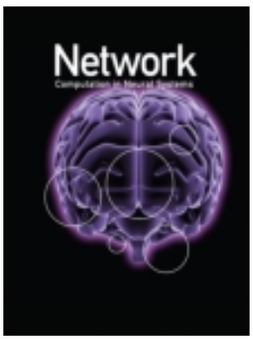

Network: Computation in Neural Systems

ISSN: 0954-898X (Print) 1361-6536 (Online) Journal homepage: http://www.tandfonline.com/loi/inet20

Hebbian learning in a model with dynamic rate-coded neurons: An alternative to the generative model approach for learning receptive fields from natural scenes

Fred H. Hamker & Jan Wiltschut

To cite this article: Fred H. Hamker & Jan Wiltschut (2007) Hebbian learning in a model with dynamic rate-coded neurons: An alternative to the generative model approach for learning receptive fields from natural scenes, Network: Computation in Neural Systems, 18:3, 249-266, DOI: 10.1080/09548980701661210

To link to this article: http://dx.doi.org/10.1080/09548980701661210

Network: Computation in Neural Systems
September 2007; 18(3): 249–266

informa
healthcare

Hebbian learning in a model with dynamic rate-coded neurons: An alternative to the generative model approach for learning receptive fields from natural scenes

FRED H. HAMKER & JAN WILTSCHUT

Department of Psychology and Otto-Creutzfeldt Center for Cognitive and Behavioral Neuroscience, Westf. Wilhelms-Universität Münster, 48149 Münster, Germany

(Received 13 March 2007; accepted 4 September 2007)

Abstract
Most computational models of coding are based on a generative model according to which the feedback signal aims to reconstruct the visual scene as close as possible. We here explore an alternative model of feedback. It is derived from studies of attention and thus, probably more flexible with respect to attentive processing in higher brain areas. According to this model, feedback implements a gain increase of the feedforward signal. We use a dynamic model with presynaptic inhibition and Hebbian learning to simultaneously learn feedforward and feedback weights. The weights converge to localized, oriented, and bandpass filters similar as the ones found in V1. Due to presynaptic inhibition the model predicts the organization of receptive fields within the feedforward pathway, whereas feedback primarily serves to tune early visual processing according to the needs of the task.

Keywords: *Natural scenes, network models, visual system, attention*

Introduction

Visual perception is thought to be based on a hierarchy of visual processing, where the complexity of the encoded feature properties grows with each level of increasing hierarchy. Theories of coding address the intriguing question of the kind

Correspondence: Fred H. Hamker, Allgemeine Psychologie Psychologisches Institut II, Westf. Wilhelms-Universität, Fliednerstrasse 21, 48149 Münster, Germany. Tel: +49 251-83 34171. Fax: +49 251-83 34173. E-mail: fhamker@uni-muenster.de

ISSN 0954-898X print/ISSN 1361-6536 online/07/030249–266
DOI: 10.1080/09548980701661210

of information encoded by the neurons. Since natural scenes are highly redundant, the core idea is to find a code which reduces the redundancy of images. In particular, efficient coding might be a fundamental constraint of visual processing (Barlow 1961; Atick and Redlich 1990; Nadal and Parga 1994). Typically, linear or approximately linear approaches are used to reduce the redundancy based on second or higher order statistics. The principal component analysis (PCA), independent component analysis (ICA), and independent factor analysis or sparse coding (IFA) have been successfully used to learn a set of basis functions (Hancock et al. 1992; Harpur and Prager 1996; Olshausen and Field 1996; Bell and Sejnowski 1997; van Hateren and van der Schaaf 1998). Especially ICA and IFA let the receptive fields converge to edge-filters, which exhibit similar properties as V1 cells. Despite this success there remain several open questions (Barlow 2001; Simoncelli 2003). From our point of view, one outstanding issue is the generalization of present approaches to higher levels of visual processing. Yet, there are only a few attempts to extend learning to model higher areas of visual processing (Rao and Ballard 1999; Hoyer and Hyvärinen 2002; Karklin and Lewicki 2003). Most of the present approaches are based on a linear generative model. Learning in this generative model grounds in the objective to minimize the error between the actual image and the predicted image (Olshausen and Field 1997; Hoyer 2003; Rehn and Sommer 2007), and typically only feedback connections are learned. In neural terms, the generative approach is analogous to an analysis–synthesis loop in which the feedback signal represents the predicted image. The residual image, the subtraction of the predicted image from the input image, is processed forward and the activity in the next (output) layer is relaxed to a stable, typically sparse representation. Explicit iterative feedforward/feedback processing has been used to learn receptive fields and, by enforcing a sparse representation, edge-filters emerge (Rao and Ballard 1999; Jehee et al. 2006).

However, the generative model approach appears to be difficult to reconcile with the idea of attention. As this might not be critical for early visual processing, higher brain areas tend to emphasize the behaviorally relevant aspects of the visual scene. The activity in higher brain areas would hardly allow to reconstruct the whole visual scene. Due to this potential limitation of the generative model approach, we started with the objective that learning should be embedded in the attentional dynamics of the network. This is inspired by the idea that attention is an emergent property of interactions between brain areas (Hamker 2005, 2006). In this framework of attention, predictive feedback implements a match enhancement (Grossberg 1980; Hamker 2004). Thus, feedback enhances the sensitivity of a neuron towards its input. This is an important difference to generative models, which use predictive feedback to compute the residual error of the prediction and the present input. While there are many studies of learning using the generative model, the match enhancement model has not been used for learning. However, our work benefits from several previous studies of Hebbian learning (von der Marlsburg 1973; Sejnowski 1977; Oja 1982; Linsker 1986). We advanced a Hebbian learning framework using presynaptic inhibition (Spratling and Johnson 2002). This learning rule has shown superior properties compared to other Hebbian learning frameworks on variations of the bar-learning task (Wiltschut et al. in preparation). We demonstrate that Hebbian learning in the match enhancement model with presynaptic inhibition not only leads to sparse representations and edge-like

receptive fields but also shows interesting coding properties such as the context dependence of a neuron's receptive field.

A model of Hebbian learning within a network for attentional processing

Architecture

Our model consists of two layers, whose neurons are bidirectionally connected with each other by feedforward (W) and feedback (A) weights (Figure 1). The image

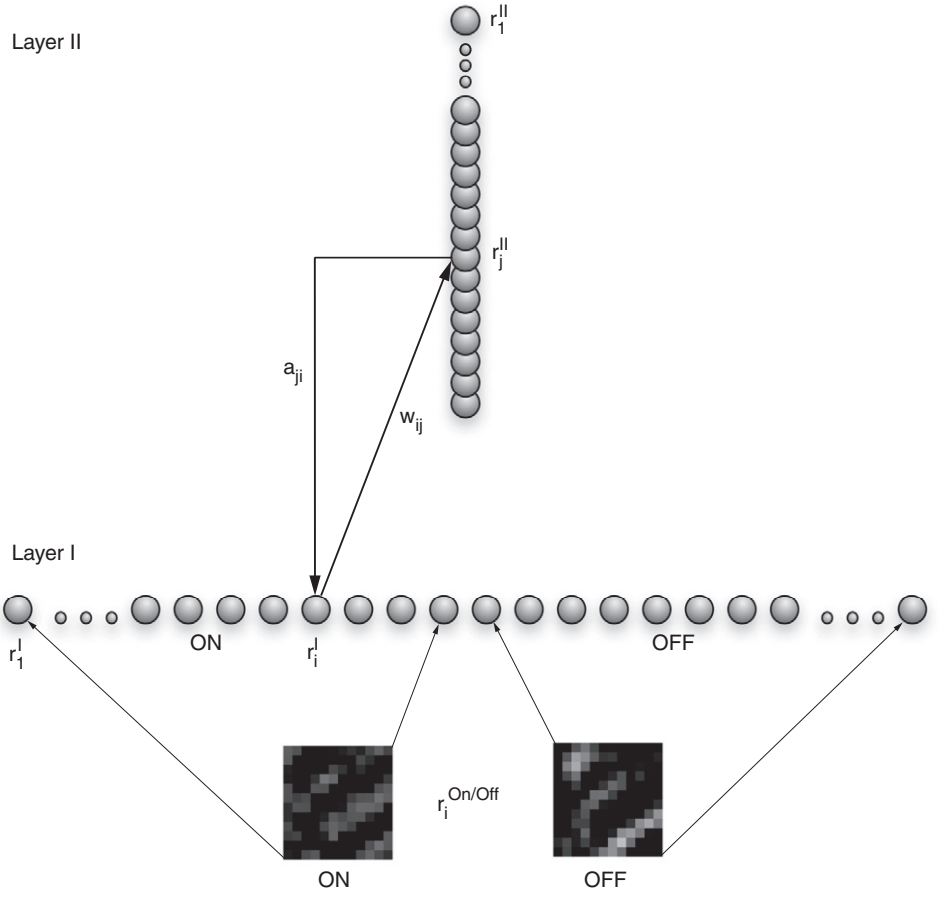

Figure 1. Our network consists of two layers. The neural dynamics implement a feedforward/feedback system, where feedback strengthens the representation of the predicted features in the input. If other top-down signals were available this network could be used to implement feature-based attention. The on/off responses are determined from the input image. The firing rate of these cells $r_i^{On/Off}$ provides the input to layer I cells which are subject to feedback control. The layer I cells represent by their activity r_i^I the content of the visual scene with an additional small increase in firing rate if the input matches the expectation from layer II. Layer II is supposed to learn a more efficient representation of the visual content. The feedforward w_{ij} and feedback weights a_{ji} are learnt simultaneously.

has been whitened/lowpass filtered (see ''Methods'') and separated into on/off channels (depending on the sign of the pixel value after filtering) which yields the input activity $r_i^{\mathrm{On/Off}}$ of layer I.

Layer II gets activated from layer I neurons, but dependent on the activity of other layer II cells. The layer II cells feed back to layer I cells and increase their gain. Due to the learning of the feedback weights, this feedback is predictive. The feedback signal enhances the sensitivity of specific neurons in the previous layer and thus leads to an ''attentional'' tuning.

Neural dynamics

We simulate the change in the firing rate of the cells with differential equations. The activity of the model units and the weight of the connection between cells are restricted to nonnegative values.

Layer I. The neurons in layer I are driven by the on- and off-cells (Figure 1). Feedback from layer II implements a gain modulation (Bayerl and Neumann 2004; Hamker 2004; Hamker 2005). A related approach has been used earlier to dynamically link features (Eckhorn et al. 1990). There is no lateral competition among the neurons in layer I, but they can receive a selective reentrant signal due to the competitive dynamics in layer II. The firing rate r_i^{I} of layer I cells is simulated by:

$$\tau \frac{\mathrm{d}r_i^{\mathrm{I}}}{\mathrm{d}t} = r_i^{\mathrm{On/Off}} \cdot \left(1 + \left(\gamma - \max_k r_k^{\mathrm{I}}\right)^+ \cdot \sum_j a_{ji} r_j^{\mathrm{II}}\right) - r_i^{\mathrm{I}} \tag{1}$$

i refers to the position of the neurons in the image space, $\tau = 10\,\mathrm{ms}$ is the time constant of the temporal dynamics, a_{ji} denotes the feedback weight from neuron j of layer II to neuron i of the first layer and $(x)^+ = \max(x, 0)$. r_i^{I} and r_i^{II} denote the strength of the firing rate for the corresponding neuron. The parameter ($\gamma = 1$) determines the influence of the feedback signal with respect to the activity in the postsynaptic layer. Please refer to Yu et al. (2002) for a discussion about a biophysical implementation of the maximum operation.

Layer II. Layer II neurons learn a combination of specific input features. Their firing rates are determined by the weighted sum of the activity in layer I and by pre-synaptic lateral inhibition (Spratling and Johnson 2002) to induce competition among cells:

$$\tau \frac{\mathrm{d}r_j^{\mathrm{II}}}{\mathrm{d}t} = \sum_i \left[w_{ij} r_i^{\mathrm{I}} \left(1 - \max_{k,k\neq j} \left(\frac{w_{ik}}{\max_m w_{mk}} \frac{r_k^{\mathrm{II}}}{\max_n r_n^{\mathrm{II}}} \right) \right)^+ \right] - r_j^{\mathrm{II}} \tag{2}$$

w_{ij} denotes the strength of the feedforward weight from neuron i of layer I to neuron j of layer II. We investigated this neuronal dynamics earlier on the bar-learning problem and observed advantageous properties in the overlap condition Wiltschut et al. (in preparation). Overlap refers to the ability to learn the discrimination of input patterns consisting of shared elements, e.g., the learning

of a representation of A and B, and one that encodes the joint occurrence of A and B by a different set of neurons.

Basically, the gain of each connection from a neuron i in layer I to a neuron j in layer II is selectively decreased if another neuron k includes the feature encoded by neuron i in its representation and if neuron k is active. Both factors of the gain, the weight and the firing rate, are normalized. The dynamic decrease of the gain avoids inhibition just from the presence of other active cells but induces competition if neurons are tuned to similar features (Figure 2). The effective weights w^e_{ij} that are driving the activation of the second layer shown in Figure 9 are determined from the constant weights w_{ij} combined with the corresponding pre-synaptic lateral inhibition:

$$w^e_{ij} = w_{ij} \cdot \left(1 - \max_{k,k\neq j}\left(\frac{w_{ik}}{\max_m w_{mk}} \frac{r^{\mathrm{II}}_k}{\max_n r^{\mathrm{II}}_n}\right)\right)^+ \tag{3}$$

Note that the effective weights of a neuron differ with respect to the presented inputs.

Learning rules

The long-term potentiation (LTP) of the connections between neurons is implemented via a Hebbian learning principle. As there are many possibilities on how to constrain the weights from a permanent increase and to implement long-term depression (LTD) we previously compared several algorithms (Wiltschut et al. in preparation). These simulations with artificial data suggest that the post/not-pre learning principle is particularly suitable for LTD. However, on natural scenes we observed that the resulting receptive fields are more smooth, if the feedback learning rule slightly differs from the feedforward one and implements LTD only by the constraint to limit the overall weight resource. LTP requires

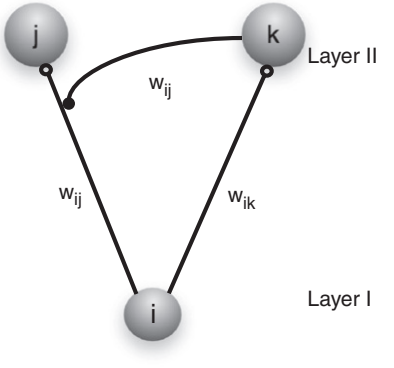

Figure 2. The concept of presynaptic inhibition (Spratling and Johnson 2002). The effective feedforward connection w_{ij} from neuron i in layer I to neuron j in layer II depends on the activity of other neurons k in layer II. If neuron k is strongly activated by the present input it prevents other cells using the same input features for which it is tuned. Thus, the presynaptic inhibition of neuron k to neuron j is input selective. The resulting dynamics is much different from a winner-take-all competition where the inhibition is not selective, since here several cells can be simultaneously activated if they are tuned to different aspects of the present input.

an above-mean activation of both, pre- and post-synaptic activities, which is well known as the covariance learning rule (Sejnowski 1977; Willshaw and Dayan 1990). Specifically we used:

$$\tau_l \frac{\mathrm{d}w_{ij}}{\mathrm{d}t} = \left(r_j^{\mathrm{II}} - \tilde{r}^{\mathrm{II}}\right)^+ \left(\left(r_i^{\mathrm{I}} - \tilde{r}^{\mathrm{I}}\right) - \alpha\left(r_j^{\mathrm{II}} - \tilde{r}^{\mathrm{II}}\right)^+ w_{ij}\right) \tag{4}$$

$$\tau_l \frac{\mathrm{d}a_{ji}}{\mathrm{d}t_l} = \left(r_i^{\mathrm{I}} - \tilde{r}^{\mathrm{I}}\right)^+ \left(\left(r_j^{\mathrm{II}} - \tilde{r}^{\mathrm{II}}\right) - \alpha\left(r_i^{\mathrm{I}} - \tilde{r}^{\mathrm{I}}\right)^+ a_{ji}\right) \tag{5}$$

$\tilde{r}$ is the mean of the activation in a particular layer (e.g., $\tilde{r}^{\mathrm{I}} = (1/N)\sum_{i=1}^{N} r_i^{\mathrm{I}}$) and $\tau_l = 250$ ms the time constant for learning. The weights are prevented from getting negative. α forces each post-synaptic cell to limit its recourses. It is primarily dependent on the number of weights and an appropriate value can be easily estimated from the stable solution of the ODE and the desired activation r_j^{II} given r_i^{I}.

Methods

We applied the model to learn receptive fields from on/off channel responses to natural scenes. In order to obtain the image data we used the software package ''nnscpack'' from Patrik Hoyer (http://www.cs.helsinki.fi/u/phoyer/code/nnscpack.tar.gz), which in turn took the natural scenes from Bruno Olshausen's ''Sparsenet'' software package (http://redwood.berkeley.edu/bruno/sparsenet). The image data consists of 10 images (512×512 pixels). To roughly simulate the characteristics of retinal ganglion cells, each image has been filtered with a zero-phase whitening/lowpass filter $R(f) = f \cdot \exp(-(f/f_0)^4)$ with $f_0 = 200$ cycles/picture (Olshausen and Field 1996). This filter attenuates low frequencies and boosts high frequencies to obtain a roughly flat amplitude spectrum across spatial frequencies. In image space, this filter has a circularly symmetric, center-surround (mexican hat) shape. For every image the same number of randomly selected patches of 12×12 pixels has been taken and used for learning. We did not define a training set with a fixed number of patches but randomly chose a patch for each trial. Each patch is divided in two different channels (on/off) and each channel is normalized to unit mean squared activation.

As the on/off channels consist of 12×12 cells, we required 288 cells for our first layer, i.e., 144 neurons obtain input from the on-, the others from the off-cells. We used 288 cells in layer II to represent the input combinations. The feedforward weights w_{ij} were initialized randomly with a mean $\bar{w} = 0.1$. The feedback weights were initialized with zero. An image patch is presented for 50 ms to let the dynamics of the system converge to a stable state. After each trial both feedback and feedforward synapses are learned according to the final firing rates of the cells. To reveal the influence of some model assumptions, we ran several additional simulations. First of all, we used a model without feedback connections. Second, we investigated the influence of the nonnegativity constraint on the weights. A control study with unconstrained weights was run for 230.000 cycles to reveal the influence of the nonnegative constraint on the weights. Third, in order to investigate the influence of α on the learning of the receptive field structure, we ran identical

simulations with respect to the sequence of randomly chosen values for the initialization and the presentation of patches using three different values for α ($\alpha = 10$, $\alpha = 50$, $\alpha = 100$). Fourth, we tested the model with a learning rule that leads to a less broad feedback projection:

$$\tau_l \frac{\mathrm{d}a_{ji}}{\mathrm{d}t_l} = \left(r_i^{\mathrm{I}} - \tilde{r}^{\mathrm{I}}\right)^+ \left(\left(r_j^{\mathrm{II}} - \tilde{r}^{\mathrm{II}}\right) - \alpha\left(r_i^{\mathrm{I}} - \tilde{r}^{\mathrm{I}}\right)^+ a_{ji}\right) \tag{6}$$

Compared to Equation 5 the feedback connection from a layer II to a layer I neuron decreases if a layer II neuron fires below average. The learning rule of Equation 6 constraints the feedback connections only to pairs which are simultaneously active for most of the time, whereas the original one allows a layer II cell to develop a more broad connectivity, since the penalty for noncorrelated activity is lower.

To compare the learned weight kernels or the receptive fields with Gabor functions, we fitted each with the following equation:

$$G(x, y; x_0, y_0, \sigma_x, \sigma_y, f, \theta, \psi) = \cos(2\pi \cdot f \cdot \hat{x} - \psi) \cdot \exp\left(-\frac{\hat{x}^2}{2\sigma_x^2} - \frac{\hat{y}^2}{2\sigma_y^2}\right) \tag{7}$$

with $\hat{x} = ((x - x_0)\cos(\theta) + (y - y_0)\sin(\theta))/12$ and $\hat{y} = -((x - x_0)\sin(\theta) + (y - y_0)\cos(\theta))/12$. We used all 144 center positions for x_0 and y_0 which are normalized to one, varied σ_x and σ_y from 0.01 to 0.3 also normalized to the patchsize in steps of 0.01. f was varied between 0 and 3 cycles/patch in 30 steps. For the orientation θ, we used values between 0 and 2π in 30 steps and the phase was varied between 0 and $3\pi/4$ in steps of $\pi/4$. The best Gabor fit was determined by the minimum of the sum of the squared difference between the normalized Gabor and the normalized receptive field.

Results

We show the results after 400.000 presentations of patches (Figure 3). The approximate shape of the kernels, however, is visible after about 70.000 presentations. The bottom-up and top-down weights converge to similar profiles (mean sum-of-squares difference between normalized weights: 0.38, 0.9-Quantil = 0.52 and 0.1-Quantil = 0.28). This result is a consequence of the learning rules (Equations 4 and 5). The learning rules are almost symmetric and pick up the correlations in the firing rates across the layers. However, we observed that the learning of the feedback weight has to be less competitive. It is advantageous to induce no LTD if the presynaptic cell (layer II) fires below threshold and the postsynaptic cell (layer I) fires above threshold (compare Equation 6). Although the firing of the presynaptic cell cannot be the cause of the postsynaptic response, LTD is not appropriate in this case since the layer II cell should participate in the encoding of multiple patterns. Otherwise, more patches with weak orientation tuning emerge. Most of the kernels are localized, oriented and bandpass, similar to several earlier approaches using a generative model. We also obtain blob-like kernels which appear absent in the classical sparse-coding model (Olshausen and Field 1996) and in ICA (Bell and Sejnowski 1997; van Hateren and van der Schaaf 1998),

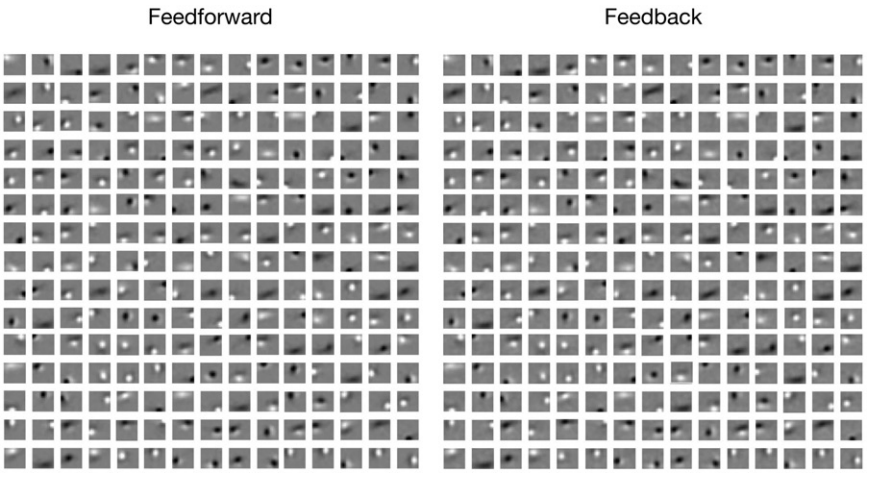

Figure 3. The learned feedforward and feedback connections of 225 example neurons after 400.000 image presentations.

but see (Li and Atick 1994) who obtained blob-like cells based on efficient coding and the additional constraint of object invariance.

To estimate the receptive field profile, including the whitening/lowpass filtering stage, we convolved the whitening/lowpass filter with the learned weights, which basically reduces the frequency but does not change the overall shape. The obtained receptive fields are well fitted with Gabor functions (mean sum-of-squares difference: 0.92, Q.9 = 1.43 and Q.1 = 0.43). Examples of the receptive field profile and the respective Gabor fit are depicted in Figure 4A. The spatial frequency (Figure 4A and B) is shifted to lower values compared to Sparsenet: (Olshausen and Field 1997). Consistent with monkey V1 experimental data (Ringach 2002), our model shows blob-like receptive fields but high frequency components are missing in the model. Other nonnegative generative coding approaches also show a tendency towards lower spatial frequencies (Hoyer 2003; Falconbridge et al. 2006). However, nonnegativity alone, at least as far as the weights are concerned, does not explain the absence of high frequencies. When we drop the nonnegativity constraint on the weights, spatial frequency increases only slightly whereas the Gaussian envelopes get much broader (Figure 5).

Although we do not explicitly enforce sparseness, the model converges to a sparse representation showing an activity distribution that is peakier than Gaussian (Figure 6). However, the degree of sparseness as measured by the kurtosis of the distribution is quite low ($K_{\text{no feedback}} = 19$, $K_{\text{feedback}} = 8$ as compared to Sparsenet $K_{\text{Sparsenet}} \approx 200$ (Rehn and Sommer 2007)). Moreover, due to whitening, the input into the model is already sparse. Control simulations with 1152 cells in layer II did not show fundamental differences in the degree of sparseness. We compared the feedforward weights of the model with and without feedback and found no major difference between both. The lower kurtosis of the activity distribution in the model with feedback appears surprising, since feedback could reinforce a sparse representation due to positive feedback. However, since many cells in layer II are active, the feedback induced by the feedforward activation is quite broad and

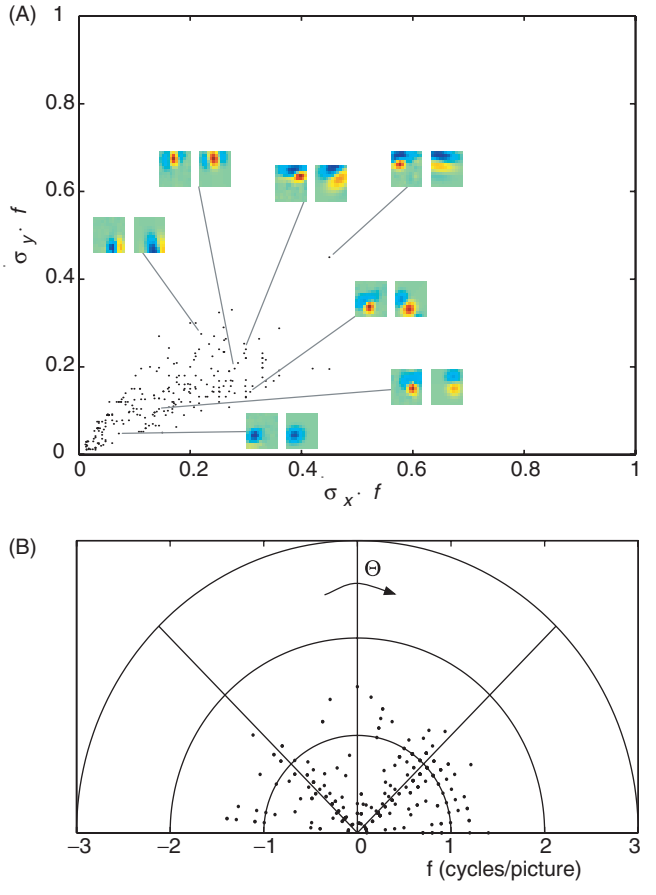

Figure 4. Receptive field properties. (A) The width and height of the Gabor envelope measured in periods of the cosine wave. A number of example receptive fields (left picture) and the respective Gabor fit (right picture) are shown along the distribution. Experimental data from (Ringach 2002) shows more cells with increased length and higher frequencies, otherwise our data is comparable with the experimentally obtained one. (B) Frequency and orientation range of the fitted Gabor receptive field functions.

allows many cells in layer I to increase in activity. Thus, an increase in sparseness would only be expected if the layer II cells strongly compete with each other. A fundamentally different picture with respect to sparseness emerges from dropping the nonnegativity constraint on the weights. Due to the feedforward inhibition the responses get very sparse ($K_{\text{neg. weights}} = 956$).

The value of α in the learning rule has no fundamental influence on the shape of the receptive field (Figure 7). For the larger weights a change in α implements a scaling of all weights with a roughly constant factor.

In order to shed more light on the influence of feedback on learning, we compared the progress of learning the feedforward weights W with and without feedback from layer II to layer I (Figure 8). The model with feedback approximates to the final weight W much faster within the critical first 100.000 trials where the overall RF structure is learned. The fine tuning, the period in which only little change takes place, however, is faster in the model without feedback.

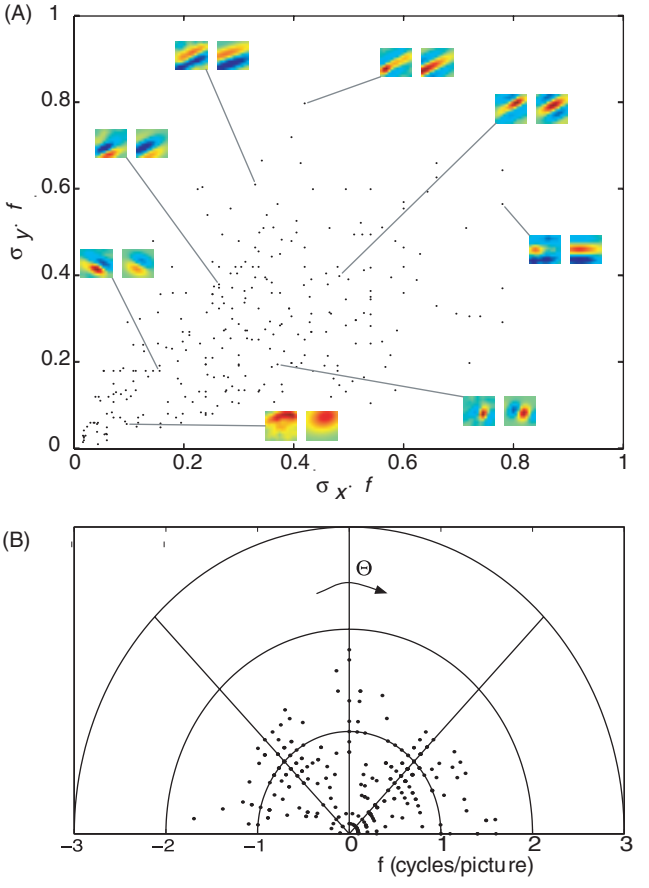

Figure 5. Receptive field properties when the nonnegative constraint on the weights is dropped. The removal of the nonnegativity constraint on the weights only slightly increases the spatial frequency but it leads to a broader envelope. (A) The width and height of the Gabor envelope measured in periods of the cosine wave. Receptive field examples (left picture) and the respective Gabor fit (right picture). The positive and the negative weights contribute almost equally to the receptive field profile for each position in the visual field. (B) Frequency and orientation range of the fitted Gabor receptive field functions.

Sparsenet (Olshausen and Field 1997) predicts a nonlinear effect for units with overlapping basis functions. This interesting property is also found in our model. The units ought to compete with each other such that the one optimally tuned for a specific stimulus suppresses the less tuned one. According to our model the effective receptive field of a cell is not fixed but depends on the stimulation, i.e. the change in the receptive field structure depends on the content of the visual input (Figure 9). Whereas the excitatory contribution describes the basic tuning of a cell, the stimulus-dependent gain decrease significantly alters a neuron's receptive field structure. If this was true, it would imply that experimental estimates of the receptive field structure depend on the mapping procedure. It is of course well known that such mapping must be done with appropriate stimuli. According to our model however, a cell's response can only be understood in the context of other cell's responses.

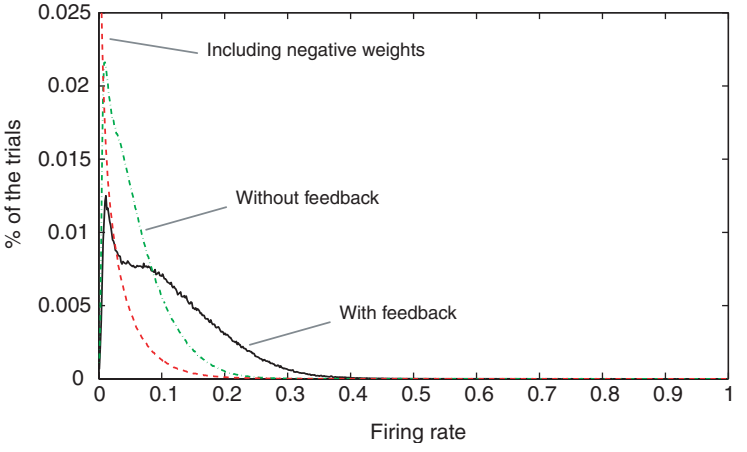

Figure 6. Sparseness of layer II activation. The graph shows the histogram of the firing rate of layer II neurons to 10.000 randomly selected patches averaged over all 288 neurons of layer II using a model that was trained with feedback (solid black line), without feedback (green dashed line) and with no constraint on the sign of the weights (dashed red line). We cut the y-axis at 0.025.

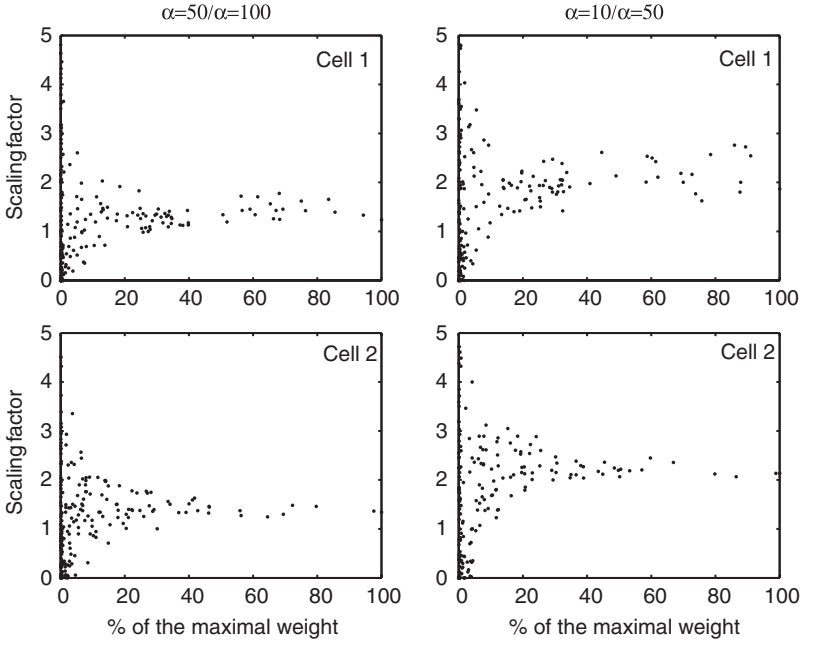

Figure 7. The influence of α on the scaling of the weights. The graphs show the scaling factor of the weights dependent on the strength of the weight for two different cells, each in the case when $\alpha = 50$ is compared to $\alpha = 100$ and when $\alpha = 10$ is compared to $\alpha = 50$. For large weights, a change in α implements a uniform scaling of the weight. α has no relevant influence on the receptive field structure.

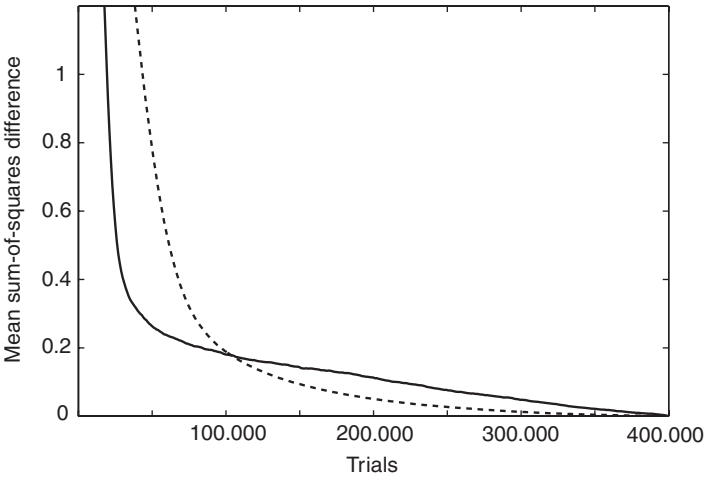

Figure 8. Convergence of learning using a model with (solid line) and without (dashed line) feedback with reference to the final feedforward weights in each condition (W). Since the weights converged in both conditions to similar final values (after 400.000 image presentations), we can compare the progress of learning by computing the mean sum-of-squares difference between the final weights and the present weights in each condition.

Our intention to introduce feedback into the model has been motivated by our earlier studies investigating the role of attention (Hamker 2005, 2006). In these models the feature space was identical across all layers in the processing hierarchy. Here, the learning of feedforward and feedback connections allows us to demonstrate how a search template selective for a particular location, orientation and frequency is transferred downwards to selectively enhance the gain of center-surround cells. Such gain increase has been often observed in experiments of visual attention (Reynolds and Chelazzi 2004), but critical experiments addressing the source and neural pathways have primarily been performed only in occulomotor areas, but see Bullier et al. (2001). We propose to experimentally test top-down gain control by the microstimulation of cells at a particular level of the cortical hierarchy and the simultaneous recording at an earlier level which receives feedback (Figure 10). Techniques of such dual stimulation and recording studies are available (Sommer and Wurtz 2004; Armstrong et al. 2006). Our model predicts that microstimulation alone should not lead to a significant enhancement of the neurons at the earlier level (except of a baseline increase), but when these neurons encode a presented stimulus the response should be enhanced by the micro-stimulation at the higher level neuron if the lower level neuron participates in the sensory representation of the higher level neuron.

Discussion

Anatomically massive feedback projections from TE to TEO, V4, V2 and even as far as V1 have been identified (Rockland and van Hoesen 1994; Rockland et al. 1994).

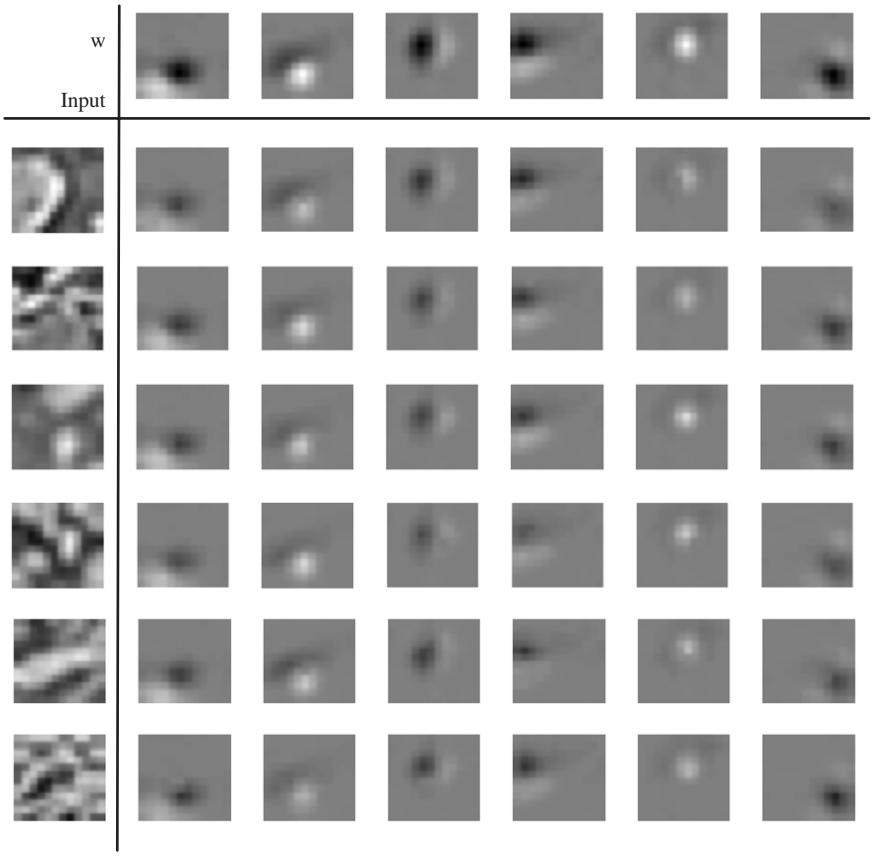

Figure 9. Effective weight depending on the input to the network. The graph shows the feedforward weights W of six different cells and their change relative to the presented input. For details refer to the "Methods" section.

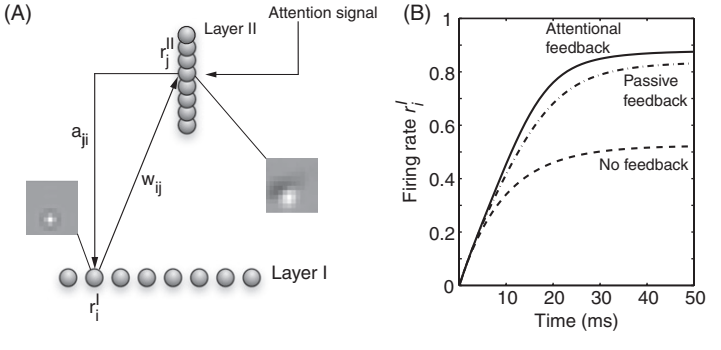

Figure 10. (A) Transfer of a top-down search template into early selective visual responses. We simulated the search for a particular feature by providing a top-down attentional signal (or alternatively an external stimulation of this cell) to a neuron in layer II. As a result, also neurons in layer I with on/off receptive fields get selectively tuned. (B) Firing rate of the example neuron in (A). Due to the positive feedback from the particular layer II cell the response is enhanced relative to the one without an attentional signal. For comparison, the response without any feedback is also shown.

It has been suggested that top-down feedback plays an important role in feature binding and attention (Lammé and Roelfsema 2000). As it stands now, there are at least two competitive models of feedback: the generative and the match-enhancement model. In the generative model a good match between the internal hypothesis and the actual input results in a weak feedforward signal and a mismatch in a strong feedforward signal. Thus, feedback primarily serves for ''explaining away'' the evidence by suppressing the activity. The match-enhancement approach, which shows some similarity to adaptive resonance (Grossberg 1980), predicts a gain increase of the feedforward signal if both signals are consistent with each other. If both signals are not consistent, no enhancement occurs, i.e., no gain change takes place (Hamker 2007). A complete model of the ventral pathway should consist of stackable modules, but also direct connections across more distant layers should be possible. Thus, the top-down connection to layer II, denoted as an attentional signal (Figure 10), should indeed be the feedback signal from a layer III module and should act like the feedback from layer II to layer I. Before these modules can be put together some issues have to be addressed. First of all, the level of activity has to be controlled by some form of homeostatic regulation (Turrigiano and Nelson 2004) to ensure that the feedforward signal neither dies out nor saturates. Since the top-down signal only acts on the gain and saturates with respect to the firing rate in the target layer (Equation 1), an explosion of the overall activity of the circuit is not much of a concern. An overall control of activity would nevertheless be necessary. Second, mechanisms of spatial and view invariance have to be addressed, which presumably require additional stages or at least adjustments of the learning rule.

The generative approach became very popular to demonstrate the learning of receptive fields, since it is relatively easy to define an objective function which can be minimized by well-known algorithms. The match-enhancement approach has been successfully used to describe attentional dynamics (Hamker 2005) but it has been unclear if feedforward and feedback weights can be learned within the dynamics of visual processing. In this study, we have shown that consistent feedforward and feedback weights can indeed be learned within the match-enhancement approach and thus we offer an alternative to the generative approach. It is too early to judge which of the two approaches is superior over the other. The generative approach allows to formulate an energy function, whereas the match-enhancement approach appears to be more general for learning weights under attentive control in higher brain areas. Our model allows to explain how more high level search templates travel in reverse direction to selectively enhance neurons encoding more detailed aspects, a mechanism we suggest being fundamental for feature-based attention. This mechanism could also be responsible for observed attentional effects in the lateral geniculate nucleus (O'Connor et al. 2002).

Despite these conceptional differences, there are also some important similarities between the generative approach and ours. According to our notation, the generative approach seeks to minimize the following energy function:

$$E = \sum_i \left(r_i^{\mathrm{On/Off}} - \sum_j a_{ij} \cdot r_j^{\mathrm{II}} \right)^2 - \lambda \sum_j S(r_j^{\mathrm{II}}) \tag{8}$$

where the first part represents the difference between the original and predicted input, and the second part enforces sparseness. The short-term, ''neural'' dynamics are described by the derivative with respect to each r_j^{II} and reads:

$$-\frac{\partial E}{\partial r_j^{\mathrm{II}}} = \sum_i r_i^{\mathrm{On/Off}} a_{ij} - \sum_i \sum_l a_{il} a_{ij} \cdot r_j^{\mathrm{II}} - \lambda \sum_j S'(r_j^{\mathrm{II}}) \tag{9}$$

The second term means that neurons with similar receptive fields compete with each other. This principle can also be found in our model (Equation 2). Despite some differences in the exact implementation, the general idea is very similar. Although the importance of an appropriate learning rule on the resulting receptive field structure cannot be neglected, the competition of cells for sensory representation (often referred to as ''explaining away'') appears to be an essential component in neural coding. Thus, inherent to both models is a dependence of the receptive field on the stimulus used to measure the receptive field. Since the effective weight depends on the response of other cells in the layer, a full understanding of the receptive field would require to know how other cells respond to the stimulus. This appears consistent with the observation that natural scenes can evoke much different responses than those estimated by synthetic stimuli (David et al. 2004). It also suggests that the receptive fields of cells in models of visual coding should be mapped with similar stimuli as it is done in experiments to allow better comparisons.

When we consider the exact implementation of mutual inhibition, both models make different predictions with respect to the development of receptive fields. In our model, mutual inhibition is already present in the feedforward path and relies on the similarity of the feedforward weights. The gain of the weights are dynamically decreased if other neurons with a similar weight kernel already successfully encode the present stimulus. In the generative model the similarity measurement is based on the feedback weights and is embedded in the analysis/synthesis loop. If it was technically possible to selectively shut off the feedback pathway, our model would predict no impairment in the development of receptive fields, whereas according to the generative model, receptive field development should be impaired. Although feedback is, according to our model, not crucial for the development of receptive fields, feedback can nevertheless alter the receptive field structure, for example when a particular object is repeatedly presented and attended to (represented with higher activity compared to other objects).

The activity and the weights in our model are constrained to nonnegative values. This is motivated by the fact that negative firing rates of neurons do not exist and the feedforward projection from LGN to layer 4 in V1 is excitatory. In the classical generative approach no constraint is imposed on the weights as well as the activity. Nonnegativity combines elementary features additively, whereas otherwise, features can cancel each other out. In addition to this biological motivation, nonnegativity has been suggested to lead to a part-based representation (Lee and Seung 1999), although evidence for this is mixed. For example, nonnegative matrix factorization (Lee and Seung 1999) does not always lead to a part-based representation (Hoyer 2004), but this property appears improved when an additional sparseness constraint is used (Hoyer 2004). Without this constraint, nonnegative matrix factorization does even not converge to oriented receptive fields when natural scenes are presented (Hoyer 2004). This suggests that sparseness is critical for nonnegative

sparse coding to show essential properties of neural coding in visual areas. However, the degree of sparseness is quite low in our model (only 11% of the one used in nonnegative matrix factorization). If we do not impose a nonnegativity constraint on the weights, we observe oriented receptive fields with broader envelopes and a much higher degree of sparseness. Now the inhibitory feedforward connections allow to suppress the cell's response which allows them to get more selective to a particular feature. The increase in selectivity is reinforced by the first part of the learning rule $((r_j^{\mathrm{II}} - \tilde{r}^{\mathrm{II}})^+(r_i^{\mathrm{I}} - \tilde{r}^{\mathrm{I}}))$, which contributes to a weight decay only, when the layer I cell fires below average and the layer II cell above average. Despite these differences in the selectivity of the cells, both versions of the model lead to receptive fields which show some of the typical characteristics observed in V1 cells.

Our model does not contain an additional, more explicit constraint to enforce sparseness like in Sparsenet. Different forms of sparseness have been recently reviewed (Rehn and Sommer 2007). According to this study, Sparsenet (Olshausen and Field 1996) implements a ''soft'' form of sparseness that limits the average neural activity. Alternatively, ''hard'' forms of sparseness enforce the proportion of neurons representing a single image being small. This ''hard'' sparse-ness has been suggested to provide a better fit with experimental data specifically showing also unoriented, blob-like receptive fields (Rehn and Sommer 2007). Since our model does not make use of postsynaptic inhibition to suppress less well-tuned cells, our empirically obtained level of sparseness is low and the distribution shows no discontinuous density of neural activity as predicted by ''hard'' sparseness models. Nevertheless, we observe oriented (low frequency) and blob-like receptive fields.

Acknowledgements

We thank Julien Vitay (Westf. Wilhelms-Universität) for rewriting a part of the code and the fitting procedure of the receptive fields in the programming language C to speed up the simulation time. This work has been supported by the German Science Foundation (DFG HA2630/4).

References

Atick JJ, Redlich A. 1990. Towards a theory of early visual processing. Neural Comput 2:308–320.

Armstrong KM, Fitzgerald JK, Moore T. 2006. Changes in visual receptive fields with microstimulation of frontal cortex. Neuron 50:791–798.

Barlow HB. 1961. Possible principles underlying the transformation of sensory messages. In: Rosenblith WA, editor. Sensory communication. Cambridge, MA: MIT Press. pp 217–234.

Barlow HB. 1998. Redundancy reduction revisited. Network 12:241–253.

Bayerl P, Neumann H. 2004. Disambiguating visual motion through contextual feedback modulation. Neural Comput 16:2041–2066.

Bell AJ, Sejnowski TJ. 1997. The 'independent components' of natural scenes are edge filters. Vis Res 37:3327–3338.

Bullier J, Hupe JM, James AC, Girard P. 2001. The role of feedback connections in shaping the responses of visual cortical neurons. Prog Brain Res 134:193–204.

David SV, Vinje WE, Gallant JL. 2004. Natural stimulus statistics alter the receptive field structure of v1 neurons. J Neurosci 24:6991–7006.

Eckhorn R, Reitboeck E, Arndt M, Dicke P. 1990. Feature linking via synchronisation among distributed assemblies: Simulations of results from Cat Visual Cortex. Neural Comput 2:293–307.
Falconbridge MS, Stamps RL, Badcock DR. 2006. A simple Hebbian/anti-Hebbian network learns the sparse, independent components of natural images. Neural Comput 18:415–429.
Grossberg S. 1980. How does the brain build a cognitive code? Psychol Rev 87:1–51.
Hamker FH. 2004. A dynamic model of how feature cues guide spatial attention. Vision Res 44:501–521.
Hamker FH. 2005. The reentry hypothesis: The putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement. Cereb Cortex 15:431–447.
Hamker FH. 2006. Modeling feature-based attention as an active top-down inference process. BioSystems 86:91–99.
Hamker FH. 2007. The mechanisms of feature inheritance as predicted by a systems-level model of visual attention and decision making. Adv Cogn Psychol 3:111–123.
Hancock PJB, Baddeley RJ, Smith LS. 1992. The principle components of natural images. Network 3:61–70.
Harpur G, Prager R. 1996. Development of low entropy coding in a recurrent network. Network: Comput Neural Syst 7:277–284.
Hoyer PO. 2003. Modeling receptive fields with non-negative sparse coding. Neurocomputing 52–54:547–552.
Hoyer PO. 2004. Non-negative matrix factorization with sparseness constraints. J Mach Learn Res 5:1457–1469.
Hoyer PO, Hyvärinen A. 2002. A multi-layer sparse coding network learns contour coding from natural images. Vision Res 42:1593–1605.
Jehee JF, Rothkopf C, Beck JM, Ballard DH. 2006. Learning receptive fields using predictive feedback. J Physiol Paris 100:125–132.
Karklin Y, Lewicki MS. 2003. Learning higher-order structures in natural images. Network 14:483–499.
Lammé VAF, Roelfsema PR. 2000. The distinct modes of vision offered by feedforward and recurrent processing. Trend Neurosci 23:571–579.
Lee DD, Seung HS. 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791.
Li Z, Atick JJ. 1994. Towards a theory of striate cortex. Neural Comput 6:127–146.
Linsker R. 1986. From basic network principles to neural architecture: Emergence of orientation-selective cells. Proc Natl Acad Sci USA 83:8390–8394.
Nadal J-P, Parga N. 1994. Nonlinear neurons in the low-noise limit: A factorial code maximizes information transfer. Network: Comput Neural Sys 5:565–581.
O'Connor DH, Fukui MM, Pinsk MA, Kastner S. 2002. Attention modulates responses in the human lateral geniculate nucleus. Nat Neurosci 5:1203–1209.
Oja E. 1982. A simplified neuron model as a principal component analyzer. J Math Biol 15:267–273.
Olshausen BA, Field DJ. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381:607–609.
Olshausen BA, Field DJ. 1997. Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Res 37:3311–3325.
Rao RP, Ballard DH. 1999. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87.
Rehn M, Sommer FT. 2007. A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. J Comput Neurosci 22:135–146.
Reynolds JH, Chelazzi L. 2004. Attentional modulation of visual processing. Annu Rev Neurosci 27:611–647.
Ringach DL. 2002. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. J Neurophysiol 88:455–463.
Rockland KS, van Hoesen GW. 1994. Direct temporal-occipital feedback connections to striate cortex (V1) in the macaque monkey. Cereb Cortex 4:300–313.
Rockland KS, Saleem KS, Tanaka K. 1994. Divergent feedback connections from areas V4 and TEO in the macaque. Visual Neurosci 11:579–600.
Sejnowski T. 1977. Storing covariance with nonlinearly interacting neurons. J Math Biol 4:303–321.
Simoncelli EP. 2003. Vision and the statistics of the visual environment. Curr Opin Neurobiol 13:144–149.

Sommer MA, Wurtz RH. 2004. What the brain stem tells the frontal cortex. I. Oculomotor signals sent from superior colliculus to frontal eye field via mediodorsal thalamus. J Neurophysiol 91:1381–1402.

Spratling MW, Johnson MH. 2002. Pre-integration lateral inhibition enhances unsupervised learning. Neural Comput 14:2157–2179.

Turrigiano GG, Nelson SB. 2004. Homeostatic plasticity in the developing nervous system. Nat Rev Neurosci 5:97–107.

van Hateren JH, van der Schaaf A. 1998. Independent component filters of natural images compared with simple cells in primary visual cortex. Proc Biol Sci 265:359–366.

von der Marlsburg C. 1973. Self-organization of orientation selective cells in the striate cortex. Kybernetic 14:85–100.

Willshaw DJ, Dayan P. 1990. Optimal plasticity in matrix memories: What goes up must come down. Neural Comput 2:85–93.

Wiltschut J, Zirnsak M, Hamker FH. (in preparation) Hebbian learning of feedforward and feedback connections in dynamic rate coded neurons.

Yu A, Giese MA, Poggio T. 2002. Biophysiologically plausible implementations of the maximum operation. Neural Comput 14:2857–2881.