# Slow feature analysis yields a rich repertoire of complex cell properties

**Pietro Berkes**

Institute for Theoretical Biology, Humboldt University, Berlin, Germany

**Laurenz Wiskott**

Institute for Theoretical Biology, Humboldt University, Berlin, Germany

In this study we investigate temporal slowness as a learning principle for receptive fields using slow feature analysis, a new algorithm to determine functions that extract slowly varying signals from the input data. We find a good qualitative and quantitative match between the set of learned functions trained on image sequences and the population of complex cells in the primary visual cortex (V1). The functions show many properties found also experimentally in complex cells, such as direction selectivity, non-orthogonal inhibition, end-inhibition, and side-inhibition. Our results demonstrate that a single unsupervised learning principle can account for such a rich repertoire of receptive field properties.

Keywords: complex cells, slow feature analysis, temporal slowness, computational model, spatiotemporal receptive fields

## 1. Introduction

Primary visual cortex (V1) is the first cortical area dedicated to visual processing. This area has been intensively studied neurophysiologically since the seminal work by Hubel and Wiesel (1962), who also introduced the standard classification of neurons in V1 into two main groups: *simple* and *complex cells*. These neurons are conceived as edge or line detectors: Simple cells respond to bars having a specific orientation and position in the visual field; complex cells also respond to oriented bars but are insensitive to their exact position.

Idealized simple and complex cells can be described by Gabor wavelets (Pollen & Ronner, 1981; Adelson & Bergen, 1985; Jones & Palmer, 1987), which have the shape of sine gratings with a Gaussian envelope function. A single Gabor wavelet used as a linear filter (Figure 1a) is similar to a simple cell, because the response depends on the exact alignment of a stimulus bar on an excitatory (positive) subfield of the wavelet. Taking the square sum of the responses of two Gabor wavelets with identical envelope function, frequency, and orientation but with a 90-deg phase difference (Figure 1b) yields a model of a complex cell that is insensitive to the exact location of the bar (following a rule similar to the relation $\sin(x)^2 + \cos(x)^2 = 1$) while still being sensitive to its orientation. We will refer to these models as the *classical models* of simple and complex cells.

This idealized picture, however, is clearly not complete. In particular, complex cells in V1 show a much richer repertoire of receptive field properties than can be explained with the classical model. For example, they show end-inhibition, side-inhibition, direction selectivity, and sharpened or broadened tuning to orientation or frequency (Hubel & Wiesel, 1962; Sillito, 1975; Schiller, Finlay, & Volman, 1976a, 1976b, 1976c; De Valois, Albrecht, & Thorell, 1982; De Valois, Yund, & Hepler, 1982; Dobbin, Zucker, & Cynade, 1987; Versavel, Orban, & Lagae, 1990; Skottun et al., 1991; DeAngelis, Freeman, & Ohzawa, 1994; Shevelev, 1998; Walker, Ohzawa, & Freeman, 1999; Ringach, Bredfeldt, Shapley, & Hawken, 2002).

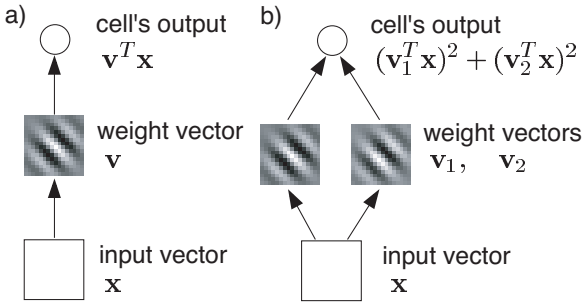



Figure 1. Classical models of simple and complex cells. (a). Simple cells respond best to oriented bars at a specific position in the visual field, and are well modeled by a linear Gabor filter (Jones & Palmer, 1987). (b). Complex cells respond to oriented bars but are insensitive to their local position. The classical model (energy model) consists of two linear Gabor filters having the same shape except for a 90-deg phase difference. The square sum of the response of the two filters yields the output (Adelson & Bergen, 1985). For comparison, orientation, frequency, and size of the subunits in this figure have been fitted to those of the optimal excitatory stimulus of the unit shown in Figure 7a.

A possible approach to the study of the organization of the visual cortex is to assume that it is philo- or ontogenetically adapted to the statistics of its input to satisfy one (or possibly more) computational objectives (e.g., Field, 1994). The neurons resulting from such an information-processing strategy would compute input-output functions that pro-

doi:10.1167/5.6.9 Received October 1, 2003; published July 20, 2005 ISSN 1534-7362 

vide particular advantages for further decoding and processing. The computational approach does not necessarily provide an explanation of the cortical mechanisms involved in the computation, but it can give a powerful functional explanation of experimental data.

In this work we investigate *temporal slowness* as a possible computational principle for the emergence of complex cell receptive fields in the visual cortex. The slowness principle is based on the observation that the environment, sensory signals, and internal representations of the environment vary on different time scales. The environment (e.g., the objects we see around us) changes usually on a relatively slow time scale. Sensory signals on the other hand, such as the responses of single receptors in the retina, vary on a faster time scale, because even a small eye movement or shift of a textured object may lead to a rapid variation of the light intensity received by a receptor neuron. The internal representation of the environment, finally, should vary on a time scale similar to that of the environment itself (i.e., on a slow time scale). If we succeed in extracting slowly varying features from the quickly varying sensory signal, then the features are likely to reflect the properties of the environment and are in addition invariant or at least robust to frequent transformations of the sensory input, such as visual translation, rotation, or zoom. Our working hypothesis in this study is that the cortex organizes according to this principle to build a consistent internal representation of the environment. To verify this hypothesis we consider a space of nonlinear input-output functions (here the space of all polynomials of degree 2), determine those functions that extract the slowest features in response to natural image sequences, and compare their properties to those of complex cells in V1 described in the literature.

An early description of this principle was given by Hinton (1989, p. 208) and early models based on temporal slowness were presented by Földiák (1991) and Mitchison (1991). Successive studies applied this principle to the extraction of disparity from stereograms (Stone, 1996) or from artificially generated simple cell outputs (Wiskott & Sejnowski, 2002) and to blind source separation (Stone, 2001). Ideas and learning rules related to temporal slowness can also be found in the works by Becker and Hinton (1993), O'Reilly and Johnson (1994), and Peng, Sha, Gan, and Wei (1998). For other studies modeling complex and simple cells, see Discussion.

The following section introduces the slow feature analysis algorithm. Section 3 presents the input data set and the methods used to analyze the results. Section 4 describes the simulation results and compares the learned functions with neurons reported in the physiological literature. In Section 5 we investigate the role of spatial transformations, the statistics of the input images, dimensionality reduction, and asymmetric decorrelation in our results with a set of control experiments. The work concludes with a discussion in Section 6. Appendix A contains additional notes to the text that concern more technical aspects of our model that might be useful to the theoretical reader but are not central to the main results.

# 2. Slow feature analysis

## 2.1 Problem statement

The learning task we want to solve is the following. Given a multi-dimensional input signal $\mathbf{x}(t)$ we want to find (scalar) functions $g_j(\mathbf{x})$, which generate output signals $y_j(t) = g_j(\mathbf{x}(t))$ from the input signals that vary as slowly as possible but carry significant information. To ensure the latter we require the output signals to have unit variance and be mutually uncorrelated. It is important to note that even though the objective is the slowness of the output signals, the process by which the output is computed from the input is very fast or in the mathematical idealization even instantaneous. Slowness can therefore not be achieved simply by low-pass filtering. Thus only if the input signal has some underlying, slowly varying causes does the system have a chance of extracting slowly varying output signals at all. It is exactly this apparent paradox of instantaneous processing on the one hand and the slowness objective on the other hand that guarantees that the extracted output signals represent relevant features of the underlying causes that gave rise to the input signal.

In more mathematical terms the problem can be stated as follows (Wiskott & Sejnowski, 2002): given a multi-dimensional input signal $\mathbf{x}(t) = (x_1(t),\ldots,x_N(t))\ t \in [t_0,t_1]$, find a set of real-valued functions $g_1(\mathbf{x}),\ldots, g_K(\mathbf{x})$ lying in a function space F so that for the output signals $y_j(t) := g_j(\mathbf{x}(t))$

$$\Delta(y_j) := \left\langle \dot{y}_j^2 \right\rangle_t \quad \text{is minimal} \tag{1}$$

under the constraints

$$\left\langle y_j \right\rangle_t = 0 \quad \text{(zero mean)}, \tag{2}$$

$$\left\langle y_j^2 \right\rangle_t = 1 \quad \text{(unit variance)}, \tag{3}$$

$$\forall i < j, \quad \left\langle y_i y_j \right\rangle_t = 0 \quad \text{(decorrelation and order)}, \tag{4}$$

with $\langle\cdot\rangle_t$ and $\dot{y}$ indicating time-averaging and the time derivative of $y$, respectively. Equation 1 introduces a measure of the temporal variation of a signal (the $\Delta$-*value* of a signal) equal to the mean of the squared derivative of the signal. This quantity is large for quickly varying signals

and zero for constant signals. We will also use a more intuitive measure of slowness, the *β-value*, defined as $\beta(y_j) = (1/2\pi)\sqrt{\Delta(y_j)}$. A sine wave with period $T$ and unit variance has a $\beta$-value of $1/T$ when averaged over an integer number of oscillations. The zero-mean Constraint 2 is present for convenience only, so that Constraints 3 and 4 take a simple form. Constraint 3 means that each signal should carry some information and avoids the trivial solution $g_j(\mathbf{x}) = 0$. Alternatively, one could drop this constraint and divide the right side of Equation 1 by the variance $\left\langle y_j^2 \right\rangle_t$. Constraint 4 forces different signals to be uncorrelated and thus to code for different aspects of the input. It also induces an order, the first output signal being the slowest one, the second being the second slowest, etc. Control Experiment 4 (Section 5.4) investigates the role of this constraint on the learned functions.

We solve the optimization problem with slow feature analysis (SFA) (Wiskott, 1998; Wiskott & Sejnowski, 2002), an unsupervised algorithm that permits us to find the optimal set of functions $g_j$ in a general finite dimensional function space and that is efficient enough to do simulations of reasonable scale in terms of size and dimensionality of the input signals. Because SFA is based on an eigenvector approach (e.g., like principal component analysis), it finds the global solutions in a single iteration and has no convergence problems.

The following three sections sketch the mathematical background and the definition of the algorithm. They are not necessary for understanding the remainder of the work. The reader less interested in the mathematical details might want to skip them and continue with Section 3. For the purposes of this study, it is sufficient to remember that slow feature analysis finds input-output functions that extract slowly varying features from a typical input signal in a nontrivial way (i.e., instantaneously and without low-pass filtering).

## 2.2 The linear case

Consider first the linear case $g_j(\mathbf{x}) = \mathbf{w}_j^T \mathbf{x}$ for some input $\mathbf{x}$ and weight vectors $\mathbf{w}_j$. In the following we assume $\mathbf{x}$ to have zero mean (i.e., $\langle \mathbf{x} \rangle_t = 0$) without loss of generality. This implies that Constraint 2 is fulfilled, because $\langle y_j \rangle_t = \langle \mathbf{w}_j^T \mathbf{x} \rangle_t = \mathbf{w}_j^T \langle \mathbf{x} \rangle_t = 0$.

We can rewrite Equations 1, 3, and 4 as

$$\begin{aligned} \Delta(y_j) &= \left\langle \dot{y}_j^2 \right\rangle_t = \left\langle (\mathbf{w}_j^T \dot{\mathbf{x}})^2 \right\rangle_t \\ &= \mathbf{w}_j^T \left\langle \dot{\mathbf{x}} \dot{\mathbf{x}}^T \right\rangle_t \mathbf{w}_j =: \mathbf{w}_j^T \mathbf{A} \mathbf{w}_j \end{aligned} \tag{5}$$

$$\begin{aligned} \left\langle y_i y_j \right\rangle_t &= \left\langle (\mathbf{w}_i^T \mathbf{x})(\mathbf{w}_j^T \mathbf{x}) \right\rangle_t \\ &= \mathbf{w}_i^T \left\langle \mathbf{x} \mathbf{x}^T \right\rangle_t \mathbf{w}_j \\ &=: \mathbf{w}_i^T \mathbf{B} \mathbf{w}_j . \end{aligned} \tag{6}$$

If we integrate Constraint 3 in the objective function 1, as suggested in the previous section, we obtain

$$\Delta(y_j) = \frac{\left\langle \dot{y}_j^2 \right\rangle_t}{\left\langle y_j^2 \right\rangle_t} = \frac{\mathbf{w}_j^T \mathbf{A} \mathbf{w}_j}{\mathbf{w}_j^T \mathbf{B} \mathbf{w}_j} \quad . \tag{7}$$

It is known from linear algebra that the weight vectors $\mathbf{w}_j$ that minimize this equation correspond to the eigenvectors of the generalized eigenvalue problem

$$\mathbf{A}\mathbf{W} = \mathbf{B}\mathbf{W}\mathbf{\Lambda} \quad , \tag{8}$$

where $\mathbf{W}$ is the matrix of the generalized eigenvectors and $\mathbf{\Lambda}$ is the diagonal matrix of the generalized eigenvalues $\lambda_1, \ldots, \lambda_N$ (e.g., see Gantmacher, 1959, Chap. 10.7, Theorems 8, 10, and 11). In particular, the vectors $\mathbf{w}_j$ can be normalized such that $\mathbf{w}_i^T \mathbf{B} \mathbf{w}_j = \delta_{ij}$, which implies that Constraints 3 and 4 are fulfilled:

$$\left\langle y_j^2 \right\rangle_t = \mathbf{w}_j^T \mathbf{B} \mathbf{w}_j = 1 \quad , \tag{9}$$

$$i \neq j, \quad \left\langle y_i y_j \right\rangle_t = \mathbf{w}_i^T \mathbf{B} \mathbf{w}_j = 0 \quad . \tag{10}$$

Note that by substituting Equation 8 into Equation 7 one obtains $\Delta(y_j) = \lambda_j$, so that by sorting the eigenvectors by increasing eigenvalues we induce an order where the most slowly varying signals have lowest indices [i.e., $\Delta(y_1) \leq \Delta(y_2) \leq \ldots \leq \Delta(y_N)$ ].

## 2.3 The general case

In the more general case of a nonlinear but finite-dimensional function space F, consider a basis $h_1, \ldots, h_M$ of F. For example, in the standard case where F is the space of all polynomials of degree $n$, the basis will include all monomials up to order $n$.

Defining the *expanded input*

$$\mathbf{h}(\mathbf{x}) := (h_1(\mathbf{x}), \ldots, h_M(\mathbf{x}))^T \quad , \tag{11}$$

every function $g \in \mathrm{F}$ can be expressed as

$$g(\mathbf{x}) = \sum_{k=1}^{M} w_k h_k(\mathbf{x}) = \mathbf{w}^T \mathbf{h}(\mathbf{x}) \quad . \tag{12}$$

This leads us back to the linear case if we assume that $\mathbf{h}(\mathbf{x})$ has zero mean (again, without loss of generality), which can be easily obtained in practice by subtracting the mean over time $\langle \mathbf{h}(\mathbf{x}) \rangle_t =: \mathbf{h}_0$ from the expanded input signal.

For example, in the case of 3 input dimensions ($N = 3$) and polynomials of degree 2 we have

$$\mathbf{h}(\mathbf{x}) = (x_1^2, x_1x_2, x_1x_3, x_2^2, x_2x_3, x_3^2, x_1, x_2, x_3)^T - \mathbf{h}_0 \quad (13)$$

and

$$g(\mathbf{x}) = w_1x_1^2 + w_2x_1x_2 + w_3x_1x_3 + w_4x_2^2 + w_5x_2x_3 + w_6x_3^2 + w_7x_1 + w_8x_2 + w_9x_3 - \mathbf{w}^T\mathbf{h}_0 \quad . \quad (14)$$

Every polynomial of degree 2 in the 3 input variables can then be expressed by an appropriate choice of the weights $w_j$.

### 2.4 The SFA algorithm

We can now formulate the slow feature analysis (SFA) algorithm (cf. Wiskott & Sejnowski, 2002):

**Nonlinear expansion:** expand the input data and subtract the mean over time to obtain the expanded signal $\mathbf{z} := \mathbf{h}(\mathbf{x}) - \mathbf{h}_0 = (h_1(\mathbf{x}), \ldots, h_M(\mathbf{x}))^T - \mathbf{h}_0$.

**Slow feature extraction:** solve the generalized eigenvalue problem

$$\mathbf{AW} = \mathbf{BW\Lambda}, \quad (15)$$

$$\text{with} \quad \mathbf{A} := \left\langle \dot{\mathbf{z}}\dot{\mathbf{z}}^T \right\rangle_t \quad (16)$$

$$\text{and} \quad \mathbf{B} := \left\langle \mathbf{z}\mathbf{z}^T \right\rangle_t . \quad (17)$$

The $K$ eigenvectors $\mathbf{w}_1, \ldots, \mathbf{w}_K$ ($K \le M$) corresponding to the smallest generalized eigenvalues $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_K$ define the nonlinear input-output functions $g_1(\mathbf{x}), \ldots, g_K(\mathbf{x}) \in \mathrm{F}$:

$$g_j(\mathbf{x}) = \mathbf{w}_j^T(\mathbf{h}(\mathbf{x}) - \mathbf{h}_0), \quad (18)$$

which satisfy Constraints 2-4 and minimize 1.

In other words, to solve the optimization problem (Equation 1), it is sufficient to compute the covariance matrix of the signals and that of their derivatives in the expanded space and then solve the generalized eigenvalue problem (Equation15). In the simulations presented here, the derivative of $\mathbf{z}(t)$ is computed by the linear approximation $\dot{\mathbf{z}}(t) \approx (\mathbf{z}(t + \Delta t) - \mathbf{z}(t))/\Delta t$ ($\Delta t = 1$ throughout this work). Simulations performed with cubic interpolation showed equivalent results.

## 3. Methods

### 3.1 Input data

Our data source consisted of 36 gray-valued natural images extracted from the natural stimuli collection of van Hateren (available online at http://hlab.phys.rug.nl/archive.html). The images were chosen by the authors to contain a variety of natural contents, including trees, flowers, animals, water, and so on. We avoided highly geometrical human artifacts. The images were preprocessed as suggested in van Hateren and van der Schaaf (1998) by block-averaging (block size 2×2) and by taking the logarithm of the pixel intensities. This procedure corrects possible calibration problems and reshapes the contrast of the images. After preprocessing, the images were 768×512 pixels large. An extensive discussion of the images and of the preprocessing can be found in van Hateren and van der Schaaf (1998).

We constructed image sequences by moving a quadratic window over the images by translation, rotation, and zoom and subsequently rescaling the frames (to compensate for the zoom) to a standard size of 16×16 pixels. The input window was not masked or weighted in any way. The initial position, orientation, and zoom for each sequence were chosen at random. The transformations were performed simultaneously, so that each frame differed from the previous one by position, orientation, and scale. If the window moved out of the image, the sequence was discarded and a new one was started from scratch. Each individual sequence was 100 frames long with a total of 250,000 frames per simulation. (The length of the sequences is irrelevant to the algorithm as long as the total number of input vectors is preserved.) Each image contributed an equal number of frames. Figure 2 shows one example sequence. The displacements per frame in horizontal and vertical direction were Gaussian-distributed with zero mean and standard deviation 3.56 pixels. The angular speed measured in radians/frame and the magnification difference (defined as the difference between the magnification factor of two successive frames) followed a Gaussian distribution with mean 0 and standard deviation 0.12 and 0.03, respectively. Other simulations showed qualitatively similar results within a reasonable range of parameters, although the distribution of some unit properties might vary. See Control Experiment 1 (Section 5.1) for a study of the influence of the individual transformations. To include temporal information, the input vectors to SFA were formed by the pixel intensities of two consecutive frames at times $t$ and $\Delta t = 1$, so that the second frame in one input vector was equal to the first frame in the next, as illustrated in Figure 2. (The time difference $\Delta t$ was the same used to compute the time derivative.) Note that with two frames as an input, processing is not strictly instantaneous anymore, but slowness still cannot be achieved by low-pass filtering.

The function space F on which SFA is performed (Section 2.1) is chosen here to be the set of all polynomials of degree 2, as discussed extensively in Section 6.2. A run with SFA requires the computation of two large covariance matrices having in the order of $O(M^2)$ elements, where $M$ is the dimension of the considered function space. In the case of polynomials of degree 2 this corresponds to a number of elements in the order of $O(N^4)$, where $N$ is the input dimension. Because this is computationally expensive, we performed a standard preprocessing step using principal

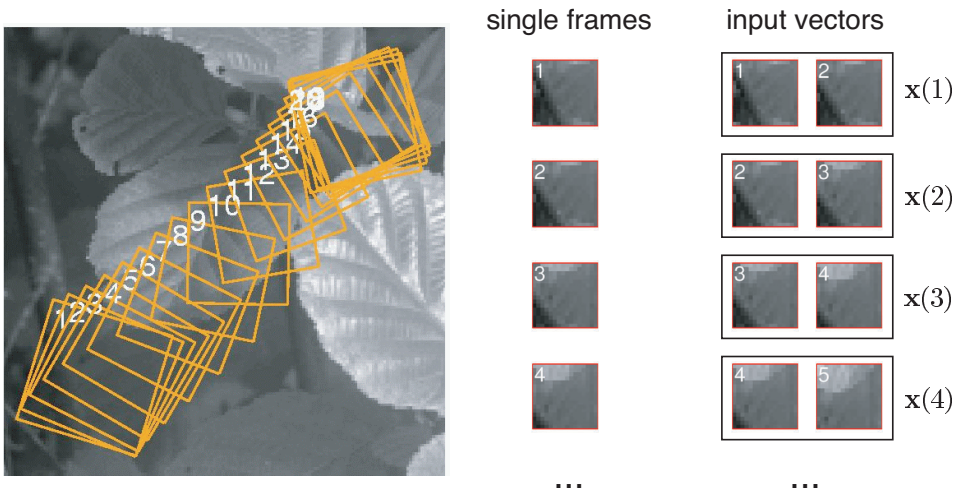



Figure 2. Natural image sequences. A close-up of one of the natural images used in the simulations (left). The numbered squares show the position, size, and orientation of the input window for a short sequence of 20 frames. The content of the window is then rescaled to 16×16 pixels (center). The input to SFA consists of pairs of successive frames (right) to include temporal information.

component analysis (PCA) to reduce the dimensionality of the input vectors from 16×16×2=512 to $N = 100$, capturing 93% of the total variance (see Appendix A.1 for additional remarks). In Control Experiment 3 (Section 5.3), we present the results of a simulation performed with smaller patches (10×10 pixels) and no dimensionality reduction.

## 3.2 Analysis methods

In the simulations presented here, SFA learns polynomials of degree 2 that applied to our visual input stimuli have the most slowly varying output (which does not imply that processing is slow; see Section 2.1). We refer to the *i-th unit*. The units are ordered by slowness (the first one being the slowest) and their outputs are mutually uncorrelated. Because the sign of a unit's response is arbitrary in the optimization problem, we have chosen it here such that the strongest response to an input vector with a given norm is positive (i.e., the magnitude of the response to the optimal excitatory stimulus, $\mathbf{x}^+$, is greater than that to the optimal inhibitory stimulus, $\mathbf{x}^-$; see below).

Because the input vectors are pairs of image patches, the functions $g_j$ can be interpreted as nonlinear spatiotemporal receptive fields and be tested with input stimuli much like in neurophysiological experiments. The units can have a spontaneous firing rate (i.e., a non-zero response to a blank visual input). As in physiological experiments we interpret an output lower than the spontaneous one as active inhibition. The absolute value of the spontaneous firing rate is fixed by the zero mean constraint (Equation 2) and has no direct interpretation.

To analyze the units, we first compute for each of them the optimal excitatory stimulus $\mathbf{x}^+$ and the optimal inhibitory stimulus $\mathbf{x}^-$, which correspond to the input that elicits the strongest positive and strongest negative output from the unit, respectively, given a constant norm $r$ of the input vector (i.e., a fixed energy constraint) (Figure 3). We choose $r$ to be the mean norm of the training vectors because we want $\mathbf{x}^+$ and $\mathbf{x}^-$ to be representative of the typical input. This is in analogy to the physiological practice of characterizing a neuron by the stimulus to which the neuron responds best (e.g., Dayan & Abbott, 2001, Chap. 2.2). Because in our model we have an explicit definition of the input-output functions of our units, $\mathbf{x}^+$ and $\mathbf{x}^-$ can be computed analytically (Berkes & Wiskott, 2005).

From the two $\mathbf{x}^+$ patches we compute the size and position of the receptive field and by Fourier analysis the preferred frequency, orientation, speed, and direction of a unit. In some units the preferred parameters for the patch at time $t$ and that at time $t + \Delta t$ are slightly different, in which case we take the mean of the two.

Although the optimal stimuli carry much information, they give only a partial view of the behavior of a unit, because these are nonlinear. To gain further insight into the response properties we use an appropriate pair of *test images* (one at time $t$ and one at time $t + \Delta t$) and compute for each unit the corresponding *response image* (Figure 4). The response image is computed by cutting a 16×16 window at each point of the test images, using it as the input to the unit and plotting its output at the corresponding point (cf. Creutzfeldt & Nothdurft, 1978).

To study the response to a range of frequencies and orientations we use a test image that consists of a circular pattern of sine waves with frequency increasing from the circumference to the center. The frequency increase is logarithmic (i.e., an equal distance along the radius corresponds to an equal frequency difference in octaves) to make the comparison with physiological data easier. We let the ring patterns move outward at the preferred speed of the unit (Figure 4a). These images contain information not only about the whole range of frequencies and orientations to which the unit responds or by which it is inhibited but also about the sensitivity of the unit to the phase of the grating.

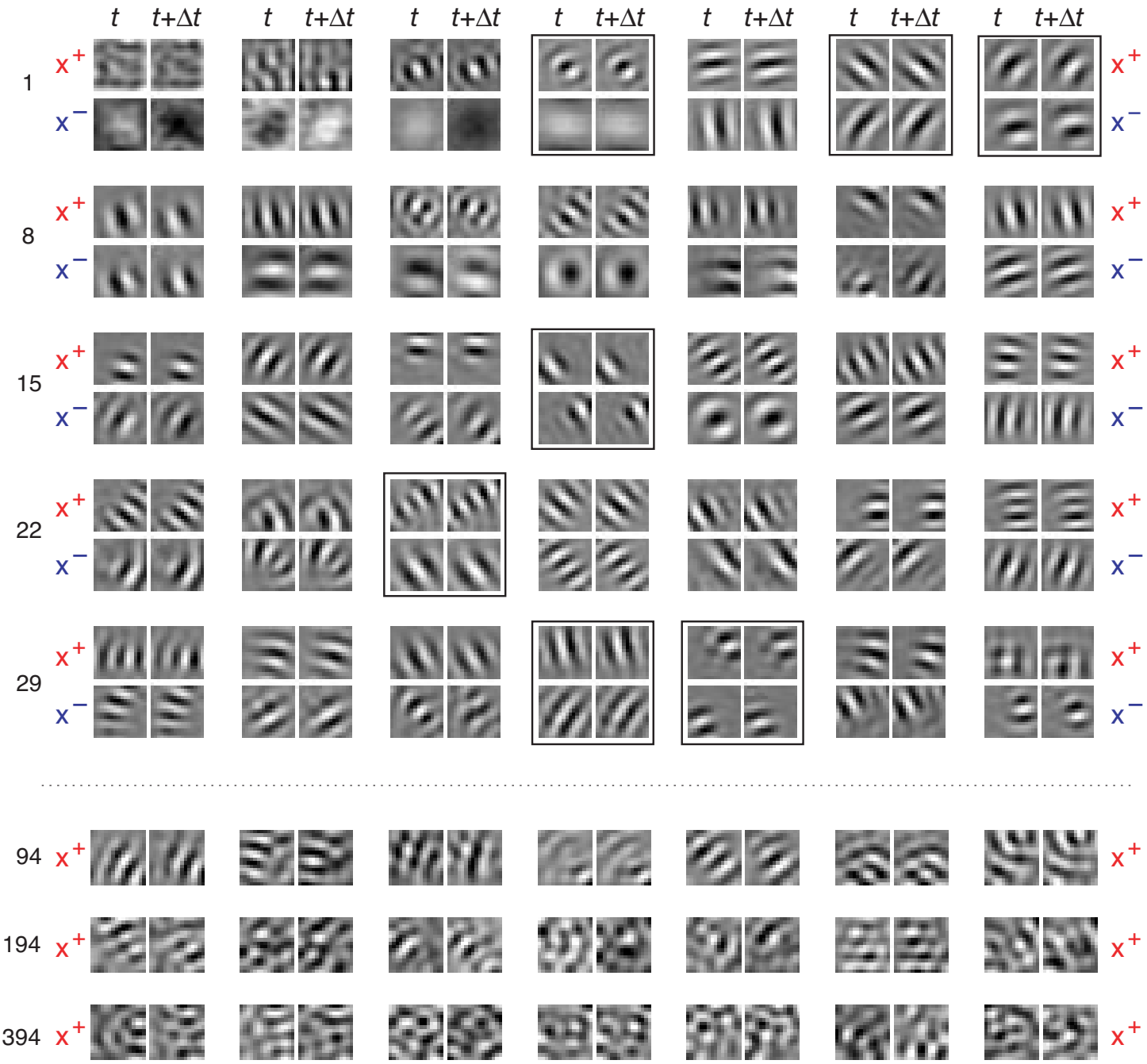



Figure 3. Optimal stimuli. Top five rows: optimal excitatory stimuli ($\mathbf{x}^+$) and optimal inhibitory stimuli ($\mathbf{x}^-$) of the first 35 units of the simulation described in the text. For most units $\mathbf{x}^+$ and $\mathbf{x}^-$ look like Gabor wavelets in agreement with physiological data. $\mathbf{x}^+$ gives information about the preferred frequency and orientation of a unit and about the size and position of its receptive field. A comparison between the patches at time $t$ and at time $t + \Delta t$ hints at the temporal structure of the receptive field (e.g., its preferred speed and direction). The units surrounded by a black frame are analyzed in more detail in Figure 7, Figure 10, and Figure 11. Bottom three rows: optimal excitatory stimuli for Units 94-100, 194-200, and 394-400. The Gabor-like shape of $\mathbf{x}^+$ begins to degrade around Unit 100, and becomes unstructured for successive units corresponding to functions with quickly varying output. (More optimal stimuli can be found online as indicated in Additional Material.)

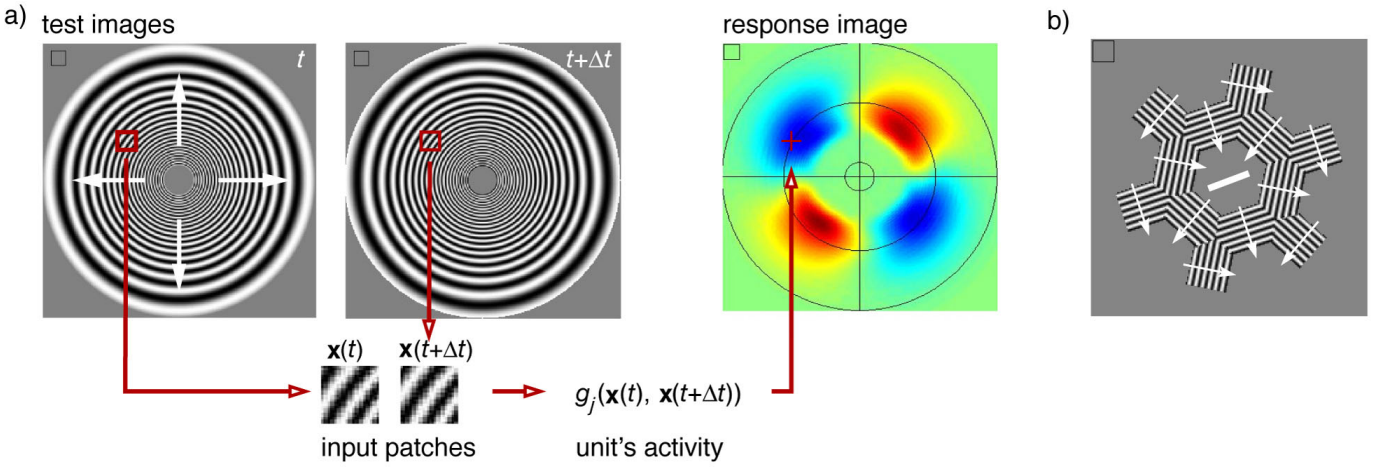



Figure 4. Test and response images. (a). This image illustrates how the response images are computed. Given a test image at time $t$ and at time $t + \Delta t$, at every position two 16×16 input patches are cut out and used as the input to the considered unit. The output is then plotted at the corresponding point of the response image. The colors are normalized such that red corresponds to excitation, blue to inhibition, and green to the spontaneous firing rate. The square at the upper left corner of the response image indicates the size of the input patches. The circular test image shown on the left is used to investigate the response of a unit to a range of frequencies and orientations. The gratings move outward at the preferred speed of the considered unit, as indicated by the white arrows. (b). Test image used to investigate end- and side-inhibition. The hexagonal shape is oriented such that it is aligned to the preferred orientation of the considered unit, indicated by the central bar. The gratings are tuned to the preferred frequency and move at the preferred speed of the unit in the direction shown by the thin arrows.

If a unit is sensitive, the response image shows oscillations in the radial direction, whereas if there are no oscillations, the unit is phase-invariant. Moreover, at two opposite points of a ring the orientation is equal but the grating is moving in opposite directions, and different responses indicate selectivity to the direction of motion. An illustrative example for the classical simple and complex cell model is shown in Figure 5. Because the gratings are curved, an additional factor due to curvature selection might be present in the response images. However, we find that in general this effect is negligible.

The circular response images are similar to the Fourier representation of neural responses used by Ringach et al. (2002) (with the radial axis inverted), but they are more informative in that they also contain information about phase-shift behavior and direction selectivity. Experimental readers might be more familiar with orientation and frequency tuning curves. The circular response images contain this information, too (a slice of the response image along the radial direction would give a frequency-tuning curve, whereas a circular slice would give an orientation-tuning curve), but in addition it shows how the unit behaves at non-optimal parameters.

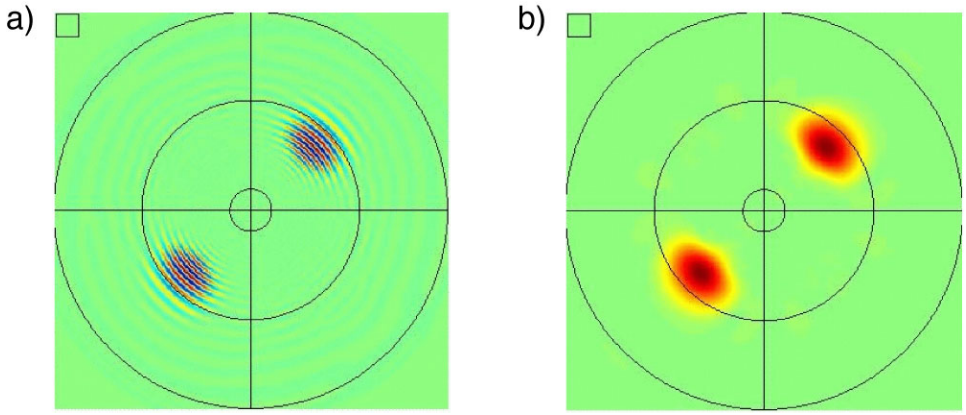



Figure 5. Response images for the classical models of simple and complex cells. (a). Response image for the classical model of simple cells (Figure 1a). The model shows a narrow orientation- and frequency-tuning and oscillations due to phase sensitivity. (b). Response image for the classical model of complex cells (Figure 1b). The orientation and frequency tuning are the same as in (a), but the oscillations have disappeared because the model is phase-insensitive.

We use hexagonal-shaped test images (Figure 4b) to investigate end- and side-inhibition in our units. The hexagon is oriented such that two of the branches are aligned to the preferred orientation as indicated by the central bar. The gratings are set to the preferred frequency and move at the preferred speed as shown by the arrows. The branches of the hexagon are useful to determine if a unit is end- or side-inhibited: on their border the receptive field is only partially filled while in the middle the grating occupies the whole input patch. If the response drops between border and center, the unit is end- or side-inhibited. Of particular interest are the branches at the preferred orientation; on the additional four branches it is possible to see if the inhibition is effective also at other orientations. The central hexagonal part contains angles at various orientations, and is useful to study the curvature selectivity of the units. If the preferred speed is not zero, in the second image there is one junction for each branch where the sine gratings of two branches do not coincide anymore. This might in principle distort the response in those areas. (Note that the hexagonal response image shown in Figure 10b has preferred speed zero, and is thus not affected.)

We additionally performed experiments with drifting sine gratings to compute various unit properties and compare them with physiological results. The sine-grating parameters were always set to the preferred ones of the considered unit. For example, the polar plots of Figure 7a-c.3 were generated by presenting sine gratings to a unit at different orientations and with frequency, speed, position, and size fixed to the preferred ones. The plots show the response of the unit normalized by the maximum (radial axis) versus orientation of the sine grating (angular direction). Unless stated otherwise, comparisons are always made with experimental data of complex cells only.

Another technique that consists in computing the invariances of the optimal stimuli (i.e., the directions in which a variation of $\mathbf{x}^+$ or $\mathbf{x}^-$ has the least effect on the output of the unit) was described in Berkes and Wiskott (2005).

# 4. Results

We now describe units obtained in a single simulation and make a comparison with corresponding cells reported in the experimental literature. For each simulation SFA extracts a complete basis of the considered function space ordered by decreasing slowness. In our case this corresponds to 5150 polynomials of degree 2. Of course, the last functions are actually the ones with the most *quickly* varying output signal and will not be considered. We use the mean $\beta$-value of the pixel intensities as a reference, because we don't want functions that compute a signal varying more quickly than their input. Figure 6 shows the $\beta$-values of the first 400 units when tested on training data or on previously unseen sequences with a total of 400,000 frames. The units remain slowly varying also on test data, and their order is largely preserved. At about Unit 100 the shape of the optimal excitatory stimulus begins to degrade (Figure 3, bottom rows) and the $\beta$-values come close to that of the input signal (>90%). For these reasons we consider here only the first 100 functions.

Although for illustrative purposes we mainly concentrate on individual units, we also find a good qualitative and quantitative match on population level between the considered units and complex cells in V1. We did not find any unit among the first 100 whose properties were in contradiction with those of neurons in V1.

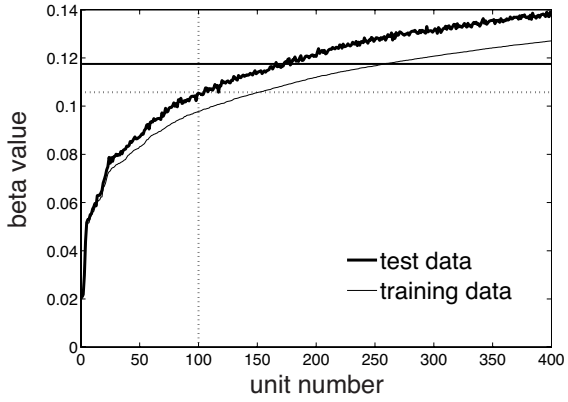



Figure 6. Beta-values. The $\beta$-values of the first 400 units when applied to the training data (thin line) or to the test data (novel sequences with a total of 400,000 frames, thick line). The units remain slow and ordered also on test data. The horizontal solid line corresponds to the mean $\beta$-value of the input signals. The horizontal dotted line corresponds to the $\beta$-value of Unit 100, which is slightly higher than 90% of that of the input signals.

## Gabor-like optimal stimuli and phase invariance

The optimal stimuli of almost all units look like Gabor wavelets (Figure 3), in agreement with physiological data. This means that the units respond best to edgelike stimuli. The response of all these units is largely invariant to phase shift as illustrated by the lack of oscillations in the response images (Figure7, Figure 10, and Figure 11). To quantify this aspect we presented to each unit a sine grating tuned to the preferred orientation, frequency, speed, length, width, and position as revealed by $\mathbf{x}^+$ and computed the relative modulation rate $F_1/F_0$ (i.e., the ratio of the amplitude of the first harmonic to the mean response). Neurons are classified as complex if their $F_1/F_0$ ratio is less than 1.0; otherwise they are classified as simple, as defined in Skottun et al (1991). All units have a modulation rate considerably smaller then 1.0 (the maximum modulation rate is 0.16) and would thus be classified as complex cells in a physiological experiment. Out of 100 units, 98 had both a Gabor-like $\mathbf{x}^+$ and phase-shift invariance; the missing two cells are described in Tonic cells. (See Appendix A.2 for additional remarks.)

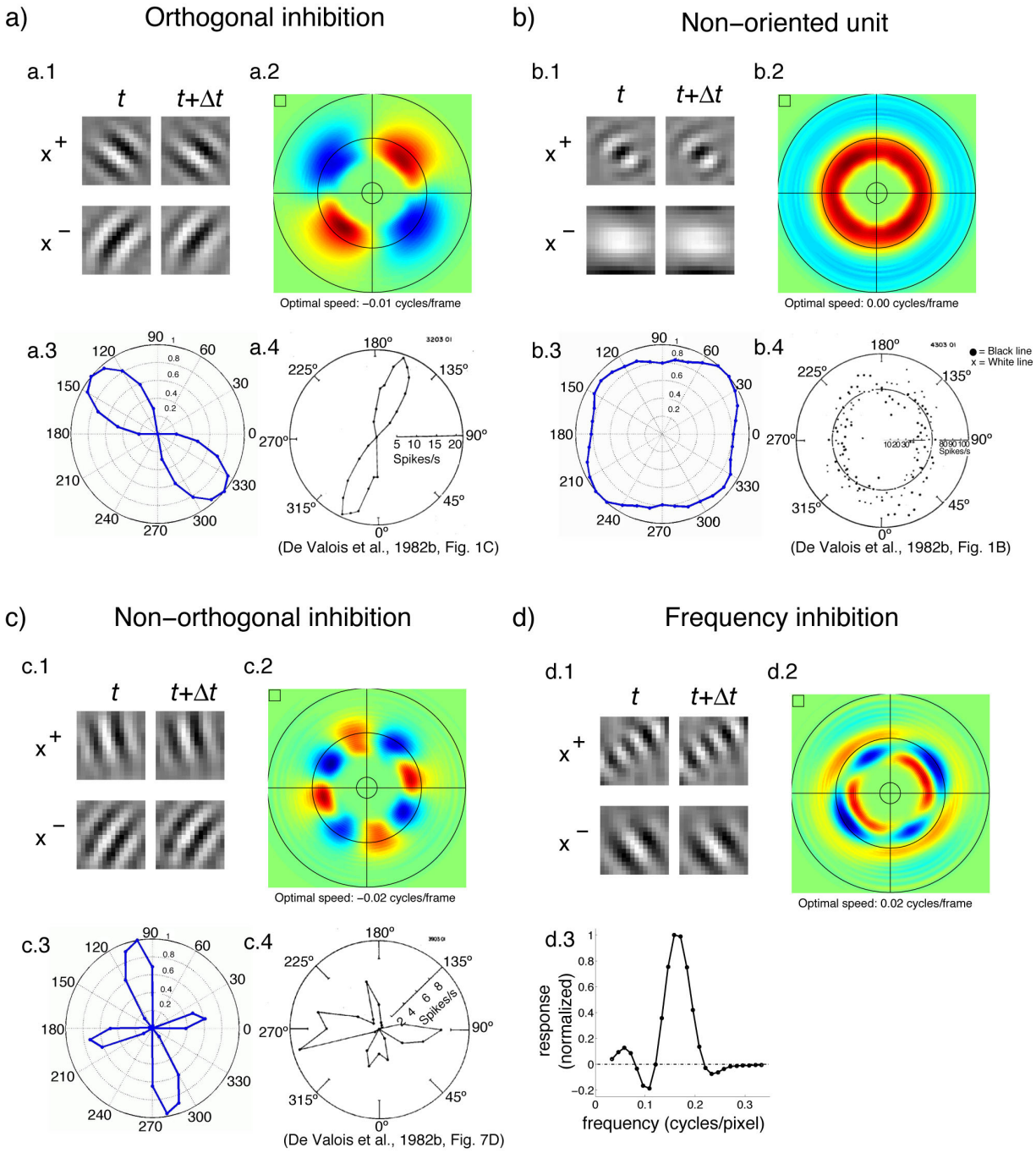



Figure 7. Active inhibition. This figure illustrates the ways in which active inhibition can shape the receptive field of a unit. Subfigures (a-d.1) show the optimal excitatory and inhibitory stimuli $\mathbf{x}^+$ and $\mathbf{x}^-$ of the considered units. Subfigures (a-d.2) show the response images corresponding to the circular test image (Figure 4a). The small square in the upper left corner represents the size of an input patch. Subfigures (a-c.3) show a polar plot of the response of the unit to a sine grating. The orientation of the grating varies along the angular direction, while the radial axis measures the response normalized to the maximum. In subfigures (a-c.4) we show for comparison equivalent plots showing the response in spikes/s of neurons of the primary visual cortex of the cat (De Valois, Yund et al., 1982). Subfigure (d.3) shows the normalized response of the unit to a sine grating of varying frequency. (a). This unit shows maximal inhibition at an orientation orthogonal to the optimal excitatory one while there is no inhibition in frequency. The unit has a relatively broad tuning to both orientation and frequency. (The cell shown in (a.4) has been classified as simple but seems to be representative for complex cells as well.) (b). Although this unit responds to edges (as shown by $\mathbf{x}^+$), the polar plot reveals that it is not selective for any particular orientation. The unit is thus classified as non-oriented. There is a slight inhibition at lower frequencies. (c). This unit has inhibitory flanks with an orientation near the preferred one. In such a case it is sometimes possible to observe a second peak of activity appearing at the orthogonal orientation, known in the experimental literature as secondary response lobe. (d). $\mathbf{x}^+$ and $\mathbf{x}^-$ of this unit have the same orientation but a different frequency. This results in a very sharp frequency tuning. On the other hand the unit responds to a broad range of orientations.

## Active inhibition (orientation and frequency tuning)

In the classical model, complex cells have no inhibition and are correspondingly restricted in their functional properties (e.g., see MacLennan, 1991). In physiological neurons, however, active inhibition is present and useful (e.g., to shape the tuning to orientation and frequency) (Sillito, 1975; De Valois, Yund et al., 1982). (See Appendix A.3 for additional remarks.)

In our model inhibition is present in most units and typically makes an important contribution to the output. As a consequence $\mathbf{x}^-$ is usually well structured and has the form of a Gabor wavelet as well (Figure 3). Its orientation plays an important role in determining the orientation-tuning. It can be orthogonal to that of $\mathbf{x}^+$ (Figure 7a), but it is often not, which results in sharpened orientation-tuning (because the response must decrease from a maximum to a minimum in a shorter interval along the orientation). On the other hand, we also find units that would be classified as complex by their response to edges and their phase invariance but which are not selective for a particular orientation (Figure 7b) (4 units out of 100). Similar cells are present in V1 and known as *nonoriented cells* (De Valois, Yund et al., 1982). Figure 8a compares the distribution of the orientation bandwidth of our units with that of complex cells reported in De Valois, Yund et al. (1982) and Gizzi, Katz, Schumer, and Movshon (1990). Our units seem to have a slightly broader tuning, but the difference is not significant (one-sided Kolmogorov-Smirnov test, $p > .8$).

When the orientation of $\mathbf{x}^-$ is very close to that of $\mathbf{x}^+$, it is possible to observe a second peak of activity at an orientation orthogonal to the optimal one (Figure 7c) known as *secondary response lobe*. Out of 100 units, 9 had this characteristic. (We classify a unit as having a secondary response lobe if on the orientation-tuning curve its output expressed in percentage of the maximal response decreases from 100% to less than 10% and then it increases again to more than 50% at the orthogonal orientation.) De Valois, Yund et al. (1982) repeatedly stress that non-orthogonal inhibition is common among neurons in V1 and show some example cells (one of which is shown in Figure 7c.4). The fraction of such cells in V1 is not reported. Ringach et al. (2002) quantitatively studied the relative angle between maximal excitation and suppression. Figure 8b shows the histograms of the angle between maximal excitation and

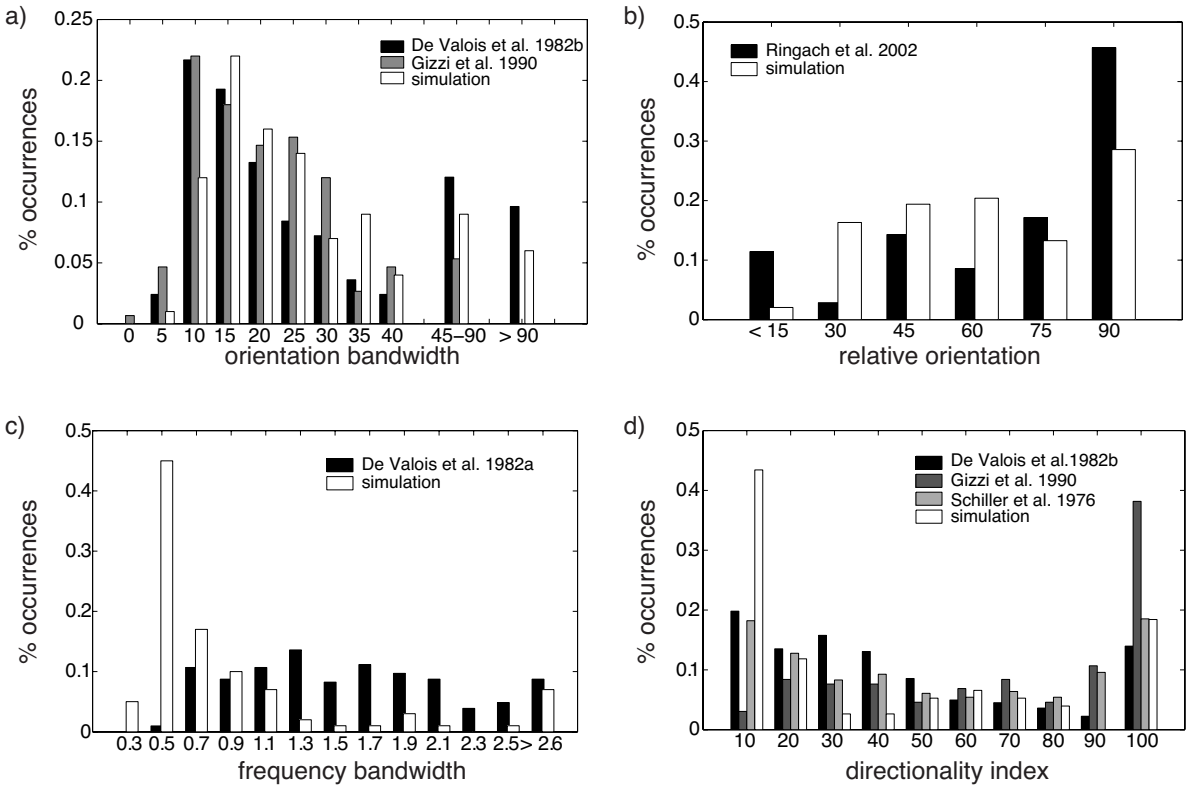



Figure 8. Population statistics. (a). The distribution of half-height orientation bandwidth in complex cells and in our simulation. (b). The distribution of the angle between the orientation of maximal excitation and maximal inhibition. The data from Ringach et al. (2002) contain simple cells as well as complex cells, which might explain the more pronounced peak at 90 deg. (c).The distribution of half-height frequency bandwidth in complex cells and in our simulation. The bandwidth is measured by the units' contrast sensitivity function as in De Valois, Albrecht et al. (1982). Units whose contrast sensitivity does not drop below 50% of the maximum are classified in the last bin. (d). The distribution of the directionality index of complex cells and in our simulation (the data from De Valois, Yund et al., 1982, also contain simple cells, but in the study it is stated that there was no significant difference between the two populations).

inhibition in our simulation and in their study. In the latter there is a more pronounced peak at orthogonal orientations. This difference might be due to the small fraction of complex cells in the population considered in the study (24 complex cells out of 75). To draw the histogram, the cells that responded to very low frequencies were discarded. (It was not specified how many complex cells remained in the final set.) To the extent simple cells are well described by linear functions, they are likely to show maximal suppression at 90 deg.

Analogous considerations also hold for the frequency domain: Here again the tuning varies from a sustained response within a range of frequencies (Figure 7a) to a sharp tuning due to active inhibition (Figure 7d). Figure 8c shows the distribution of frequency bandwidth in octaves in our simulation and in complex cells as reported by De Valois, Yund et al. (1982). The bandwidth was computed from the units' contrast sensitivity like in the cited study. Complex cells have a rather flat distribution, while the bandwidth of our units is concentrated between 0.3 and 1.4 octaves with a peak at 0.5. The reason for this difference is not yet entirely clear (but see Appendix A.4).

Figure 9 shows the joint distribution of frequency and orientation bandwidth in V1 in our simulation and as reported by De Valois, Albrecht et al. (1982). Because the marginal distribution of frequency bandwidth is different, in our case the data points are more concentrated in the left part of the graph. However, the two distributions are similar in that they have a large scatter and no strong correlation between orientation and frequency bandwidth. (The data from De Valois, Albrecht et al., 1982, also contains simple cells. The distribution of frequency and orientation bandwidth was found to be similar in simple and in complex cells (De Valois, Albrecht et al., 1982; De Valois, Yund et al., 1982), but the correlation of the two variables within the two groups is not reported.)

## End- and side-inhibition

Some of the complex cells in V1 are selective for the length (*end-inhibited* cells) or width (*side-inhibited* cells) of their input. While in normal cells the extension of a grating at the preferred orientation and frequency produces an increase of the response up to a saturation level, in these cells the response drops if the grating extends beyond a certain limit (DeAngelis et al., 1994).

End- and side-inhibition are present also in our simulation (Figure 10). We computed for each unit a quantitative measure of its degree of end- or side-inhibition by presenting sine gratings of different length and width (keeping all other parameters equal to the preferred ones). We define the end- and side-inhibition index as in DeAngelis et al. (1994) by the decrease of the response in percentage between optimal and asymptotic length and width, respectively. Out of 100 units, 10 had an end-inhibition index greater than 20%, and 7 units of 100 had a side-inhibition index greater than 20%. In contrast to DeAngelis et al. (1994) we found only 2 units that showed large (>20%) end- and side-inhibition simultaneously.

End- and side-inhibited units can sometimes be identified by looking directly at the optimal stimuli. In these cases $\mathbf{x}^+$ fills only one-half of the window while the missing half is covered by $\mathbf{x}^-$ with the same orientation and frequency (Figure 10a.1 and b.1). In this way, if we extend the stimulus into the missing half, the output receives an inhibitory contribution and drops. This receptive field organization is compatible with that observed in V1 by Walker et al. (1999) in that inhibition is asymmetric and is tuned to the same orientation and frequency as the excitatory part.

A secondary characteristic of end- and side-inhibited cells in V1 is that they are sometimes selective for different signs of curvature (Dobbins et al., 1987; Versavel et al., 1990). This can be observed in our simulation (e.g., in Figure 10b.2 where the dashed circles indicate two opposite curvatures). One of them causes the unit to respond strongly while the other one inhibits it.

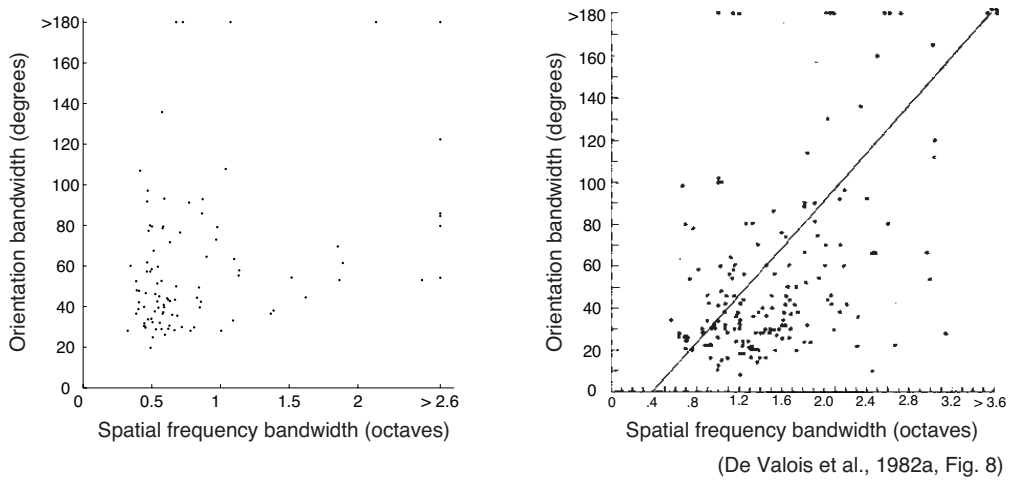
## Direction selectivity

Complex cells in V1 are sensitive to the motion of the presented stimuli. Some of them respond to motion in both directions while others are direction-selective (Hubel & Wiesel, 1962; Schiller et al., 1976a; De Valois, Yund et al., 1982; Gizzi et al., 1990). Similarly, in our model some units are strongly selective for direction (Figure 11) while others are neutral. In the latter case the optimal speed may be non-zero but the response is nearly equal for both directions.

We measure direction selectivity by the *directionality index* given by $DI = (1 - R_{np}/R_p) \cdot 100$ with $R_p$ and $R_{np}$ being the response in the preferred and in the nonpreferred direction, respectively (Gizzi et al., 1990). The index is 0 for bidirectional units and 100 for units that respond only in one direction of motion. Figure 8d shows the histogram of *DI* in our simulation compared to three distributions from the physiological literature. The distributions are quite similar, although there is a more pronounced peak for bidirectional units. (See Appendix A.5 for additional remarks.)

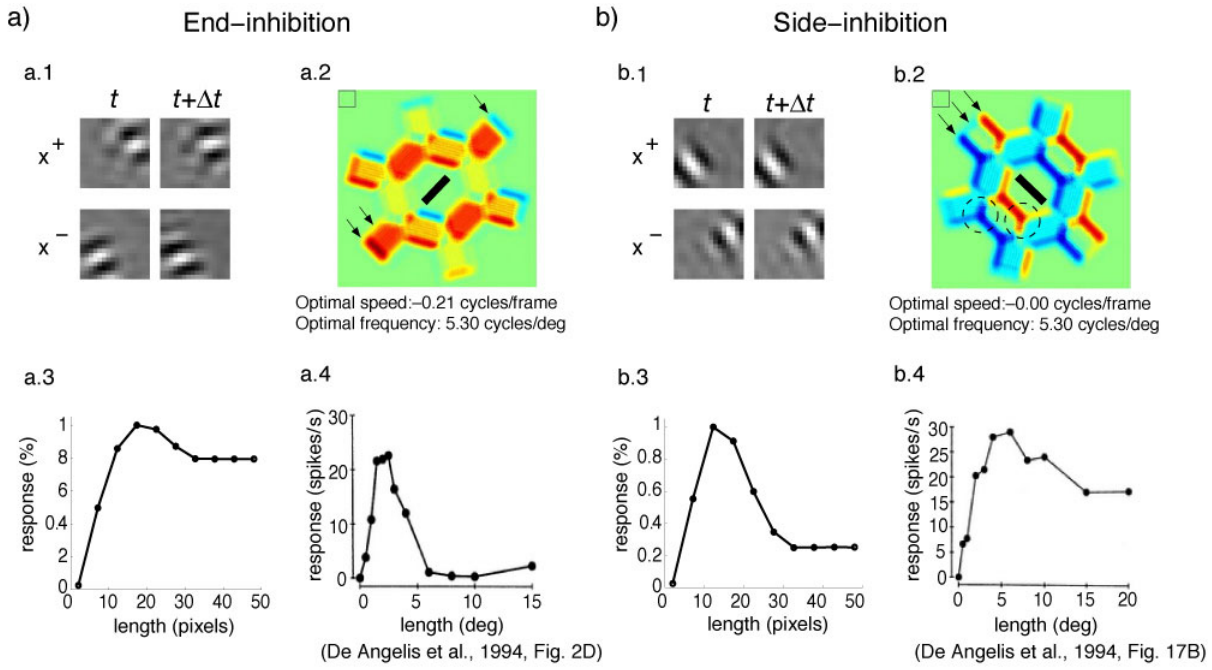
## Tonic cells

The first unit in every simulation codes for the mean pixel intensity and is thus comparable to the *tonic cells* found in V1 (Schiller et al., 1976a). Tonic cells respond to either bright or dark stimuli but do not need a contour to respond. We find in addition one unit (the second one) that responds to the squared mean pixel intensity and therefore to both bright and dark stimuli. (See Appendix A.6 for additional remarks.)




(De Valois et al., 1982a, Fig. 8)

Figure 9. Frequency and orientation bandwidth. This figure compares the joint distribution of frequency and orientation bandwidth in our simulation (left) and in De Valois, Albrecht et al. (1982) (right). The two distributions are similar in that they have a large scatter and no strong correlation between orientation and frequency bandwidth. (The data set on the right contains simple and complex cells. The distribution of orientation bandwidth was found to be similar for both classes, but the correlation of the two variables within the two groups was not reported.)




(De Angelis et al., 1994, Fig. 2D) (De Angelis et al., 1994, Fig. 17B)

Figure 10. End- and side-inhibition. This figure illustrates end-and side-inhibition in our simulation. Subfigures (a-b.1) show the optimal stimuli $\mathbf{x}^+$ and $\mathbf{x}^-$ of the considered units. Subfigures (a-b.2) show the response image corresponding to a hexagonal test image (Figure 4b). The small square in the upper left corner represents the size of an input patch. Subfigures (a-b.3) show the response of the unit to a sine grating with varying length or width, respectively. For comparison, equivalent plots of the response in spikes/s of end- and side-inhibited complex cells published in DeAngelis et al. (1994) are shown in subfigures (a-b.4). The four curves (a-b.3 and a-b.4) have similar shapes and mainly differ in their inhibition index (the ratio between maximal and asymptotic response), which has a broad distribution in all four cases. (a). This unit is end-inhibited. The optimal stimuli indicate how the receptive field is organized: the optimal excitatory stimulus fills only one half of $\mathbf{x}^+$ while the missing half is in $\mathbf{x}^-$. Inhibition is thus asymmetric and tuned to the same orientation and frequency as excitation, in agreement with Walker et al. (1999). From left to right, the first arrow in the hexagonal response image indicates the point of maximal response, when only the right part of the input window is filled. In the region indicated by the second arrow, the grating fills the whole input window; the response is decreased, which corresponds to end-inhibition. The third arrow indicates the region where only the left part of the input window is filled, and the unit is inhibited. (b). This is a side-inhibited unit. The interpretation of $\mathbf{x}^+$ and $\mathbf{x}^-$ and of the arrows is similar to that in (a), except that the grating extends laterally, which corresponds to side-inhibition. The two dashed circles surround two regions with opposite curvature. The unit responds strongly in one case and is inhibited in the other, which indicates curvature selectivity.

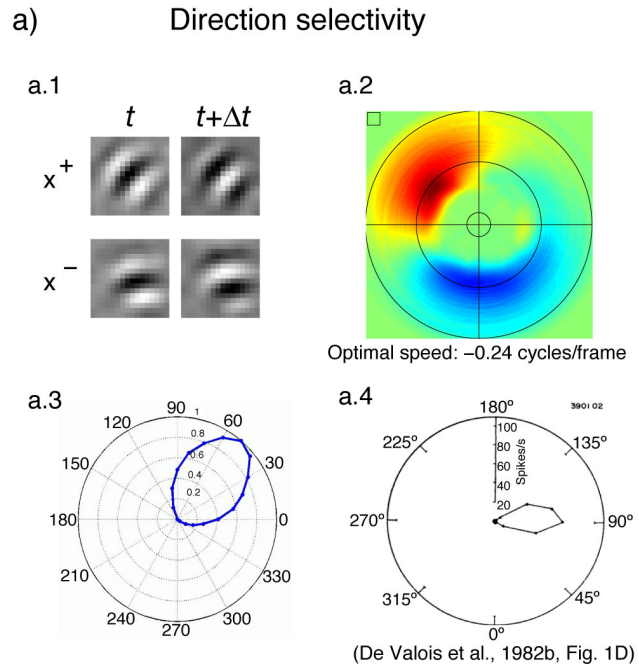



Figure 11. Direction selectivity. This figure shows a direction-selective unit. See the caption of Figure 7 for the description of the subfigures. The two wavelets in $\mathbf{x}^+$ at time $t$ and at time $t + \Delta t$ are identical except for a phase shift. This means that the unit responds best to a moving edge. This is confirmed by the response image, which shows different responses at opposite points of a ring. At those points the orientation is equal but the grating is moving in opposite directions.

## Response to complex stimuli

As can be inferred from the invariances (see Berkes & Wiskott, 2002), some units give a near-optimal output (>80% of the optimum) in response to corners and T-shaped stimuli and are related to the V1 cells described in Shevelev (1998). In a physiological experiment these stimuli could be classified as the optimal ones if the contrast instead of the energy is kept constant (e.g., a T-shaped stimulus has a larger energy than a bar with the same length and contrast). Other units respond to one sign of curvature only. These behaviors are often associated with end- or side-inhibition, as described above. In a few cases the two wavelets in the optimal stimuli have a slightly different orientation or frequency at time $t$ and at time $t + \Delta t$, which indicates a more complex behavior in time, such as selectivity for rotation or zoom.

## Relations between slowness and behavior

Although the precise shape and order of the units can vary in different simulations, there seem to be relations between the slowness of unit responses and the receptive field properties. The slowest units are usually less selective for orientation and frequency, have orthogonal inhibition, and their preferred speed is near zero. Units with non-orthogonal inhibition, direction selectivity, and end- or side-inhibition predominate in a faster regime. It would be interesting to see if similar relations can also be found experimentally by comparing the temporal variation of the response of a neuron stimulated by natural scenes and its receptive field properties. We could not find any relation between preferred orientation and slowness.

# 5. Control experiments

We performed a set of control experiments to assess the role of spatial transformations, the statistics of the input images, dimensionality reduction, and asymmetric decorrelation in our results.

## 5.1 Control experiment 1

In this set of experiments we investigated the role of the three spatial transformations used in our simulation: translation, rotation, and zoom. The settings for these experiments are identical to those of the main simulation described in Section 4 except that to achieve a reasonable total simulation time the input vectors consisted of single frames (instead of pairs of consecutive frame). The input dimension is reduced accordingly to $N = 50$, so that the proportion betweens input and reduced dimensions is the same as in the main simulation. We analyzed the first 50 units for each experiment. We performed a first simulation to be used as a reference using all three spatial transformations, followed by six simulations where only one or two transformations where used: translation only, rotation only, zoom only, translation and rotation, translation and zoom, and rotation and zoom. The parameters used for the transformations are the same as in the main simulation.

Figure 12a-h shows the optimal excitatory stimuli of all seven simulations. It can be seen that optimal stimuli similar to Gabor wavelets appear only in simulations with translation, including the one with only translation. Sine-grating experiments show that all units have phase-shift invariance (all relative modulation rates are smaller than 0.27). Translation is thus a necessary and sufficient spatial transformation to obtain complex cell receptive fields from natural images with SFA. On the other hand, the optimal stimuli in the simulation with only translation seem to occupy the whole patch in contrast to the more localized optimal stimuli in the simulations where translation is combined with the other transformations. Zoom and especially rotation are necessary to obtain more localized receptive fields. (This is not directly evident from Figure 12 because of the small number of optimal stimuli shown. Images containing more optimal stimuli can be found online; see Additional Material.) In simulations including translation but no rotation, the distribution of orientation bandwidth is skewed toward small bandwidths. High bandwidths therefore seem to be a consequence of the amount of rotation included in the

simulation, as one would expect because it results in an improved tolerance to changes in orientation.

Functions learned with rotation only and with both rotation and zoom show optimal stimuli with a circular structure (cf. Kohonen, Kaski, & Lappalainen, 1997). Functions learned with zoom only have optimal stimuli with small white/black spots in the center of the receptive field. These two last receptive field structures are not found in V1.

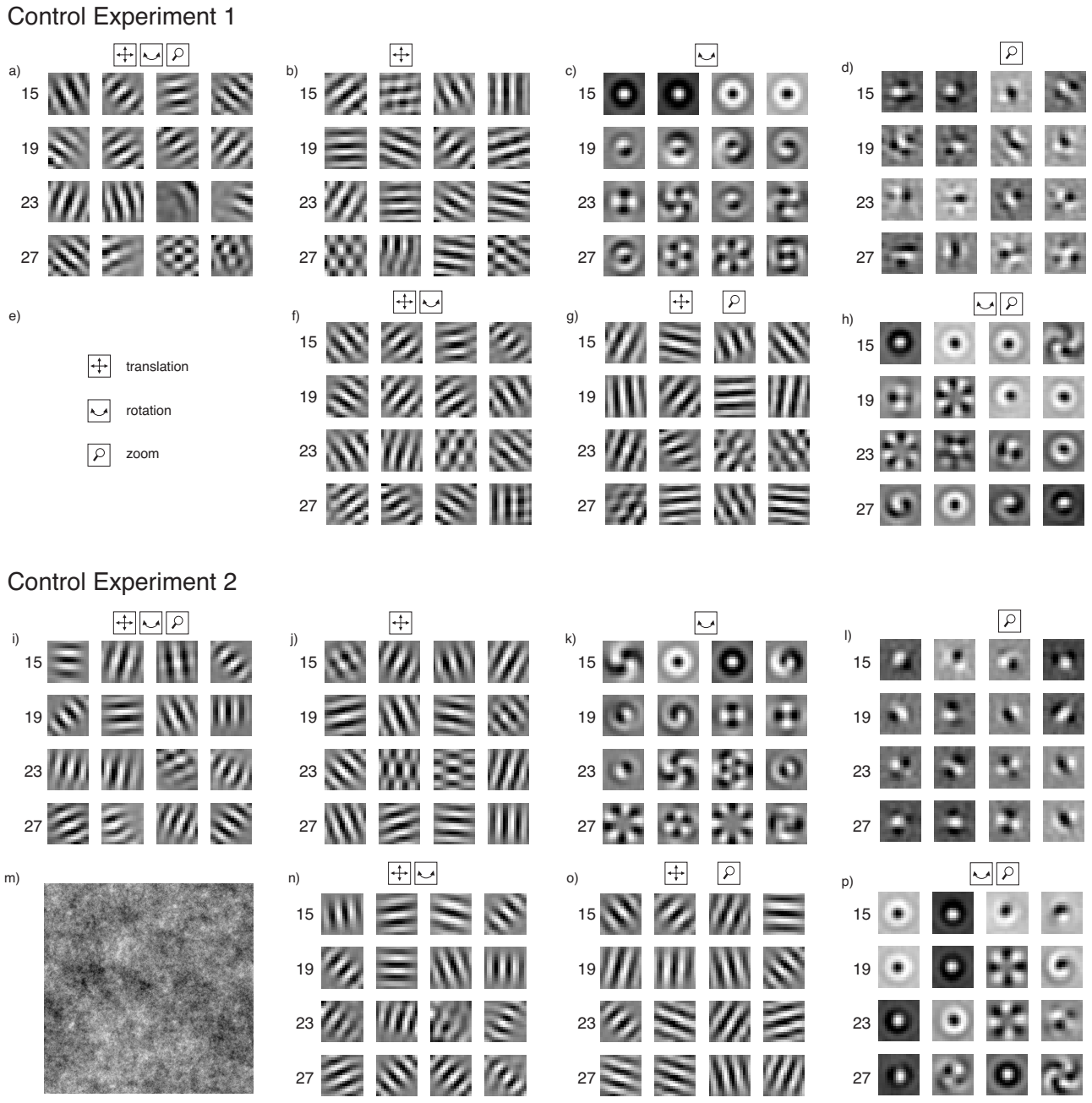



Figure 12. Control Experiments 1 and 2. This figure shows the optimal excitatory stimuli of Units 15-30 for Control Experiments 1 and 2. One or more icons representing the spatial transformations applied in a particular experiment are displayed on the top of each plot. A legend for the icons can be found in (e). (a-h). Optimal excitatory stimuli for Control Experiment 1. In this set of experiments we investigated the role of the three spatial transformations used in our simulation. The input image sequences were constructed from natural images like those in the main simulation, and all combinations of one, two, or three spatial transformations were applied. The results show that translation is a necessary and sufficient spatial transformation to obtain complex cell characteristics. (i-p). Optimal excitatory stimuli for Control Experiment 2. In this set of experiments we investigated the role of the spatial statistics of natural images in our results. The input images were replaced by colored noise images with a power spectrum equal to that of natural images. The results suggest that spatial second-order statistics are sufficient to obtain complex cell characteristics.

## 5.2 Control experiment 2

In this set of experiments we investigated the role of the spatial statistics of natural images in our results. The settings are identical to those of Control Experiment 1 except that instead of natural images we used colored noise images with a $1/f^2$ power spectrum, similar to the one shown in Figure 12m. The statistical properties of such images are equivalent to those of natural images up to the second order (Ruderman & Bialek, 1994; Dong & Atick, 1995). For each experiment we generated 36 new noise images that replaced the natural ones. The results of these experiments were almost identical to those of Control Experiment 1. The experiments including translation show Gabor-like optimal stimuli (Figure 12i-p) and phase-shift invariance. All relative modulation rates are smaller than 0.14 except for two units (one in the experiment with all transformations and one in the experiment with translation and rotation), which have a modulation rate of 1.47 and 1.53, respectively. The distributions of the various parameters are very similar to those obtained in Control Experiment 1. This suggests that spatial second-order statistics are sufficient to learn complex cell receptive fields. In principle our model considers spatial statistics up to the fourth order because the matrices **A** and **B** (Equations 16 and 17) contain products of monomials of degree 2.

## 5.3 Control experiment 3

This experiment was performed to exclude an influence of the dimensionality reduction on our results (see also Appendix A.1). The settings of the simulation are identical to those used in the main simulation except that the input vectors consisted in single frames only and, most importantly, that the input patches were 10×10 pixels large and no dimensionality reduction was performed. The translation speed was reduced by 10/16th, so that the proportion with the patch size was preserved. The first 100 units were analyzed.

The first two units were classified as tonic units. All other units have Gabor-like optimal stimuli (Figure 13) and phase-shift invariance (maximum modulation rate 0.23), excluding Units 9 and 10, which are described below. Units 4 and 11 have checkerboardlike optimal excitatory stimuli. Further analysis reveals that they are nonoriented units that only respond to very high frequencies. Units 9 and 10 have optimal excitatory stimuli with one bright region in a corner. The optimal inhibitory stimuli are similar, but their bright corner is opposite to the excitatory one. The role of these units is unclear, but it is possible that they give a nonlinear response to a luminance gradient along the diagonal. This experiment shows that the learning of complex cell receptive fields is not a consequence of the dimensionality reduction step.

## 5.4 Control experiment 4

In our mathematical formulation of the slowness principle the units are learned "one after the other" (Constraint 4) in the sense that in an online implementation of SFA the units would be learned using an asymmetric decorrelation term (i.e., unit $j$ would be adapted to optimize Equation 1 and to be decorrelated to all units $i$ with $i < j$). In a biological system it might seem more realistic to use symmetric decorrelation, where each unit is adapted to optimize Equation 1 and to be decorrelated to all other units.

In this control experiment we relax Constraint 4 and mix the units of the main simulation by an orthonormal transformation. The resulting units still satisfy Constraints (2-4) and span the slowest subspace (i.e., they minimize the *total* variance of the derivative in a 100-dimensional subspace). However, the asymmetry that is inherent in the algorithm and induces the order is no longer effective, so none of the units is distinguished over the others anymore.

All resulting units have Gabor-like optimal stimuli and phase-shift invariance (maximum modulation rate 0.37) and would thus be classified as complex cells. However, their response images are more unstructured than in the main simulation and they sometimes show a few peaks at different orientations and frequencies, which is not consistent with the behavior of complex cells (e.g., see the plots reported in Ringach et al., 2002). Moreover, the distribution of orientation bandwidth is skewed toward small bandwidths, and in the distribution of the relative angle between excitation and suppression the peak at 90 deg is missing and there is a maximum at 45 deg. It thus seems that asymmetric decorrelation is necessary to obtain the more structured results found in the main simulation. Note, however, that every breaking in the symmetry of decorrelation would make the units converge to the asymmetric solution. Because perfect symmetry is difficult to enforce in a biological system, the asymmetric case might be more realistic.

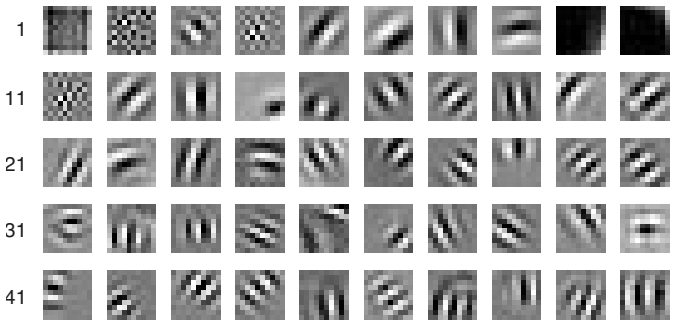



Figure 13. Control Experiment 3. This figure shows the optimal excitatory stimuli of Units 1-50 for Control Experiment 3. This experiment was performed without dimensionality reduction to exclude a possible influence on the main simulation's results. The results show that the learning of complex cell receptive fields is not a consequence of the dimensionality reduction step.

# 6. Discussion

In this work we have shown that SFA applied to natural image sequences learns a set of functions that have a good qualitative and quantitative match with the population of complex cells in V1. In the following section we discuss other theoretical models of complex cells. In Section 6.2 we discuss the properties of the chosen function space, present a neural network architecture equivalent to a given polynomial of degree 2, and compare it with the neural networks used in other studies. We conclude with some remarks about other learning rules.

## 6.1 Other theoretical studies

Several theoretical studies have successfully reproduced the basic properties of simple cells (Olshausen & Field, 1996; Bell & Sejnowski, 1997; Hoyer & Hyvärinen, 2000; Szatmáry & Lõrincz, 2001; Olshausen, 2002; Einhäuser, Kayser, Körding, & König, 2002; Hurri & Hyvärinen, 2003a) or complex cells (Hyvärinen & Hoyer 2000; Körding, Kayser, Einhäuser, & König, 2004) in models based on the computational principles *sparseness* (Olshausen & Field, 1996; Olshausen, 2002), *statistical independence* (Bell & Sejnowski, 1997; Hoyer & Hyvärinen, 2000; Hyvärinen & Hoyer, 2000; Szatmáry & Lõrincz, 2001), or *slowness* (Einhäuser et al., 2002; Hurri & Hyvärinen, 2003a; Körding et al., 2004). Among the simple cell models, two included direction selectivity (Szatmáry & Lõrincz, 2001; Olshausen, 2002), two color selectivity (Hoyer & Hyvärinen, 2000; Einhäuser et al., 2002), and one disparity (Hoyer & Hyvärinen, 2000). Most of these models focused on one particular aspect of the behavior of cells in V1. In particular, the two complex cell models (Hyvärinen & Hoyer, 2000; Körding et al., 2004) learned units that were equivalent to the classical model and thus reproduced only the Gabor-like receptive fields and the phase-shift invariance. One important limitation of these models was that they assumed linear or nonlinear but simple neural network architectures that belong to a function set much smaller than the one we consider (see Section 6.2). None of the nonlinear models included inhibition while many of the illustrated complex cell behaviors are impossible to obtain without it.

Hashimoto (2003) learned quadratic forms (without the linear term) from natural image sequences using three computational principles: independent component analysis (ICA), a gradient descent variant of SFA, and an objective function that maximizes the sparseness of the derivatives of the output. In the experiments performed using ICA, Hashimoto obtained a set of units corresponding to the squared output of simple cells. The results obtained with the gradient descent variant of SFA showed a few units with complex cell properties; most of them were not structured. Although it is difficult to make a direct comparison because only the largest eigenvectors of two of the quadratic forms are reported, the results are in contradiction with the ones presented in this work. It is possible that the size of the input patches (8×8 pixel) was too small in comparison with the transformations of the image sequences that were used, so that two consecutive frames would have had almost no correlation. It is also possible that the gradient descent converged to a local minimum. In this respect it would be interesting to compare the $\beta$-value of the quadratic forms. The experiments performed by maximizing the sparseness of the derivatives learned some units with complex cell properties (including a few with structured inhibition) and others with the characteristics of squared simple cells. These results seemed in general more structured than the ones obtained with the SFA variant. It would be interesting to explore the relation between these two objective functions further.

The studies mentioned up to now learned visual receptive fields directly from the pixel intensities of natural images or movies. Zetzsche and Röhrbein (2001), on the other hand, considered as an input the response of a set of artificial simple cells (i.e., linear Gabor wavelets, whose outputs were split into an ON (positive) and an OFF (negative) pathway by half-wave rectification). The two pathways were then used as an input to PCA or to ICA. The main result of the study is that PCA applied to the output of Gabor wavelets having the same orientation and frequency but different positions in the visual field learns units with simple and others with complex cell characteristics. Additionally, some units of both classes showed end-inhibition. Another experiment performed by applying ICA to the output of Gabor wavelets with same orientation, different positions, and even- and odd-symmetric filter properties produced simple cells, some of which were end-inhibited and others side-inhibited. It is known that the Gabor wavelets used to form the first layer can be learned directly from natural images (e.g., by ICA), so that these results could in principle be obtained directly from pixel intensities. In this case, however, an additional criterion should be provided to group the resulting wavelets, so that only the ones with equal orientation and frequency are connected to a second-layer unit.

To our knowledge the model presented here is the first one based directly on input images that is able to learn a population of units with a rich repertoire of complex cell properties, such as active inhibition, secondary response lobes, end-inhibition, side-inhibition, direction selectivity, tonic cell behavior, and curvature selectivity.

## 6.2 Function space and equivalent network architecture

We performed SFA on the function space F of all polynomials of degree 2 mainly because of limited computational resources. In principle one would like to consider a function space as large as possible. On the other hand, neurons have computational constraints, too, and thus considering too large a set could lead to unrealistic results. This leads to an interesting question: Which functions of its input can a neuron compute? In other words, in which function space does the input-output map of a neuron lie?

Lau, Stanley, and Dan (2002) have fitted the weights of a nonlinear two-layer neural network to the output of complex cells. They found that the relation between the linear output of the subunits and the output of the complex cell is approximately quadratic (mean exponent $2.3 \pm 1.1$). This result describes which functions the neurons do compute and not which ones it could compute, which would determine the function space. It suggests, however, that considering the space of polynomials of degree 2 might be sufficient. Kayser, Körding, and König (2003) adapted the weights and the exponents of a neural network similar to the classical model of complex cells using an objective function based on the slowness principle (but see Section 6.3 for some remarks regarding the definition of slowness). The exponent of most of the units converged to 2. This experiment also suggests that quadratic nonlinearities might be an appropriate choice in our context. Polynomials of degree 2 also correspond to a Volterra expansion up to the second order of the spatio-temporal receptive field of a neuron (e.g., see Dayan & Abbott, 2001, Sect. 2.2) when time is discretized in small steps. Such an approximation has been used with some success to describe complex cells (e.g., Touryan, Lau, & Dan, 2002).

Polynomials of degree 2 are closely related to but at the same time much more general than the functions corresponding to the neural networks used in standard studies in the field (Hyvärinen & Hoyer, 2000; Hyvärinen & Hoyer, 2001; Körding et al., 2004; also see Section 6.1). Those models usually rely either on linear networks, which lie in the space of polynomials of degree 1, or on networks with one layer of a fixed number of linear units (2 to 25) followed by a quadratic nonlinearity (Figure 14b), which form a small subset of the space of polynomials of degree 2. This can be seen from the following considerations. Each polynomial of degree 2 can be written as an inhomogeneous quadratic form $g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{f}^T\mathbf{x} + c$, where $\mathbf{H}$ is an $N \times N$ matrix, $\mathbf{f}$ is an $N$-dimensional vector, and $c$ is a constant. For example for $N = 2$,

$$\begin{aligned} & w_1 x_1^2 + w_2 x_1 x_2 + w_3 x_2^2 \quad + w_4 x_1 + w_5 x_2 + c \\ &= \frac{1}{2}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \underbrace{\begin{pmatrix} 2w_1 & w_2 \\ w_2 & 2w_3 \end{pmatrix}}_{\mathbf{H}} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \underbrace{\begin{pmatrix} w_4 \\ w_5 \end{pmatrix}^T}_{\mathbf{f}} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + c \quad . \end{aligned} \tag{20}$$

As also noticed by Hashimoto (2003), for each quadratic form there exists an equivalent two-layer neural network, which can be derived by rewriting the quadratic form using its eigenvector decomposition:

$$\begin{aligned} g(\mathbf{x}) &= \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{f}^T\mathbf{x} + c \\ &= \frac{1}{2}\mathbf{x}^T\mathbf{V}\mathbf{D}\mathbf{V}^T\mathbf{x} + \mathbf{f}^T\mathbf{x} + c \\ &= \frac{1}{2}(\mathbf{V}^T\mathbf{x})^T\mathbf{D}(\mathbf{V}^T\mathbf{x}) + \mathbf{f}^T\mathbf{x} + c \\ &= \sum_{k=1}^{N} \frac{\mu_k}{2}(\mathbf{v}_k^T\mathbf{x})^2 + \mathbf{f}^T\mathbf{x} + c \quad , \end{aligned} \tag{21}$$

where $\mathbf{V}$ is the matrix of the eigenvectors $\mathbf{v}_k$ of $\mathbf{H}$ and $\mathbf{D}$ is the diagonal matrix of the corresponding eigenvalues $\mu_k$, so that $\mathbf{V}^T\mathbf{H}\mathbf{V}=\mathbf{D}$. One can thus define a neural network with a first layer formed by a set of $N$ linear subunits $s_k(\mathbf{x}) = \mathbf{v}_k^T\mathbf{x}$ followed by a quadratic nonlinearity weighted by the coefficients $\mu_k/2$. The output neuron sums the contribution of all subunits plus the output of a direct linear connection from the input layer (Figure 14a). Because the eigenvalues can be negative, some of the subunits give an

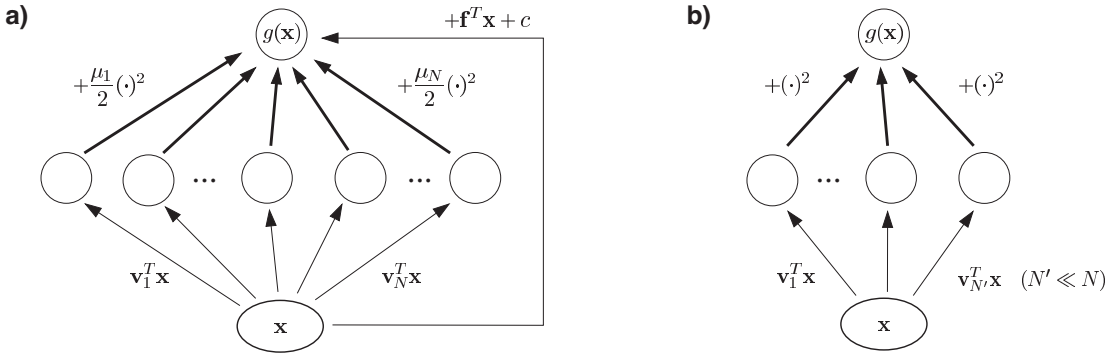



Figure 14. Equivalent neural network. (a). Neural network architecture equivalent to a polynomial of degree 2. The first layer consists of linear subunits whose outputs are squared and weighted. The output neuron on the second layer sums the contribution of all subunits and the output of a direct linear connection from the input layer. (The ellipse in the input layer represents a multidimensional input.) (b). Simpler neural network used in some theoretical studies. The output of the linear subunits is squared but not weighted and can give only an excitatory (positive) contribution to the output. There is no direct linear connection between input and output layer.

inhibitory contribution to the output. The weight vectors $\mathbf{v}_k$ are uniquely determined only if the eigenvalues of $\mathbf{H}$ are all different; otherwise the decomposition of $\mathbf{H}$ is arbitrary in the subspace that corresponds to the multiple eigenvalue. Moreover, the coefficients to the subunits are fixed only if one assumes that the weight vectors have unit norm.

The equivalent neural network clarifies the relation between the general case of polynomials of degree 2 (Figure 14a) and the smaller neural networks used in other modeling studies and described above (Figure 14b). An evident difference is the lack of a direct linear contribution to the output and the size of the networks: in our study each learned function has $N = 100$ subunits, which is much larger than the fixed numbers used in the studies mentioned above. The most important difference, however, is related to the normalization of the weights. In the theoretical studies cited above, the weights are normalized to a fixed norm and the activity of the subunits is not weighted. In particular, because there are no negative coefficients, no inhibition is possible.

The equivalent neural network shows that the choice of the space of all polynomials of degree 2 is compatible with the hierarchical organization of the visual cortex first proposed by Hubel and Wiesel (1962) in the sense that every learned function can be implemented by a hierarchical model similar to the energy model. The learning of the linear subunits would be modulated by the application of the slowness principle to the complex cell (see Section 6.3). According to this interpretation, the subunits would correspond to simple cells, and their receptive fields should thus look like Gabor wavelets. However, typically only a few subunits (those corresponding to the largest eigenvalues) are structured like simple cells (see also Berkes & Wiskott, 2005).

A possible alternative would be that simple cells are learned by a parallel computational principle and then grouped and weighted by the slowness principle to form complex cells. A similar distinct grouping step has been used in Zetzsche and Röhrbein (2001) and Hurri and Hyvärinen (2003b). Computational principles that have led to simple cells are sparseness, statistical independence, and slowness (see Section 6.1).

Although the function space of polynomials of degree 2 is mathematically attractive and has proved to be appropriate in experimental and theoretical studies as discussed above, it is not able to encompass all input-output nonlinearities of visual neurons. Divisive contrast gain control (Ohzawa, Sclar, & Freeman, 1982), saturation effects, and pattern adaptation are examples of nonlinear effects present in the visual cortex that cannot be realized.

## 6.3 Relation to other learning rules

We would like to point out that the definition of *slowness* in the models described in Einhäuser et al. (2002), Hurri and Hyvärinen (2003a), Körding et al. (2004), and in the present work are different to some extent. In Körding et al. (2004) the weights of neural networks equivalent to the classical model of complex cells are adapted by gradient descent to optimize a decorrelation and a slowness term. The slowness term is defined by the *mean* of the $\Delta$-values in Equation 1. If one fully enforces the decorrelation constraint (Equation 4), the units found by this rule lie in the subspace of the most slowly varying functions, but they are unique only up to an orthogonal transformation (i.e., by mixing the resulting functions through a rotation in the space of polynomials one would find different but equally optimal units). In the cited study, however, the architecture of the neural networks imposes additional constraints in the sense that the polynomials that the networks can compute form a subset and not a subspace of the space of polynomials of degree 2. This implies that an arbitrary rotation could lead to functions that do not lie in the subset and are thus not representable by such neural networks (K. P. Körding, personal communication, 2003). This argument shows that the two objective functions are different in some aspects.

In Einhäuser et al. (2002) and Hurri and Hyvärinen (2003a) the temporal variation of the output is minimized after a rectifying nonlinearity. When the nonlinearity is symmetric (e.g., when squaring or taking the absolute value of the output), solutions oscillating strongly around zero can be optimal because the sign is canceled out by the rectification. Also in the nonsymmetric case the solutions found are different from the ones extracted by SFA. Further investigations are needed to compare the different definitions and to find a unifying framework. In this context it is interesting to notice that in our model the learning rule at the level of the subunits is similar to the one proposed by Hurri and Hyvärinen (2003a) plus some cross-correlation terms. As shown in Appendix A.7, if we consider a neural network like that of Figure 14b and we expand the SFA objective function (Equation 1) at the level of the subunits, we obtain the equivalent objective function

$$\sum_{i=1}^{N'} s_i(t)^2 s_i(t-1)^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^{N'} s_i(t)^2 s_j(t-1)^2 , \qquad (22)$$

which has to be maximized ($s_i(t)$ is the activity of subunit $i$ at time $t$). The first term of Equation 22 is equal to the objective function proposed by Hurri and Hyvärinen (2003a) and is based on a computational principle related to temporal slowness and called temporal coherence. According to that principle the *energy* of the output has to be similar at

successive time steps. Hurri and Hyvärinen (2003a) showed that temporal coherence applied to natural video sequences learns simple cell receptive fields. The second term of Equation 22 maximizes the coherence of the energy of *different* subunits at successive timesteps. As a consequence, the subunits are encouraged to code for frequently occurring transformations of the features represented by the others. According to this analysis, it is tempting to conclude that temporal slowness at the level of complex cells modulates temporal coherence at the level of simple cells.

Temporal and spatial slowness are closely related concepts. For example, in our model, temporal slowness could be reformulated as a spatial one by adapting each unit to respond in a similar way to neighboring visual regions. The slowness objective could thus be reformulated as a spatial optimization criterion (Wiskott & Sejnowski, 2002). However, the former seems more natural to us and easier to implement in a biological system.

As mentioned above, another proposed computational principle is based on the sparseness of the output of a cell or on the independence of the outputs of a set of cells, which turns out to be equivalent in this context (Hyvärinen, Karhunen, & Oja, 2001, Chap. 21.2). Sparse codes are advantageous because they increase the signal-to-noise ratio, improve the detection of "suspicious coincidences," and allow effective storage in associative memories (Field, 1994). The sparseness of a code can be measured by its kurtosis, where higher kurtosis corresponds to a sparser code (Willmore & Tolhurst, 2001). Interestingly, the kurtosis of our units (mean kurtosis $12.85 \pm 3.46$) is much higher than that of their input (mean kurtosis $0.42 \pm 0.04$). This is due to the selectivity characteristics of the units. They can therefore take advantage of the benefits of a sparse representation without being explicitly optimized for it. Figure 15 shows an excerpt of the activity trace of a unit and the distribution of its output in response to 400,000 test frames. In a complex cell in V1 the activity would be half rectified at some threshold because the firing rate cannot be negative.

Statistical independence can be defined on the basis of higher order statistics, like in the studies cited above, or on the basis of second-order temporal statistics of the input signals. In this case, the correlation between different input signals at different time delays is minimized (Molgedey & Schuster, 1994; Belouchrani, Abed Meraim, Cardoso, & Moulines, 1997; Ziehe & Müller, 1998). Blaschke, Wiskott, and Berkes (2004) investigated the relation between SFA and second-order ICA and proved that in the linear case if only one time delay is considered both methods are equivalent.

## 6.4 Conclusion

In summary we have shown that slowness leads to a great variety of complex cell properties found also in physiological experiments. Our results demonstrate that such a rich repertoire of receptive field properties can be accounted for by a single unsupervised learning principle. Our results suggest that there is a relation between the behavior of a neuron and the slowness of its output. It will be interesting to see whether this prediction will be confirmed experimentally. Earlier modeling studies with SFA (Wiskott & Sejnowski, 2002) have shown that translation, scale, and other invariances can also be learned for whole objects in a hierarchical network of SFA modules: When trained with moving random 1D objects, such a network learns to represent the *what* and the *where* information (i.e., the identity

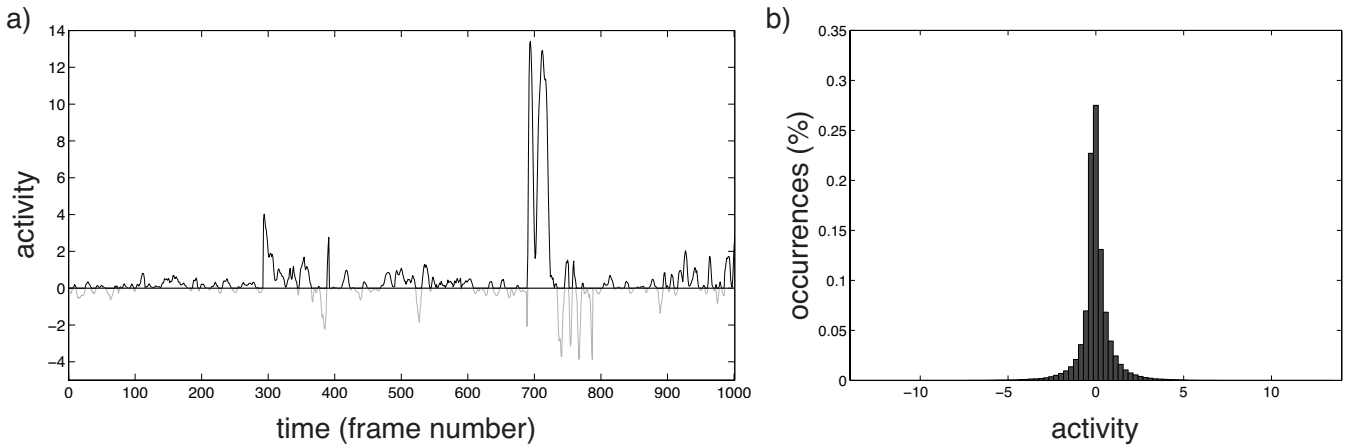



Figure 15. Unit activity. (a). Excerpt of the activity trace of Unit 5. In a complex cell in V1, the activity would be half rectified at some threshold because the firing rate cannot be negative. In the plot, the spontaneous activity level has been put to zero, and negative values corresponding to inhibition have been plotted in gray. (b). Distribution of the activity of Unit 5 in response to 400,000 test frames. The kurtosis of the output of this unit is 19.74, which is much higher than that of its input (mean kurtosis $0.42 \pm 0.04$).

and position) of novel objects in an invariant fashion, results that have been derived also analytically in Wiskott (2003). This suggests that slowness might be a rather general learning principle in the visual and possibly also other perceptual systems.

# Additional material

Additional material and software concerning the model described in this study are available at http://itb.biologie.hu-berlin.de/~berkes/slowness/slowness_index.shtml.

# Appendix A

This appendix contains additional notes on the more technical aspects of our model that might be useful to the theoretical reader but are not central to the main results of this study.

## A.1 Dimensionality reduction by PCA

In our model, the dimensionality of the input vectors is reduced by PCA. This corresponds to a low-pass filtering of the input patches because it is known that the principal components of natural images are linear filters of increasing frequency (Hancock, Baddeley, & Smith, 1992). The exact form of the filters learned in the PCA step, however, is completely irrelevant (and thus not shown), because SFA is independent of any linear transformation of its input. An arbitrary linear mix of the principal components would lead to identical results. Due to the self-similar structure of natural images (Ruderman & Bialek, 1994; Dong & Atick, 1995), it is in principle equivalent to work with low-pass filtered large patches or with small patches with no preprocessing. Large, low-pass filtered patches, however, are smoother and easier to analyze, especially in experiments with drifting sine gratings. In smaller patches, higher frequencies are represented as alternating positive and negative values. This raw sampling has undesired effects, especially for diagonal orientations, where the highest frequencies assume a checkerboardlike appearance whose orientation is often ambiguous. Moreover, the anisotropy due to the square shape of the pixels has more influence on measurements. Control Experiment 3 (Section 5.3) shows that there is no major qualitative difference between results obtained with large, low-passed filtered patches or with smaller unprocessed patches.

## A.2 Receptive field localization

The optimal stimuli are somewhat localized (Figure 3), especially for end- or side-inhibited units. However, their size is necessarily relative to that of the input patches (i.e., by making the input patches larger we would expect to obtain larger optimal stimuli), because there is nothing in the algorithm nor in the statistics of natural images that would define an absolute scale. This is analogous to what happens in the linear case for PCA, which also produces a set of filters extending over the whole image patches when applied to natural images. In contrast, the wavelets learned by independent component analysis (ICA) (e.g., Bell & Sejnowski, 1997) are more localized and do not scale with input patch size (but might depend on the resolution used, because the frequencies of the learned filters seem to cluster around the highest possible frequency). The difference between PCA and ICA suggests that if we would replace the decorrelation constraint (Equation 4) (like in PCA) with an independence constraint (like in ICA) we might expect to find more localized filters with a fixed absolute scale.

## A.3 Inhibition in cortical neurons

The exact shape and tuning of inhibition in cortical neurons are usually difficult to determine experimentally from the firing rate, which cannot be negative. Experiments studying inhibition must rely on the membrane potential, increase the neuron's firing by superimposing a "conditioning stimulus," or block inhibition with pharmacological manipulations. Each of these methods has specific drawbacks. For example, adaptation to the conditioning stimulus influences the orientation-tuning of the neurons (Dragoi, Rivadulla, & Su, 2001). Ringach et al. (2002) applied a new reverse-correlation technique and found that suppression and enhancement have similar magnitudes and that peak enhancement tends to be slightly larger than suppression. This is compatible with our results (note that peak enhancement is larger by construction). They also showed a positive correlation between suppression and orientation selectivity. Because they used nonlocalized, oriented stimuli, it is impossible to say if inhibition was oriented or localized. It was also not possible to make precise statements on the feedback/feedforward or broadly/narrowly tuned structure of inhibition, although it was found to be compatible with a tuned, feedforward, additive inhibition like the one present in our model. Walker et al. (1999) showed that the inhibitory part of end- and side-inhibited cells in V1 is localized and oriented, which is compatible with our results (Section 4). In the two cells reported in Walker et al. (1999), the inhibitory part has also the same orientation- and frequency-tuning as the excitatory part.

## A.4 Frequency-tuning, digitalization, and dimensionality reduction

The difference between the two distributions of frequency bandwidths in Figure 8c might be partly due to digitalization and dimensionality reduction. Our input patches are 16×16 pixels large, which means that the maximal bandwidth of our units is 3 octaves (from 1 to 8 cycles/patch, somewhat more on the diagonal). However,

we reduced the number of input dimensions to 100 for both time steps using PCA. There are thus about 50 components per input patch (assuming that the two patches are independent). Because the principal components of natural images are linear filters of increasing frequency (see Appendix A.1), we are only considering the 7×7 central Fourier components (because we have $50 = 7 \cdot 7 + 1$ principal components). This corresponds to frequencies from 1 to 4 cycles/patch and thus to at most 2 octaves. The actual bandwidth would in general be much smaller because to reach the theoretical limit the response of a unit at the two extreme frequencies of 1 and 4 cycles/patch would need to be exactly half of the maximum response. Simulations with a higher number of input components could yield a broader distribution.

### A.5 Direction selectivity and velocity distribution

We observed in other simulations (data not shown) that the distribution of the direction selectivity index depends on the distribution of velocities in the input sequences. Direction selectivity disappears for a velocity distribution that includes mostly very small translations and increases if it is skewed toward larger translations. A better estimation of the real-world distribution of velocities (both of the observer and of the environment) could improve the match between the histograms. One would perhaps need to increase the size of the input patches, because it limits the maximum velocity.

### A.6 Tonic cells

The first two units in our simulation code for the mean pixel intensity and for the squared mean pixel intensity. It might be argued that the first two units code for such simple features argues against the slowness principle, because they might seem "uninteresting" when compared with successive units described in Section 4. However, although simple, the features coded by the first two cells might be fundamental ones, just like the first terms in a Taylor expansion.

### A.7 Derivation of the relation to temporal coherence

As proved in Blaschke et al. (2004), if the first derivative is approximated by the time difference $\dot{y}_j(t) \approx y_j(t) - y_j(t-1)$, it is equivalent to minimize Equation 1 or to maximize the expression

$$\left\langle y_j(t) y_j(t-1) \right\rangle_t \quad , \tag{23}$$

because

$$\begin{aligned} \left\langle \dot{y}_j^2 \right\rangle_t &= \left\langle (y_j(t) - y_j(t-1))(y_j(t) - y_j(t-1)) \right\rangle_t \\ &= \left\langle y_j(t)^2 \right\rangle_t + \left\langle y_j(t-1)^2 \right\rangle_t \\ &\quad -2\left\langle y_j(t) y_j(t-1) \right\rangle_t \\ &= 2 - 2\left\langle y_j(t) y_j(t-1) \right\rangle_t \quad , \end{aligned} \tag{24}$$

where in the last step we applied the unit variance Constraint 3.

For a neural network like that shown in Figure 14b, we can express Equation 23 at the level of the subunits by expanding the output $y_j$:

$$\begin{aligned} \left\langle y_j(t) y_j(t-1) \right\rangle_t &= \left\langle \left( \sum_{k=1}^{N'} s_k(t)^2 \right) \left( \sum_{l=1}^{N'} s_l(t-1)^2 \right) \right\rangle_t \\ &= \left\langle \sum_{k,l=1}^{N'} s_k(t)^2 s_l(t-1)^2 \right\rangle_t \\ &= \left\langle \sum_{k=1}^{N'} s_k(t)^2 s_k(t-1)^2 \right\rangle_t \\ &\quad + \left\langle \sum_{\substack{k,l=1 \\ k \neq l}}^{N'} s_k(t)^2 s_l(t-1)^2 \right\rangle_t \end{aligned} \tag{25}$$

where in the last step we split the sum over all terms into one sum over all terms with equal indices and one sum over all terms with different indices. $N'$ is the number of the subunits.

As discussed in Section 6.2, the first term is equal to the objective function proposed by Hurri and Hyvärinen (2003a) and maximizes the correlation of the energy of the output of each subunit, whereas the second term maximizes the correlation of the energy of the output of different subunits.

## Acknowledgments

This work has been supported by a grant from the Volkswagen Foundation. We thank Reinhard Eckhorn, Jack Gallant, Peter König, and Dario Ringach for fruitful discussions. We also thank Tim Gollisch, Andreas Herz, and Martin Stemmler for useful comments on an earlier version of this manuscript.

Commercial relationships: none.
Corresponding author: Pietro Berkes.
Email: p.berkes@biologie.hu-berlin.de.
Address: Humboldt University Berlin, Institute for Theoretical Biology, Berlin, Germany.

## References


Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal Optical Society of America A, 2*, 284-299. [PubMed]

Becker, S., & Hinton, G. E. (1993). Learning mixture models of spatial coherence. *Neural Computation, 5*, 267-277.

Bell, A. J., & Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Research, 37*, 3327-3338. [PubMed]

Belouchrani, A., Abed Meraim, K., Cardoso, J.- F., & Moulines, E. (1997). A blind source separation technique based on second order statistics. *IEEE Transactions on Signal Processing, 45*, 434-444.

Berkes, P., & Wiskott, L. (2002). Applying slow feature analysis to image sequences yields a rich repertoire of complex cell properties. In J. R. Dorronsoro (Ed.), *Artificial neural networks (ICANN 2002 Proceedings)* (pp. 81-86). Berlin: Springer-Verlag.

Berkes, P., & Wiskott, L. (2005). On the analysis and interpretation of inhomogeneous quadratic forms as receptive fields. *Cognitive Sciences EPrint Archive* (CogPrints), 4081, http://cogprints.org/4081/. [Article]

Blaschke, T., Wiskott, L., & Berkes, P. (2004). *What is the relation between independent component analysis and slow feature analysis?* Manuscript submitted for publication.

Creutzfeldt, O. D., & Nothdurft, H. C. (1978). Representation of complex visual stimuli in the brain. *Naturwissenschaften, 65*, 307-318. [PubMed]

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge: MIT Press.

De Valois, R., Albrecht, D., & Thorell, L. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research, 22*, 545-559. [PubMed]

De Valois, R., Yund, E., & Hepler, N. (1982). The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research, 22*, 531-544. [PubMed]

DeAngelis, G., Freeman, R., & Ohzawa, I. (1994). Length and width tuning of neurons in the cat's primary visual cortex. *Journal of Neurophysiology, 71*, 347-374. [PubMed]

Dobbins, A., Zucker, S. W., & Cynader, M. S. (1987). End-stopped neurons in the visual cortex as a substrate for calculating curvature. *Nature, 329*, 438-441. [PubMed]

Dong, D. W., & Atick, J. J. (1995). Statistics of natural time-varying images. *Network: Computation in Neural Systems, 6*, 354-358.

Dragoi, V., Rivadulla, C., & Sur, M. (2001). Foci of orientation plasticity in visual cortex. *Nature, 411*, 80-86. [PubMed]

Einhäuser, W., Kayser, C., Körding, K., & König, P. (2002). Learning multiple feature representation from natural image sequences. In J. R. Dorronsoro (Ed.), *Artificial neural networks (ICANN 2002 Proceedings)* (pp. 21-26). Berlin: Springer-Verlag.

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

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation, 3*, 194-200.

Gantmacher, F. R. (1959). *Matrix theory* (Vol. 1). New York: AMS Chelsea Publishing.

Gizzi, M. S., Katz, E., Schumer, R. A., & Movshon, J. A. (1990). Selectivity for orientation and direction of motion of single neurons in cat striate and extrastriate visual cortex. *Journal of Neurophysiology, 63*, 1529-1543. [PubMed]

Hancock, P. J., Baddeley, R. J., & Smith, L. S. (1992). The principal components of natural images. *Network: Computation in Neural Systems, 3*, 61-70.

Hashimoto, W. (2003). Quadratic forms in natural images. *Network: Computation in Neural Systems, 14*, 765-788. [PubMed]

Hinton, G. (1989). Connectionist learning procedures. *Artificial Intelligence, 40*, 185-234.

Hoyer, P., & Hyvärinen, A. (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems, 11*, 191-210. [PubMed]

Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology, 160*, 106-154. [PubMed]

Hurri, J., & Hyvärinen, A. (2003a). Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation, 15*, 663-691. [PubMed]

Hurri, J., & Hyvärinen, A. (2003b). Temporal and spatiotemporal coherence in simple-cell responses: A generative model of natural image sequences. *Network: Computation in Neural Systems, 14*, 527-551. [PubMed]

Hyvärinen, A., & Hoyer, P. (2000). Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation, 12*, 1705-1720. [PubMed]

Hyvärinen, A., & Hoyer, P. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research, 41*, 2413-2423. [PubMed]

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis.* New York: Wiley-Interscience.

Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology, 58*, 1233-1257. [PubMed]

Kayser, C., Körding, K. P., & König, P. (2003). Learning the nonlinearity of neurons from natural visual stimuli. *Neural Computation, 15*, 1751-1759. [PubMed]

Kohonen, T., Kaski, S., & Lappalainen, H. (1997). Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM. *Neural Computation, 9*, 1321-1344.

Körding, K., Kayser, C., Einhäuser, W., & König, P. (2004). How are complex cell properties adapted to the statistics of natural scenes? *Journal of Neurophysiology, 91*, 206-212. [PubMed]

Lau, B., Stanley, G. B., & Dan, Y. (2002). Computational subunits of visual cortical neurons revealed by artificial neural networks. *Proceedings of the National Academy of Sciences U.S.A., 99*, 8974-8979. [PubMed]

MacLennan, B. (1991). Gabor representation of spatiotemporal visual images. *Technical Report CS-91-144*, Computer Science Department, University of Tennessee, Knoxville, TN.

Mitchison, G. (1991). Removing time variation with the anti-Hebbian differential synapse. *Neural Computation, 3*, 312-320.

Molgedey, L., & Schuster, G. (1994). Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters, 72*, 3634-3637. [PubMed]

Ohzawa, I., Sclar, G., & Freeman, R. (1982). Contrast gain control in the cat visual cortex. *Nature, 298*, 266-268. [PubMed]

Olshausen, B. (2002). Sparse codes and spikes. In R. P. N. Rao, B. A. Olshausen, & M. S. Lewicki (Eds.), *Probabilistic models of the brain: Perception and neural function.* Cambridge: MIT Press.

Olshausen, B., & Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*, 607-609. [PubMed]

O'Reilly, R. C., & Johnson, M. H. (1994). Object recognition and sensitive periods: A computational analysis of visual imprinting. *Neural Computation, 6*, 357-389.

Peng, H. C., Sha, L. F., Gan, Q., & Wei, Y. (1998). Energy function for learning invariance in multilayer perceptron. *Electronics Letters, 34, 292-294*.

Pollen, D., & Ronner, S. (1981). Phase relationship between adjacent simple cells in the visual cortex. *Science, 212*, 1409-1411. [PubMed]

Ringach, D. L., Bredfeldt, C. E., Shapley, R. M., & Hawken, M. J. (2002). Suppression of neural responses to nonoptimal stimuli correlates with tuning selectivity in macaque V1. *Journal of Neurophysiology, 87*, 1018-1027. [PubMed]

Ruderman, D. L., & Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Physical Review Letters, 73*, 814-817. [PubMed]

Schiller, P., Finlay, B., & Volman, S. (1976a). Quantitative studies of single-cell properties in monkey striate cortex. I. Spatiotemporal organization of receptive fields. *Journal of Neurophysiology, 39*, 1288-1319. [PubMed]

Schiller, P., Finlay, B., & Volman, S. (1976b). Quantitative studies of single-cell properties in monkey striate cortex. II. Orientation specificity and ocular dominance. *Journal of Neurophysiology, 39*, 1320-1333. [PubMed]

Schiller, P., Finlay, B., & Volman, S. (1976c). Quantitative studies of single-cell properties in monkey striate cortex. III. Spatial frequency. *Journal of Neurophysiology, 39*, 1334-1351. [PubMed]

Shevelev, I. A. (1998). Second-order features extraction in the cat visual cortex: Selective and invariant sensitivity of neurons to the shape and orientation of crosses and corners. *BioSystems, 48*, 195-204. [PubMed]

Sillito, A. (1975). The contribution of inhibitory mechanisms to the receptive field properties of neurons in the striate cortex of the cat. *Journal of Physiology, 250*, 305-329. [PubMed]

Skottun, C., De Valois, R., Grosof, D., Moshnov, J., Albrecht, D., & Bonds, A. (1991). Classifying simple and complex cells on the basis of response modulation. *Vision Research, 31*, 1079-1086. [PubMed]

Stone, J. V. (1996). Learning perceptually salient visual parameters using spatiotemporal smoothness constraints. *Neural Computation, 8*, 1463-1492. [PubMed]

Stone, J. V. (2001). Blind source separation using temporal predictability. *Neural Computation, 13*, 1559-1574. [PubMed]

Szatmáry, B., & Lõrincz, A. (2001). Independent component analysis of temporal sequences subject to constraints by lateral geniculate nucleus inputs yields all three major cell types of the primary visual cortex. *Journal of Computational Neuroscience, 11*, 241-248. [PubMed]

Touryan, J., Lau, B., & Dan, Y. (2002). Isolation of relevant visual features from random stimuli for cortical complex cells. *Journal of Neuroscience, 22*, 10811-10818. [PubMed]

van Hateren, J., & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London B, 265*, 359-366. [PubMed]

Versavel, M., Orban, G. A., & Lagae, L. (1990). Responses of visual cortical neurons to curved stimuli and chevrons. *Vision Research, 30*, 235-248. [PubMed]

Walker, G., Ohzawa, I., & Freeman, R. (1999). Asymmetric suppression outside the classical receptive field of the visual cortex. *The Journal of Neuroscience, 19*, 10536-10553. [PubMed]

Willmore, B., & Tolhurst, D. (2001). Characterizing the sparseness of neural codes. *Network: Computation in Neural Systems, 12*, 255-270. [PubMed]

Wiskott, L. (1998). Learning invariance manifolds. In L. Niklasson, M. Bodén, & T. Ziemke, (Eds.), *Proceedings of the International Conference on Artificial Neural Networks (ICANN 1998), Skövde* (pp. 555-560). London: Springer-Verlag.

Wiskott, L. (2003). Slow feature analysis: A theoretical analysis of optimal free responses. *Neural Computation, 15*, 2147-2177. [PubMed]

Wiskott, L., & Sejnowski, T. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation, 14*, 715-770. [PubMed]

Zetzsche, C., & Röhrbein, F. (2001). Nonlinear and extra-classical receptive field properties and the statistics of natural scenes. *Network: Computation in Neural Systems, 12*, 331-350. [PubMed]

Ziehe, A., & Müller, K.- R. (1998). TDSEP—an efficient algorithm for blind separation using time structure. *8th International Conference on Artificial Neural Networks (ICANN 1998)* (pp. 675–680). Berlin: Springer Verlag.