_i(t)$. This is in contrast to ICA using spatiotemporal basis vectors,[36,38] where the generation of the data is linear and a single coefficient is used for the whole spatiotemporal basis vector. Therefore our model is *not* similar to ICA using $a_i(x, y)\phi(t)$ as spatiotemporal basis vectors. In fact, a spatiotemporal bubble is more similar to the activity of a complex cell with a space–time-separable receptive field,[39] although it is still quite different from conventional models of complex cells[40] because of the random generation of $z_i(t)$ at every cell and time point.

B. Estimation of the Model

In the estimation of the spatiotemporal bubble model, we can use the same ideas as those in the case of temporal bubbles and topographic ICA,[21,31] in particular the approximation of the density function in relation (16).

Here the approximation has an interesting neurophysiological interpretation. We assume that the simple cell outputs are rectified by taking squares (energies), and these are fed to complex cells. The pooling weights from simple cells to complex cells are fixed by using the assumption that complex cells pool only outputs of simple cells that are nearby on the topographic grid. The outputs of complex cells are sums of squares inside a small spatial region ("neighborhood"). While the above is similar to topographic ICA, the bubble model also pools over time. Thus we define the output of a "bubble detector" at grid point i and time point t as

$$b_i(t) = \sum_{\tau} \sum_{j=1}^{n} h(i, j)\phi(\tau)\langle w_j, I_{t-\tau}\rangle^2. \tag{24}$$

This is basically like a complex cell (energy detector) whose output is pooled over time. The output can be considered a simple, though quite crude, estimator of the variance process $u_i(t)$.

We can now approximate the likelihood of our model using the outputs of such feature detectors as

$$\log L(w_1, \ldots, w_n; I(x, y, t)) \approx \sum_{t=0}^{T} \sum_{i=1}^{n} G(b_i(t)). \tag{25}$$

The bubble pooling given by $h(i, j)$ and ϕ 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 typically convex to enforce sparseness of bubbles, as was discussed above.

In practice, it is not really necessary to compute the value of the bubble detector for all values of t. For computational convenience, one would rather sample spatiotemporal patches $I_k(x, y, t)$ from the image sequence and compute only one output of the bubble detector for each spatiotemporal patch. If we denote the temporal extent of the patch by $0, \ldots, T$, this approach gives

$$b_{ik} = \sum_{j=1}^{n} h(i, j) \sum_{t=0}^{T} \phi(T/2 - t)\langle w_j, I_{kt}\rangle^2, \tag{26}$$

where $\langle w_j, I_{kt}\rangle = \Sigma_{x,y} w_j(x, y), I_k(x, y, t)$. One would then use b_{ik} instead of $b_i(t)$ in relation (25). Note that since the sampling of a spatiotemporal patch automatically introduces limits for temporal integration, one can also define ϕ to be identical to 1 in this case, which is assumed in the following.

Learning the representation can then be accomplished by a gradient ascent of the approximative likelihood. This gives the following algorithm:

1. Whiten the image sequence spatially.
2. Sample K spatiotemporal patches $I_k(x, y, t)$ of length $T + 1$ for $k = 1, \ldots, K$.
3. Do a gradient step for all i and x, y:

$$w_j(x, y) \leftarrow w_j(x, y) + \mu \sum_{k=1}^{K} \sum_{t=0}^{T} I_k(x, y, t) \times \sum_{i=1}^{n} h(i, j)\langle w_j, I_{kt}\rangle g(b_{ik}),$$

where g is the derivative of G and μ is a small step size.

4. Orthogonalize the matrix $\mathbf{W}$, whose columns are the filters w_j:

$$\mathbf{W} \leftarrow (\mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1/2}\mathbf{W}.$$

5. Go back to step 3 if not converged.

As the nonlinearity g, we typically use $-(b + \epsilon)^{-1/2}$, where ϵ is a small constant added for stability. This corresponds to a stabilized version of $G(b) = -\sqrt{b}$ (up to an irrelevant scaling constant 1/2). Additional preprocessing, such as removal of the dc component, temporal decorrelation, or normalization, may also be useful.[16]

C. Experiment 2: Spatiotemporal Bubbles in Natural Image Sequences

In this experiment, we applied the spatiotemporal bubble model on natural image sequences.

1. Methods

Data were obtained from the same database as that in experiment 1. We took 70,000 spatial patches of 16×16 gray-scale pixels at five consecutive time points. The two-dimensional topography was defined as a 14×14 rectangular grid. The dimension of the data was accordingly reduced by principal component analysis (PCA) to $196 = 14 \times 14$ dimensions to reduce noise and aliasing artifacts.[33] The neighborhood function h was defined as being 1 inside a spatial window of 3×3 units and 0 outside that window. The topography was toroidal; i.e., the

borders of the topographic rectangle were joined together to avoid border effects.[21] The temporal kernel ϕ was defined as equal to 1 inside the sampling window and 0 elsewhere; i.e., the temporal integration interval was of length 5.

As preprocessing, we first temporally decorrelated the data as in previous work[16] and then removed the dc component of each patch.

We estimated the basis by maximization of the objective function in relation (25), as explained at the end of Subsection 4.B. The nonlinearity $G(b) = \sqrt{b + \epsilon}$ with $\epsilon = 0.00001$ was used. Optimization was performed by gradient ascent with symmetric orthogonalization.[5,10]

2. Results and Discussion

The topographic basis estimated is shown in Fig. 8. The basis is quite similar to the one estimated by the topographic ICA model.[21,31] The basis consists of Gabor-like patches, not unlike those obtained by maximization of sparseness or temporal coherence. The topography is also similar to the one obtained by topographic ICA: The orientation and the location of the feature within the patch change smoothly when moving on the topographic grid. Low-frequency patches are spatially segregated from the rest.

Thus an activity bubble in this basis consists of the temporally coherent activation of basis patches (simple cells) that have similar orientation and rather similar location. This corresponds to a basic element of visual input: a luminance contour that is of a given orientation and frequency and is inside a small spatiotemporal patch, possibly slightly moving.

The basis looks quite similar to the one obtained by topographic ICA, but this is not disappointing, since it means only that the estimation of the basis was already quite well accomplished by topographic ICA. To see the new contribution of our model, we need to consider the spatiotemporal properties of the representation. To facilitate the visualization of the results, we use in the following a one-dimensional basis of 81 units and a neighborhood range of five units. The basis is shown in Fig. 9(a).

We attempted to reproduce Fig. 6 (lower right), which was an artificial illustration of "bubbleness," but this time by using real natural image sequence data. Figure 9(b) shows the outputs of the units for two different short image sequences [the sequences are shown in Fig. 9(c)]. The outputs are thresholded for illustration. Here we can see that the data indeed produce spatiotemporal activity bubbles, that is, clusters of activity that are both

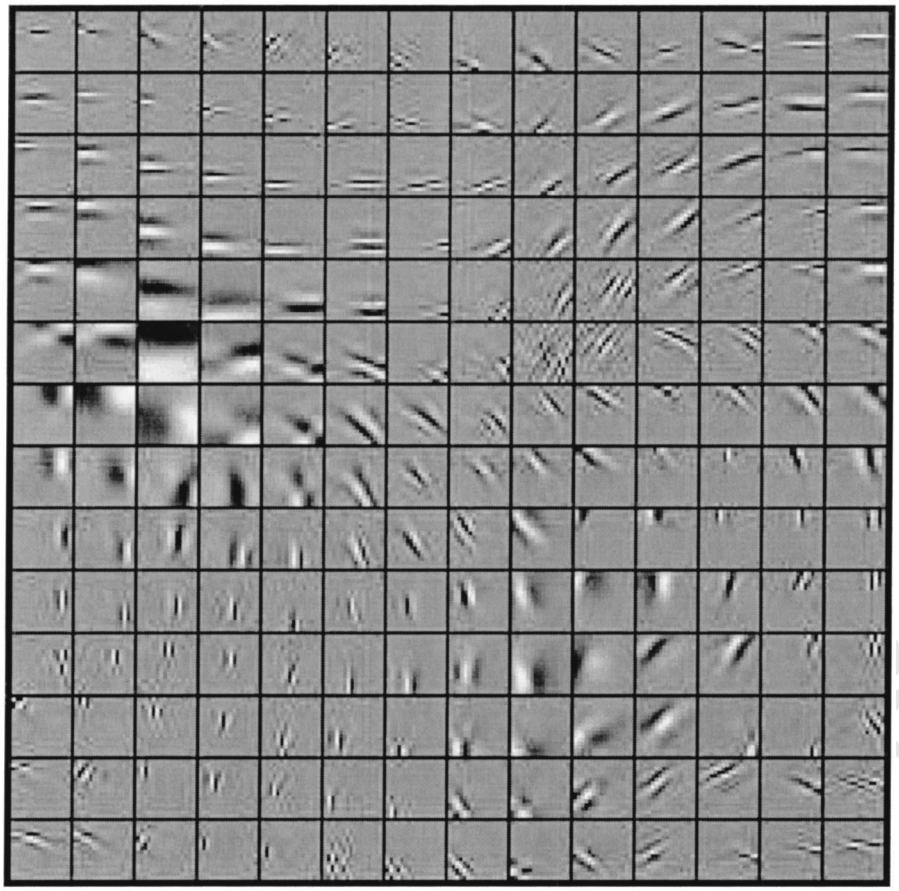

Fig. 8. Spatial basis vectors estimated by our model from natural image sequences. The results are very similar to what was found by topographic ICA.[31] Note to reviewers: The conversion of this figure to pdf for electronic submission may have produced unpleasant artifacts (shadow lines etc.) In that case, please see the original PostScript figure at: http://www.cis.hut.fi/aapo/papers/bubblebasis.eps

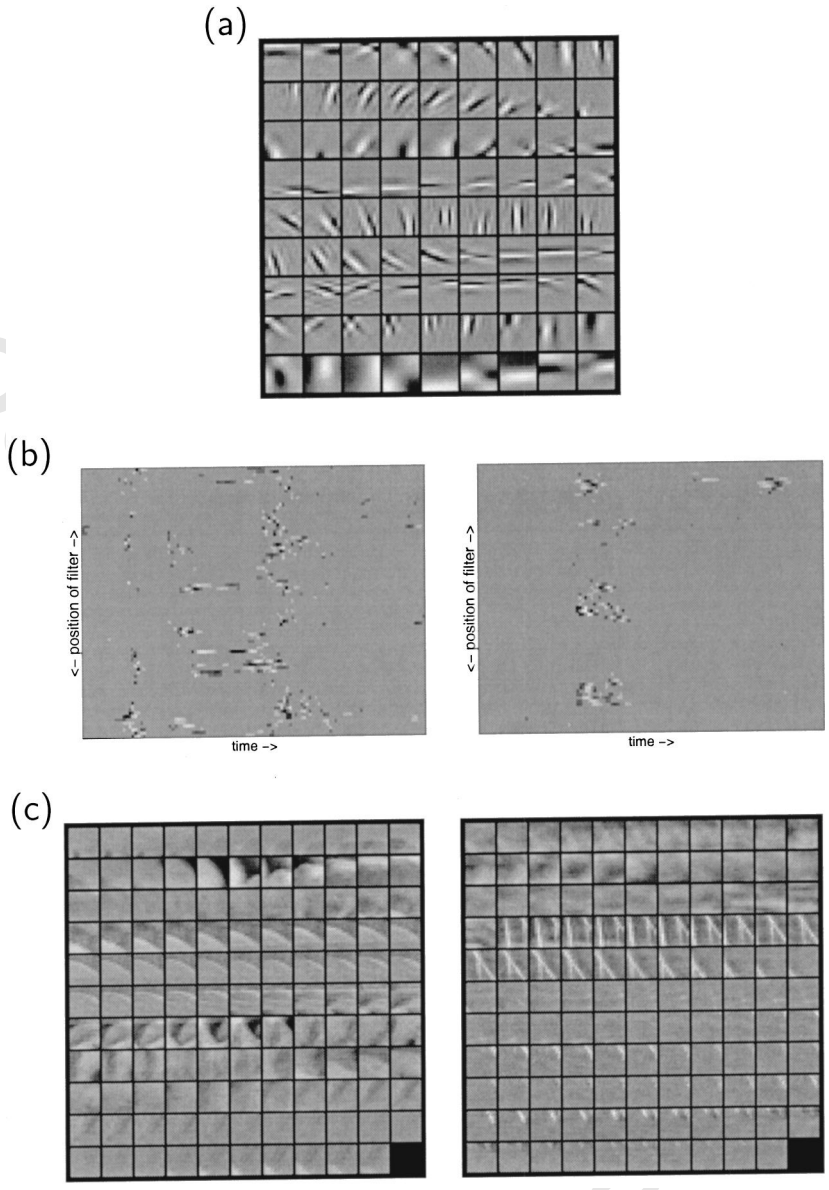

Fig. 9. Bubbles that emerge from image sequences. (a) We used a representation with one-dimensional topography to be able to visualize the results (shown arranged on a two-dimensional grid for reasons of space). (b) Outputs of the cells for two different input image sequences, coded as gray-scale values (gray = 0). The vertical axis is the cell index, and the horizontal axis is the time index. One can clearly see the bubblelike quality of the data. (c) Image sequences used as input in (b) (again, shown in two dimensions for reasons of space).

spatially and temporally contiguous. Thus we see emergence of spatiotemporal bubbles in the topographic representation, as postulated in our model.

5. EXTENSION TO SPATIOTEMPORAL RECEPTIVE FIELDS

A. Extension of the Model

In the preceding sections, we considered only spatial basis vectors and filters. This was for purposes of simplicity, both conceptual and computational. The principle of bubbles can be directly used in the case of spatiotemporal receptive fields, however.

It is important to note that bubbleness is a property of the representational units, such as linear filters or neurons (see Fig. 6). It does not depend on how those representational units compute their outputs, i.e., whether they are based on spatial or spatiotemporal filters. In fact, the same applies also for the principle of temporal coherence.[16]

Thus, to model spatiotemporal receptive fields, we do not need to change the model of the simple cell responses s_i. We need only to change the way the input data are generated from the s_i in the generative model, as in previous extensions of ICA to the spatiotemporal domain.[36,38] This is accomplished by replacing Eq. (23) by

$$I(x, y, t) = \sum_{i=1}^{n} \sum_{\tau} a_i(x, y, \tau) s_i(t - \tau). \qquad (27)$$

Thus the basis vectors are spatiotemporal. In other

words, they are basis vectors in the space of spatiotemporal patches.

While the framework using spatiotemporal bubbles and spatial basis vectors already modeled some of the spatiotemporal properties of the data, using spatiotemporal basis vectors makes the model richer. In particular, spatiotemporal basis vectors are required to properly represent motion. A spatiotemporal basis vector that is space–time inseparable[39] is able to represent a contour element moving in a specific direction. Such direction selectivity cannot be accomplished by basis vectors that have only spatial extent.

B. Experiment 3: Bubbles with Spatiotemporal Basis Vectors

1. Methods

We used the same database as that in the previous experiments, preprocessed by temporal decorrelation. The input consisted of 40,000 spatiotemporal patches whose spatial extent was 11×11 pixels and contained eight consecutive time points (sampling was constrained to consist of couples of temporally consecutive patches, as explained below). The dimension of the data was reduced to $289 = 17 \times 17$ by PCA. The neighborhood function on the 17×17 grid was 1 inside a spatial square of 3×3 units. The removal of the spatial dc component was here replaced by the removal of the mean dc component over the spatiotemporal patch.

In principle, the outputs $s_i(t)$ of the spatiotemporal filters would be computed by inverting Eq. (27). In practice, to reduce the computational load, we sample spatiotemporal patches from images, so that at each spatiotemporal sampling point, we take m temporally consecutive spatiotemporal patches. We take $m = 2$ to further reduce the computational load. Let us denote the temporal length of a single patch by T, the sample index of each temporally consecutive couple of spatiotemporal patches by $k = 1,...,K$, and pixel values in the couple of patches by $I_k(x, y, \tau)$, where $\tau = 1,...,2T$. We computed the two outputs of a spatiotemporal filter given by

$$s_i(k, 1) = \sum_{x,y} \sum_{\tau=1}^{T} W_i(x, y, \tau) I_k(x, y, \tau), \tag{28}$$

$$s_i(k, 2) = \sum_{x,y} \sum_{\tau=1}^{T} w_i(x, y, \tau) I_k(x, y, T + \tau), \tag{29}$$

where the indices 1 and 2 distinguish the first and second patches in each couple of consecutive samples. Thus the objective function had the form

$$\log L(w_1,..., w_n; I(x, y,)) = \sum_{k=1}^{K} \sum_{i=1}^{n} G\left(\sum_{j=1}^{n} h(i, j)\{[s_j(k, 1)]^2 + [s_j(k, 2)^2]\} \right). \tag{30}$$

2. Results

The results are shown as animation at the web address http://www.cis.hut.fi/aapo/papers/bubbleanimation.gif.

These spatiotemporal basis vectors are quite similar to those obtained previously.[36,38] They are Gabor-like basis vectors, most of them temporally modulated. Approximately half of them are space–time separable, and half of them are inseparable.[39] The separable ones are often constant over time, some showing weak temporal modulation. The topography is mainly organized on the axis of separable–inseparable, the frequency (spatial and temporal frequencies tend to be strongly related), and the direction of motion. A topographic organization by direction of motion has been observed in the cortex as well.[41]

6. DISCUSSION

A. Benefits of Bubble Coding

Here we propose some reasons why the visual system should use the model based on spatiotemporal bubbles. Some of these are admittedly speculative, and further research is necessary to investigate their validity.

Better prior model. First and foremost, we expect the bubble model to provide a better internal model of the structure of natural stimuli compared with that gained from previous work. If we consider visual processing in a Bayesian framework,[9] it is paramount to use statistical models of the input that are as accurate as possible. It is difficult to conclusively demonstrate that the bubble model is better than previous ones, and we have not been able to do so in this paper. Nevertheless, some indications that this claim might be true have been provided in this paper: the better separation capability of the temporal bubble model in experiment 1 and the basic fact that the model combines, for the first time, the three properties of natural images.

Better denoising by bubble thresholding. One useful application of such an internal model can be found in noise reduction. Noise in cell outputs can be reduced by using the Bayesian framework. If the ICA model is used, this leads to coring and shrinkage methods,[6,7,42] in which, basically, the outputs of the linear filter outputs are thresholded. More sophisticated models of the statistics lead to more sophisticated noise reduction methods.[43] In topographic ICA and related models,[21,23,43] information on the activation of the neighboring cells would be used in the noise reduction, so that if the neighboring cells are active, the denoising threshold is decreased. In other words, it is the output of a complex cell that is thresholded instead of the output of a simple cell; a simple cell output is then either set to zero or left unchanged, depending on whether the relevant complex cell outputs exceeded the threshold. Our bubble model further brings the temporal aspect to such a scheme: What is thresholded is essentially a temporally integrated output of a neighborhood, that is, the magnitude of a bubble.

Better rate coding. Temporal coherence can be motivated by rate coding: If the output of a linear filter is coded by the firing rate of a simple cell (or possibly two cells tuned for opposite polarities), the output of the linear filter must have some temporal coherence.[16] If there were no temporal coherence and the output of the linear filter changed very rapidly, the "readout" of the firing rate would be impossible: The Poisson noise that is inherent in such an operation would be too strong. Rate coding

using bubbles is particularly efficient in this respect because there can be both temporal and spatial (population) pooling of firing rates, as nearby cells are coding for more or less the same thing.

Minimum wiring length. Bubble coding is also related to minimum wiring length.[44] Responses of cells that are strongly dependent often need to be considered together in further computations; this is the case for the noise reduction operation discussed above, as well as computation of invariant features discussed below, and many other operations such as contrast gain control.[45,46] The topographic spatiotemporal organization makes response pooling computationally faster and simpler, since the pooling area in many operations is directly given by the topographic structure: Cells whose outputs need to be pooled together are close to each other on the cortical surface. This reduces the need for wiring (neural connections). In contrast to the case of topographic ICA, we assume in the bubble model that not only cells that fire simultaneously, but also cells whose firings are correlated over time, need to send signals to each other. This means that if cell A is a good predictor of the firing of cell B, the information about firing cell A should be sent to cell B.

More invariant features. The pooled activations of bubbles may be more interesting for higher visual processing than the activations of single cells. In fact, temporal coherence has earlier been proposed as a principle for learning invariant features,[12,14] and topographic ICA leads to emergence of features that are invariant to phase, being very similar to complex cell responses. Thus the bubbles are strongly connected to low-level invariant features. The main difference between the bubble model and conventional models of complex cells is the temporal pooling inherent in the bubbles. Thus bubbles can be considered an improved version of complex cell outputs: Invariance is enhanced by integrating the outputs over time.

B. Related Models

The concept of temporal bubbles is similar to those found in the literature on blind source separation. The technique of blind source separation by nonstationary variance[47] assumes a smoothly changing variance signal.[48] This formulation can be shown to be related to energy correlation.[17] The econometric models based on autoregressive conditional heteroscedasticity are also closely related.[49]

The concept of purely spatial bubbles, on the other hand, is closely related not only to our previous models on complex cells and topography[20,21,31] but also to models related to gain control[22] and earlier models on complex cells and energy correlation.[18]

Olshausen[38] proposed that single spikes could signal the onset of a spatiotemporal basis vector. The model included sparsification so that a single edge element moving across the image patch could be coded by the single spike. This seems to be in stark contrast to our model, in which an edge element would elicit a spatiotemporal bubble in the cell population. The contradiction can be solved, at least partially, by noting that the sparsification process includes nonlinear interactions between the cells after they have performed the initial linear filtering. Thus it is conceivable that such a sparsification process could reduce the number of active cells in a bubble, as well as the number of spikes that they fire. In the extreme case, the number of active cells could be reduced to 1, and so would the number of spikes. Thus we would obtain an estimate of the original bubble process $u_i(t)$. Olshausen's model can thus be conceived as a possible nonlinear extension of our linear filtering model.

An alternative approach to the formulation of the bubble model could be based on the autoregressive model instead of the moving-average model that we used here. We have previously used a related autoregressive model[37] because it offers some technical advantages. In particular, it is possible to estimate the pooling weights $h(i, j)$ as well. However, the moving-average formulation used here seems to be more realistic and also easier to interpret.

C. Extensions and Limitations

Our contribution is to combine the three properties of linear filter outputs that are well-known in the natural image statistics literature: sparseness, correlation of energies, and temporal coherence. To accomplish this in the simplest manner possible, we have neglected some important properties considered by other models; some of these omissions will be discussed next.

Using an overcomplete basis is often considered necessary in image modeling.[50–52] That is, there should be more basis vectors a_i than there are pixels, and, correspondingly, more components s_i. We have not included this property in our model mainly because it increases the computational complexity quite considerably.[51,52] Moreover, it seems that using an overcomplete basis does not change the properties of the individual basis vectors too much, so the qualitative properties of the results are unlikely to be affected. Yet, in any practical application of the model, overcomplete bases may be necessary. A simple way of making the obtained basis overcomplete is cycle spinning, i.e., including all the possible translations of the basis vectors in a larger basis set.

Recently, ICA and sparse coding models have been extended to multiscale representations (wavelets),[53] which also allows the modeling of whole images instead of small patches. Such an extension could be possible with our model as well. In particular, our model could be modified to incorporate multiresolution properties on the level of the bubbles as well, i.e., bubbles of different sizes. Currently, the size of the bubbles, as given by the function h, is fixed.

Another property that we have not included in our model is nonnegativity.[54,55] This might be important, even on the low level of simple cells, if one wanted to add more biological realism to the model[16,56] by considering some basic nonlinear properties of simple cells. However, as long as simple cells are modeled as linear filters, this may not be important.

The extension of our model to further modalities such as color and stereopsis may be possible by just adding the relevant properties to the data, as has been successfully done in previous work.[57–59]

7. CONCLUSION

We have proposed a new framework for the low-level statistical structure of natural image sequences, based on the notion of spatiotemporal activity bubbles. This is a unifying theoretical framework that combines the properties of sparseness (the bubbles being sparse), topography (the bubbles having spatial continuity), and temporal coherence (the bubbles having temporal continuity).

APPENDIX A: DERIVATION OF RELATION (16)

The joint density of $s(t)$ and $u(t)$ can be expressed as

$$p(s(t),\ u(t);\ t\ =\ 1,...,\ T)\ =\ \prod_t\ p_s\left(\frac{s(t)}{v(t)}\right)\frac{p_u(u(t))}{v(v)}, \tag{A1}$$

where p_u is the marginal density of $u(t)$ and p_s is the conditional marginal density of $s(t)$ given all the $u(t)$, for variance fixed to unity. The marginal density of s can be obtained by integrating out the $u(t)$. Unfortunately, such an integration is intractable. To obtain a useful approximation, we first fix the density p_s to be Gaussian, as discussed above, and we define the nonlinearity f as $f(v)\ =\ v^{-1/2}$. The main motivation for this latter definition is algebraic simplicity that makes a simple approximation possible. Then the marginal density $p(s(t);\ t\ =\ 1,...,\ T)$ equals

$$\int\ \frac{1}{\sqrt{2\pi}^T}\exp\left\{-\frac{1}{2}\sum_t\ [s(t)]^2\left[\sum_\tau\ \phi(\tau)u(t\ -\ \tau)\right]\right\}$$
$$\times\ \prod_t\ p_u(u(t))\left[\sum_\tau\ \phi(\tau)u(t\ -\ \tau)\right]^{1/2}\mathrm{d}u(1)\cdots\mathrm{d}u(T), \tag{A2}$$

which can be manipulated to give (by making the change of variables $t'\ =\ t\ -\ \tau$)

$$\int\ \frac{1}{\sqrt{2\pi}^T}\exp\left(-\frac{1}{2}\sum_{t'}\ u(t')\sum_t\ \{[s(t)]^2\phi(t\ -\ t')\}\right)$$
$$\times\ \prod_{t'}\ p_u(u(t'))\left[\sum_\tau\ \phi(\tau)u(t'\ -\ \tau)\right]^{1/2}\mathrm{d}u(1)\cdots\mathrm{d}u(T). \tag{A3}$$

Now we use the simple approximation

$$\left[\sum_\tau\ \phi(\tau)u(t'\ -\ \tau)\right]^{1/2}\ \approx\ \sqrt{\phi(0)u(t')}. \tag{A4}$$

This is actually a lower bound, and thus our approximation will be a lower bound of the likelihood as well. Now the probability is factorizable with respect to $u(t')$, and we can integrate component by component. This gives us the following approximation:

$$p(s(t);\ t\ =\ 1,...,\ T)\ \sim\ \prod_t\ \exp\left\{G\left(\sum_\tau\ \phi(\tau)[s(t\ -\ \tau)]^2\right)\right\}, \tag{A5}$$

where the scalar function G is obtained from p_u by

$$G(y)\ =\ \log\int\ \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}uy\right)p_u(u)\sqrt{\phi(0)u}\,\mathrm{d}u. \tag{A6}$$

The function G has a similar role to that of the log density of the independent components in classic ICA. The formula for G in Eq. (A6) can be exactly evaluated only in special cases. One such case is obtained if the $u(t)$ are obtained as squares of standardized Gaussian variables. Straightforward calculation then gives the following function:

$$G_0(y)\ =\ -\log(1\ +\ y)\ +\ \tfrac{1}{2}\log\pi^2\phi(0). \tag{A7}$$

In ICA, it is well-known that the exact form of the log density does not affect the consistency of the estimators, as long as the overall shape of the function is correct. This is probably true here as well. Simulations that we have performed support this conjecture.[21] The G in Eq. (A7) is a typical convex function, and thus we use the same convex functions as G as those in ICA and related models. However, the above derivation shows what is the (approximately) optimal G as a function of the probability densities in the model.

ACKNOWLEDGMENTS

We thank Patrik Hoyer, Bruno Olshausen, Konrad Körding, Laurenz Wiskott, Dan Kersten, and Eero Simoncelli for comments and interesting discussions. Funding was provided by the Academy of Finland (Academy Fellow position for A. Hyvärinen and project 48593 to J. Väyrynen) and the Helsinki Graduate School in Computer Science and Engineering to J. Hurri.

Corresponding author Aapo Hyvärinen may be reached at the location on the title page or by e-mail, aapo.hyvarinen@hut.fi; or fax, 358-9-7554892. E-mail addresses for Jarmo Hurri and Jaakko Väyrynen are jarmo.hurri@hut.fi and jjvayryn@mail.cis.hut.fi.

REFERENCES

1. H. B. Barlow, "Single units and sensation: a neuron doctrine for perceptual psychology?" Perception **1**, 371–394 (1972).
2. D. J. Field, "What is the goal of sensory coding?" Neural Comput. **6**, 559–601 (1994).
3. E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," Annu. Rev. Neurosci. **24**, 1193–1216 (2001).
4. B. A. Olshausen, "Principles of image representation in visual cortex," in *The Visual Neurosciences*, L. M. Chalupa and J. S. Werner, eds. (MIT Press, Cambridge, Mass., 2003).
5. A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis* (Wiley Interscience, New York, 2001).
6. A. Hyvärinen, "Sparse code shrinkage: denoising of non-

gaussian data by maximum likelihood estimation," Neural Comput. **11**, 1739–1768 (1999).
7. E. P. Simoncelli and E. H. Adelson, "Noise removal via bayesian wavelet coring," in *Proceedings of the Third IEEE International Conference on Image Processing* (Institute of Electrical and Electronics Engineers, New York, 1996), pp. 379–382.
8. G. E. Hinton and Z. Ghahramani, "Generative models for discovering sparse distributed representations," Philos. Trans. R. Soc. London, Ser. B **352**, 1177–1190 (1997).
9. D. C. Knill and W. Richards, eds., *Perception as Bayesian Inference* (Cambridge U. Press, Cambridge, UK, 1996).
10. A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," IEEE Trans. Neural Netw. **10**, 626–634 (1999).
11. B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," Nature (London) **381**, 607–609 (1996).
12. P. Földiák, "Learning invariance from transformation sequences," Neural Comput. **3**, 194–200 (1991).
13. C. Kayser, W. Einhäuser, O. Dümmer, P. König, and K. Körding, "Extracting slow subspaces from natural videos leads to complex cells," in *Proceedings of the International Conference on Artificial Neural Networks (ICANN2001)*, (Springer-Varlag, **LOCATION**, 2001), pp. 1075–1080.
14. L. Wiskott and T. J. Sejnowski, "Slow feature analysis: unsupervised learning of invariances," Neural Comput. **14**, 715–770 (2002).
15. P. Berkes and L. Wiskott, "Applying slow feature analysis to image sequences yields a rich repertoire of complex cell properties," in *Proceedings of the International Conference on Artificial Neural Networks (ICANN2002)* (Springer-Verlag, **CITY**, 2002), pp. 81–86.
16. J. Hurri and A. Hyvärinen, "Simple-cell-like receptive fields maximize temporal coherence in natural video," Neural Comput. **15**, **XXXX–XXXX** (2003).
17. A. Hyvärinen, "Blind source separation by nonstationarity of variance: cumulant-based approach," IEEE Trans. Neural Netw. **12**, 1471–1474 (2001).
18. C. Zetzsche and G. Krieger, "Nonlinear neurons and high-order statistics: new approaches to human vision and electronic image processing," in *Human Vision and Electronic Imaging IV*, B. Rogowitz and T. V. Pappas, eds., Proc. SPIE **3644**, 2–33 (1999).
19. E. P. Simoncelli and O. Schwartz, "Modeling surround suppression in V1 neurons with a statistically-derived normalization model," in *Advances in Neural Information Processing Systems 11*, M. S. Kearns, S. A. Solla, and D. A. Cohn, eds. (MIT Press, Cambridge, Mass., 1999), pp. 153–159.
20. A. Hyvärinen and P. O. Hoyer, "Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspeces," Neural Comput. **12**, 1705–1720 (2000).
21. A. Hyvärinen, P. O. Hoyer, and M. Inki, "Topographic independent component analysis," Neural Comput. **13**, 1527–1558 (2001).
22. O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control," Nat. Neurosci. **4**, 819–825 (2001).
23. M. J. Wainwright, E. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," Appl. Comput. Harmon. Anal. **11**, 89–123 (2001).
24. D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," J. Physiol. (London) **195**, 215–243 (1968).
25. G. G. Blasdel, "Orientation selectivity, preference, and continuity in monkey striate cortex," J. Neurosci. **12**, 3139–3161 (1992).
26. D. H. Hubel and T. N. Wiesel, "Functional architecture of macaque monkey visual cortex (Ferrier Lecture)," Proc. R. Soc. London, Ser. B **198**, 1–59 (1977).
27. R. B. H. Tootell, M. S. Silverman, S. L. Hamilton, E. Switkes, and R. L. De Valois, "Functional anatomy of macaque striate cortex. V. Spatial frequency," J. Neurosci. **8**, 1610–1624 (1988).
28. T. Kohonen, *Self-Organizing Maps* (Springer, **LOCATION**, 1995).
29. N. V. Swindale, "The development of topography in the visual cortex: a review of models," Network **7**, 161–247 (1996).
30. C. von der Malsburg, "Self-organization of orientation-sensitive cells in the striate cortex," Kybernetik **14**, 85–100 (1973).
31. A. Hyvärinen and P. O. Hoyer, "A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images," Vision Res. **41**, 2413–2423 (2001).
32. A. J. Bell and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," Vision Res. **37**, 3327–3338 (1997).
33. J. H. van Hateren and A. van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex," Proc. R. Soc. London, Ser. B **265**, 359–366 (1998).
34. P. Comon, "Independent component analysis—a new concept?" Signal Process. **36**, 287–314 (1994).
35. C. Jutten and J. Hérault, "Blind separation of sources. Part I: An adaptive algorithm based on neuromimetic architecture," Signal Process. **24**, 1–10 (1991).
36. J. H. van Hateren and D. L. Ruderman, "Independent component analysis of natural image sequences yields spatiotemporal filters similar to simple cells in primary visual cortex," Proc. R. Soc. London, Ser. B **265**, 2315–2320 (1998).
37. J. Hurri and A. Hyvärinen, "A two-layer temporal generative model of natural video exhibits complex-cell-like pooling of simple cell outputs," in *Computational Neuroscience: Trends in Research 2003*, E. De Schutter, ed. (Elsevier, Amsterdam, The Netherlands, 2003).
38. B. A. Olshausen, "Sparse codes and spikes," in *Statistical Theories of the Brain*, R. Rao and B. A. Olshausen, eds. (MIT Press, Cambridge, Mass. 2001).
39. R. C. Emerson, J. R. Bergen, and E. H. Adelson, "Directionally selective complex cells and the computation of motion energy in cat visual cortex," Vision Res. **32**, 203–218 (1992).
40. D. Pollen and S. Ronner, "Visual cortical neurons as localized spatial frequency filters," IEEE Trans. Syst. Man Cybern. **SMC-13**, 907–916 (1983).
41. M. Welicky, W. H. Bosking, and D. Fitzpatrick, "A systematic map of direction preference in primary visual cortex," Nature (London) **379**, 725–728 (1996).
42. P. O. Hoyer and A. Hyvärinen, "A multi-layer sparse coding network learns contour coding from natural images," Vision Res. **42**, 1593–1605 (2002).
43. J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Adaptive Wiener denoising using a Gaussian scale mixture model in the wavelet domain," in *Proceedings of the International Conference on Image Processing* (Society for Imaging Science and Technology, Springfield, Va., 2001), pp. **XX–XX**.
44. R. Durbin and G. Mitchison, "A dimension reduction framework for understanding cortical maps," Nature (London) **343**, 644–647 (1990).
45. W. S. Geisler and D. G. Albrecht, "Cortical neurons: isolation of contrast gain control," Vision Res. **32**, 1409–1410 (1992).
46. D. Heeger, "Normalization of cell responses in cat striate cortex," Visual Neurosci. **9**, 181–198 (1992).
47. K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," Neural Networks **8**, 411–419 (1995).
48. D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of non-stationary sources," in *Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)* (**PUBLISHER**, **CITY**, 2000), pp. 187–193.
49. R. F. Engle, ed., *ARCH: Selected Readings* (Oxford U. Press, Oxford, UK, 1995).
50. B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?" Vision Res. **37**, 3311–3325 (1997).
51. A. Hyvärinen and M. Inki, "Estimating overcomplete inde-

pendent component bases from image windows," J. Math. Imaging Vision **17**, 139–152 (2002).
52. A. Pece, "The problem of sparse image coding," J. Math. Imaging Vision **17**, 87–106 (2002).
53. B. A. Olshausen, P. Sallee, and M. S. Lewicki, "Learning sparse image codes using a wavelet pyramid architecture," in *Advances in Neural Information Processing Systems*, (MIT Press, Cambridge, Mass., 2001), Vol. 13, pp. 887–893.
54. P. Paatero and U. Tapper, "Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values," Environmetrics **5**, 111–126 (1994).
55. D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature (London) **401**, 788–791 (1999).
56. P. O. Hoyer, "Modeling receptive fields with non-negative sparse coding," in *Computational Neuroscience: Trends in Research 2003*, E. De Schutter, ed. Elsevier, Amsterdam, The Netherlands, 2003.
57. D. R. Tailor, L. H. Finkel, and G. Buchsbaum, "Color-opponent receptive fields derived from independent component analysis of natural images," Vision Res. **40**, 2671–2676 (2000).
58. T. Wachtler, T-W. Lee, and T. J. Sejnowski, "Chromatic structure of natural scenes," J. Opt. Soc. Am. A **18**, 65–77 (2001).
59. P. O. Hoyer and A. Hyvärinen, "Independent component analysis applied to feature extraction from colour and stereo images," Network Comput. Neural Syst. **11**, 191–210 (2000).

PROOF COPY [A-8652] 002307JOA