To appear in: Probabilistic Models of the Brain: Perception and Neural Function. R.P.N. Rao, B.A. Olshausen, M.S. Lewicki, Eds. MIT Press. 2002.

13 Sparse Codes and Spikes

Bruno A. Olshausen

Introduction

In order to make progress toward understanding the sensory coding strategies employed by the cortex, it will be necessary to draw upon guiding principles that provide us with reasonable ideas for what to expect and what to look for in the neural circuitry. The unifying theme behind all of the chapters in this book is that *probabilistic inference*—i.e., the process of inferring the state of the world from the activities of sensory receptors and a probabilistic model for interpreting their activity—provides a major guiding principle for understanding sensory processing in the nervous system. Here, I shall propose a model for how inference may be instantiated in the neural circuitry of the visual cortex, and I will show how it may help us to understand both the form of the receptive fields found in visual cortical neurons as well as the nature of spiking activity in these neurons.

In order for the cortex to perform inference on retinal images, it must somehow implement a generative model for explaining the signals coming from optic nerve fibers in terms of hypotheses about the state of the world (Mumford, 1994). I shall propose here that the neurons in the primary visual cortex, area V1, form the first stage in this generative modeling process by modeling the structure of images in terms of a linear superposition of basis functions (figure 13.1). One can think of these basis functions as a simple "feature vocabulary" for describing images in terms of additive functions. In order to provide a vocabulary that captures meaningful structure within time-varying images, the basis functions are adapted according to an unsupervised learning procedure that attempts to form a representation of the incoming image stream in terms of *sparse, statistically independent* events. Sparseness is desired because it provides a simple description of the structures occurring in natural image sequences in terms of a small number of vocabulary elements at any point in time (Field, 1994). Such representations are also useful for forming associations at later stages of processing (Foldiak, 1995; Baum, 1988). Statistical independence reduces the redundancy of the code, in line with Barlow's hypothesis for achieving a repre-

sentation that reflects the underlying causal structure of the images (Barlow, 1961; 1989).[1]

I shall show here that when a sparse, independent code is sought for time-varying natural images, the basis functions that emerge resemble the receptive field properties of cortical simple-cells in both space and time. Moreover, the model yields a representation of time-varying images in terms of sparse, spike-like events. It is suggested that the spike trains of sensory neurons essentially serve as a *sparse code in time*, which in turn forms a more efficient and meaningful representation of image structure. Thus, a single principle may be able to account for both the receptive properties of neurons and the spiking nature of neural activity.

The first part of this chapter presents the basic generative image model for static images, and discusses how to relate the basis functions and sparse activities of the model to neural receptive fields and activities. The second part applies the model to time-varying images and shows how space-time receptive fields and spike-like representations emerge from this process. Finally, I shall discuss how the model may be tested and how it would need to be further modified in order to be regarded as a fully neurobiologically plausible model.

Sparse Coding of Static Images

Image model

In previous work (Olshausen & Field, 1997), we described a model of V1 simple-cells in terms of a linear generative model of images (figure 13.1*a*). According to this model, images are described in terms of a linear superposition of basis functions plus noise:

$$I(x,y) = \sum_i a_i \, \phi_i(x,y) + \nu(x,y) \; . \tag{13.1}$$

An image $I(x,y)$ is thus represented by a set of coefficient values, a_i, which are taken to be analogous to the activities of V1 neurons. Importantly, the basis set is *overcomplete*, meaning that there are more basis functions (and hence more a_i's) than effective dimensions in the images. Overcompleteness in the representation is important because it allows for the joint space of position, orientation, and spatial-frequency to

1. Although it is not possible in general to achieve complete independence with the simple linear model we propose here, we can nevertheless seek to reduce statistical dependencies as much as possible over both space (i.e., across neurons) and time.

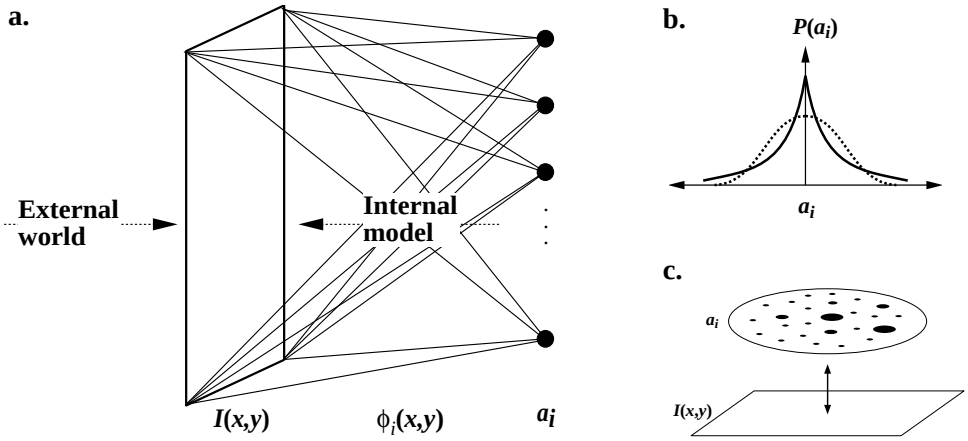

Figure 13.1. Image model. *a*, Images of the environment are modeled as a linear superposition of basis functions, ϕ_i, whose amplitudes are given by the coefficients a_i. *b*, The prior probability distribution over the coefficients is peaked at zero with heavy tails as compared to a Gaussian of the same variance (overlaid as dashed line). Such a distribution would result from a sparse activity distribution over the coefficients, as depicted in *c*.

be tiled smoothly without artifacts (Simoncelli et al., 1992). More generally though, it allows for a greater degree of flexibility in the representation, as there is no reason to believe a priori that the number of causes for images is less than or equal to the number of pixels (Lewicki & Sejnowski, 2000).

With non-zero noise, ν, the correspondence between images and coefficient values is probabilistic—i.e., some solutions are more probable than others. Moreover, when the basis set is overcomplete, there are an infinite number of solutions for the coefficients in equation 13.1 (even with zero noise), all of which describe the image with equal probability. This degeneracy in the representation is resolved by imposing a prior probability distribution over the coefficients. The particular form of the prior imposed in our model is one that favors an interpretation of images in terms of sparse, independent events:

$$P(\mathbf{a}) = \prod_i P(a_i) \tag{13.2}$$

$$P(a_i) = \frac{1}{Z_S} e^{-S(a_i)} \tag{13.3}$$

where S is a non-convex function that shapes $P(a_i)$ so as to have the requisite "sparse" form—i.e., peaked at zero with heavy tails, or positive kurtosis—as shown in figure 13.1*b*. The posterior probability of the coefficients for a given image is then

$$P(\mathbf{a}|\mathbf{I}, \theta) \propto P(\mathbf{I}|\mathbf{a}, \theta) P(\mathbf{a}|\theta) \tag{13.4}$$

$$P(\mathbf{I}|\mathbf{a},\theta) = \frac{1}{Z_{\lambda_N}} e^{-\frac{\lambda_N}{2}|\mathbf{I}-\mathbf{\Phi}\mathbf{a}|^2} \tag{13.5}$$

$$P(\mathbf{a}|\theta) = \prod_i \frac{1}{Z_S} e^{-S(a_i)} \tag{13.6}$$

where $\mathbf{\Phi}$ is the basis function matrix with columns ϕ_i and λ_N is the inverse of the noise variance σ_ν^2. θ denotes the entire set of model parameters $\mathbf{\Phi}$, λ_N, and S.

Since the relation between images and coefficients is probabilistic, there is not a single unique solution for choosing the coefficients to represent a given image. One possibility, for example, is to choose the mean of the posterior distribution $P(\mathbf{a}|\mathbf{I},\theta)$. This is difficult to compute, though, since it requires some form of sampling from the posterior. The solution we propose here is to choose the coefficients that maximize the posterior distribution (MAP estimate)

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} P(\mathbf{a}|\mathbf{I},\theta) \tag{13.7}$$

which is accomplished via gradient ascent on the log-posterior:

$$\begin{aligned}
\dot{\mathbf{a}} &\propto \nabla_{\mathbf{a}} \log P(\mathbf{a}|\mathbf{I},\theta) \\
&= -\nabla_{\mathbf{a}} \left[\frac{\lambda_N}{2}|\mathbf{I}-\mathbf{\Phi}\mathbf{a}|^2 + \sum_i S(a_i) \right] && (13.8) \\
&= \lambda_N \mathbf{\Phi}_i^T \, \mathbf{e} - S'(\mathbf{a}) \; . && (13.9)
\end{aligned}$$

where $\mathbf{e}$ is the residual error between the image and the model's reconstruction of the image, $\mathbf{e} = \mathbf{I} - \mathbf{\Phi}\mathbf{a}$. When S is a non-convex function appropriate for encouraging sparseness, such as $\beta \log(1 + (a_i/\sigma)^2)$, or $\beta|a_i/\sigma|^q, q \leq 1$, its derivative, S', provides a form of non-linear self-inhibition for coefficient values near zero. A recurrent neural network implementation of this differential equation (13.9) is shown in figure 13.2.

Learning

The basis functions of the model are adapted by maximizing the average log-likelihood of the images under the model, which is equivalent to minimizing the model's estimate of code length, $\mathcal{L}$:

$$\mathcal{L} = -\langle \log P(\mathbf{I}|\theta) \rangle \tag{13.10}$$

where

$$P(\mathbf{I}|\theta) = \int P(\mathbf{I}|\mathbf{a},\theta) \, P(\mathbf{a}|\theta) \, d\mathbf{a} \; . \tag{13.11}$$

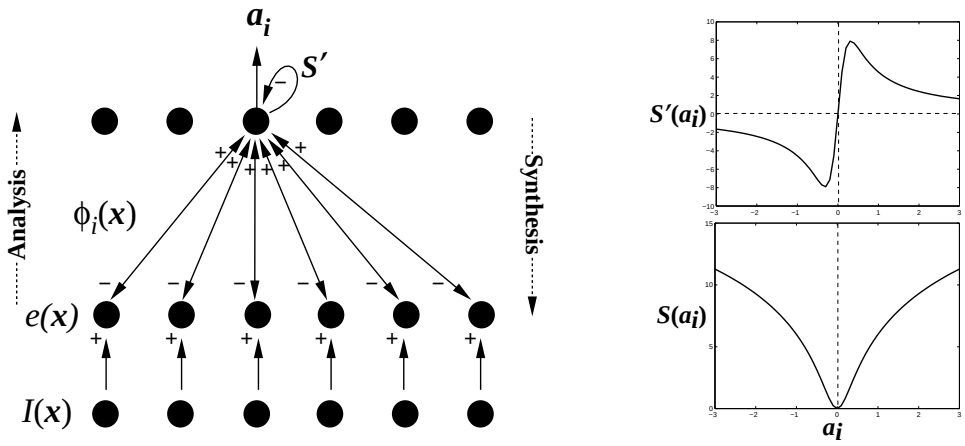

Figure 13.2. A simple network implementation of inference. The outputs a_i are driven by a sum of two terms. The first term takes a spatially weighted sum of the current residual image using the function $\phi_i(\vec{x})$ as the weights. The second term applies a non-linear self-inhibition on the outputs according to the derivative of S, that differentially pushes activity towards zero. Shown at right is the derivative of the sparse cost function $S(a_i) = \beta \log(1 + (a_i/\sigma)^2)$, $\beta = 2.5$, $\alpha = 0.3$.

$\mathcal{L}$ provides an upper bound estimate of the entropy of the images, which in turn provides a lower bound estimate of code length.

A learning rule for the basis functions may be obtained via gradient descent on $\mathcal{L}$:

$$\Delta\mathbf{\Phi} \propto -\frac{\partial \mathcal{L}}{\partial \mathbf{\Phi}} \tag{13.12}$$

$$= \lambda_N \left\langle \langle \mathbf{e}\, \mathbf{a}^T \rangle_{P(\mathbf{a}|\mathbf{I},\theta)} \right\rangle \ . \tag{13.13}$$

Thus, the basis functions are updated by a Hebbian learning rule, where the residual error $\mathbf{e}$ constitutes the pre-synaptic input and the coefficients $\mathbf{a}$ constitute the post-synaptic outputs. Instead of sampling from the full posterior distribution, though, we utilize an simpler approximation in which a single sample is taken at the posterior maximum, and so we have

$$\Delta\mathbf{\Phi} \propto \langle \mathbf{e}\, \hat{\mathbf{a}}^T \rangle \ . \tag{13.14}$$

The price we pay for this approximation is that the basis functions will grow without bound, since the greater their norm, $|\phi_i|$, the smaller each a_i will become, thus decreasing the sparseness penalty in (13.8). This trivial solution is avoided by rescaling

the basis functions after each learning step (13.14) so that their L2 norm, $g_i = |\phi_i|_{L2}$, maintains an appropriate level of variance on each corresponding coefficient a_i:

$$g_i^{new} = g_i^{old} \left[\frac{\langle a_i^2 \rangle}{\sigma^2} \right]^{\alpha}, \tag{13.15}$$

where σ is the scaling parameter used in the sparse cost function, S. This method, although an approximation to gradient descent on the true objective $\mathcal{L}$, has been shown to yield solutions similar to those obtained with more accurate techniques involving sampling (Olshausen & Millman, 2000).

Does V1 do sparse coding?

When the model is adapted to static, whitened[2] natural images, the basis functions that emerge resemble the Gabor-like spatial profiles of cortical simple-cell receptive fields (figure 13.3, similar results were also obtained with van Hateren & Ruderman's ICA model). That is, the functions become spatially localized, oriented, and bandpass (selective to structure at different spatial scales). Because all of these properties emerge purely from the objective of finding sparse, independent components for natural images, the results suggest that the receptive fields of V1 neurons have been designed according to a similar coding principle. The result is quite robust, and has been shown to emerge from other forms of independent components analysis (ICA). Some of these also make an explicit assumption of sparseness (Bell & Sejnowski, 1997; Lewicki & Olshausen, 1999) while others seek only independence among the coefficients, in which case sparseness emerges as part of the result (van Hateren & van der Schaaf, 1998; Olshausen & Millman, 2000).

We are comparing the basis functions to neural receptive fields[3] here because they are the feedforward weighting functions used in computing the outputs of the model, a_i (see figure 13.2). However, it is important to bear in mind that the outputs are not computed purely via this feedforward weighting function, but also via a non-linear, recurrent computation (13.9), the result of which is to *sparsify* neural activity. Thus, a neuron in our model would be expected to respond less often than one that simply

2. Whitening removes second-order correlations due to the $1/f^2$ power spectrum of natural images, and it approximates the type of filtering performed by the retina (see Atick & Redlich, 1992).

3. It should be noted that term 'receptive field' is not well-defined, even among physiologists. Oftentimes it is taken to mean the feedforward, linear weighting function of a neuron. But in reality, the measured receptive field of a neuron reflects the sum total of all dendritic non-linearities, output non-linearities, as well as recurrent computations due to horizontal connections and top-down feedback from other neurons.

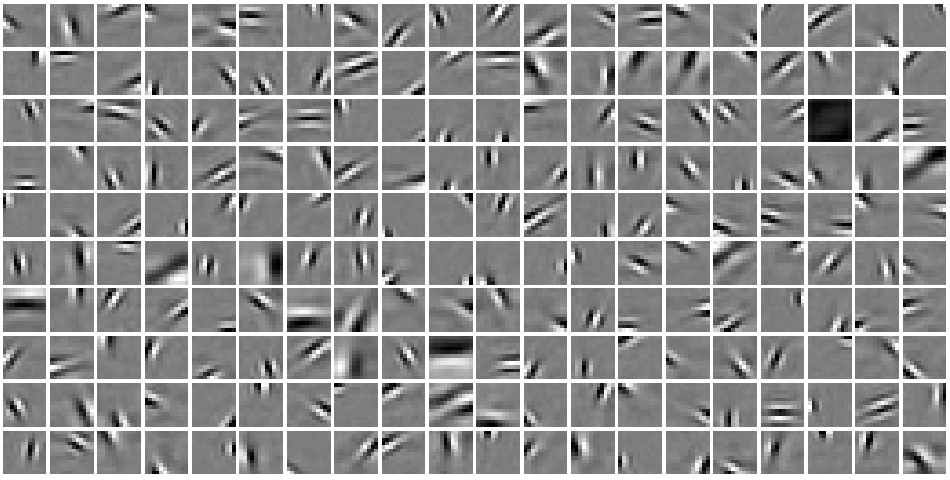

Figure 13.3. Basis functions learned from static natural images. Shown is a set of 200 basis functions which were adapted to 12×12 pixel image patches, according to equations (13.14) and (13.15). Initial conditions were completely random. The basis set is approximately $2\times$'s overcomplete, since the images occupy only about $3/4$ of the dimensionality of the input space. (See Olshausen & Field, 1997, for simulation details.)

computes the inner product between a spatial weighting function and the image, as shown in figure 13.4*a*.

How could one tell if V1 neurons were actively sparsifying their activity according to the model? One possibility is to measure a neuron's receptive field via reverse correlation, using an artificial image ensemble such as white noise, and then use this measured receptive field to predict the response of the neuron to natural images via convolution. If neural activities were being sparsified as in the model, then one would expect the actual responses obtained with natural images to be non-linearly related to those predicted from convolution, as shown in figure 13.4*c*. The net effect of this non-linearity is that it tends to suppress responses where the basis function does not match well with the image, and it amplifies responses where the basis function does match well. This form of non-linearity is qualitatively consistent with the "expansive power-function" contrast response non-linearity observed in simple cells (Albrecht & Hamilton, 1982; Albrecht & Geisler, 1991). Note however that this response property emerges from the sparse prior in our model, rather than having been assumed as an explicit part of the response function. Whether or not this response characteristic is due to the kind of dynamics proposed in our model, as opposed to the application of a fixed pointwise non-linearity on the output of the neuron, would require more complicated tests to resolve.

The above method assumes that the analog valued coefficients in the model (or positively rectified versions of these quantities) correspond to spike rate. However, recent studies have demonstrated that spike rates, which are typically averaged over

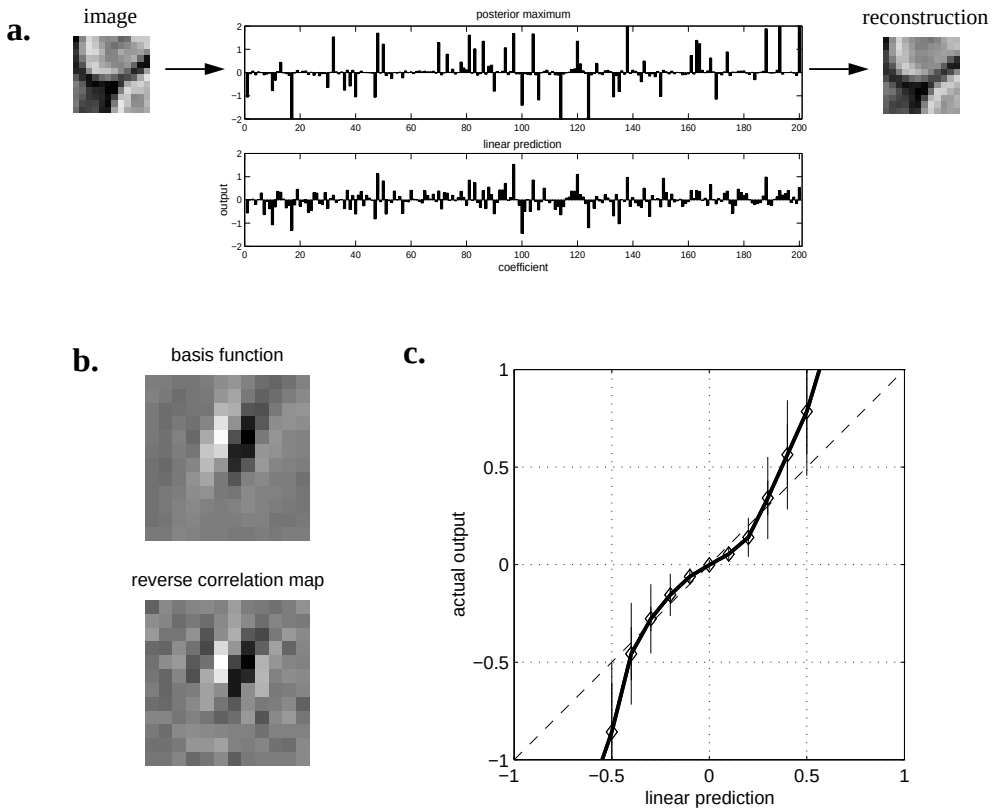

Figure 13.4. Effect of sparsification. *a*, An example 12 × 12 image and its encoding obtained by maximizing the posterior over the coefficients. The representation obtained by simply taking the inner-product of the image with the best linear predicting kernel for each basis function is not nearly as sparse by comparison. *b*, Shown is one of the learned basis functions (row 6, column 7 of figure 13.3) together with its corresponding "receptive field" as mapped out via reverse correlation with white noise (1440 trials). *c*, The response obtained by simply convolving this function with the image is non-linearly related to the actual output chosen by posterior maximization. Specifically, small values tend to get suppressed and large values amplified (the solid line passing through the diamonds depicts the mean of this relationship, while the error bars denote the standard deviation).

epochs of 100 ms or more, tend to vastly underestimate the temporal information contained in neural spike trains (Rieke et al., 1997). In addition, we are faced with the fact that the image on the retina is constantly changing due to both self-motion (eye, head and body) and the motions of objects in the world. The model as we have currently formulated it is not well-suited to deal with such dynamics, since the procedure for maximizing the posterior over the coefficients requires a recurrent computation, and it is unlikely that this will complete before the input changes

appreciably. In the next section, we show how these issues may be addressed, at least in part, by reformulating the model to deal directly with time-varying images.

Sparse Coding of Time-Varying Images

Image model

We can reformulate the sparse coding model to deal with time-varying images by explicitly modeling the image stream $I(x, y, t)$ in terms of a superposition of space-time basis functions $\phi_i(x, y, \tau)$. Here we shall assume shift-invariance in the representation over time, so that the same basis function $\phi_i(x, y, \tau)$ may be used to model structure in the image sequence around any time t with amplitude $a_i(t)$. Thus, the image model may be expressed as the convolution of a set of time-varying coefficients, $a_i(t)$, with the basis functions:

$$I(x, y, t) = \sum_i \sum_{t'} a_i(t')\, \phi_i(x, y, t - t') + \nu(x, y, t) \tag{13.16}$$

$$= \sum_i a_i(t) * \phi_i(x, y, t) + \nu(x, y, t) \tag{13.17}$$

The model is illustrated schematically in figure 13.5.

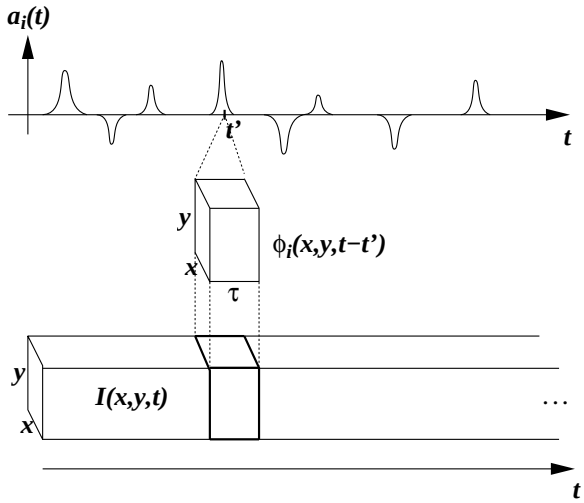

Figure 13.5. Image model. A movie $I(x, y, t)$ is modeled as a linear superposition of spatio-temporal basis functions, $\phi_i(x, y, \tau)$, each of which is localized in time but may be applied at any time within the movie sequence.

The coefficients for a given image sequence are computed as before by maximizing the posterior distribution over the coefficients

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} P(\mathbf{a}|\mathbf{I}, \theta) \tag{13.18}$$

which is again achieved by gradient descent, leading to the following differential equation for determining the coefficients:

$$\dot{a}_i(t) \propto \lambda_N \sum_{x,y} \phi_i(x, y, t) \star e(x, y, t) - S(a_i(t)) \tag{13.19}$$

$$e(x, y, t) = I(x, y, t) - \sum_i a_i(t) * \phi_i(x, y, t) \ . \tag{13.20}$$

where $\star$ denotes correlation. Note however that in order to be considered a causal system, $\phi(x, y, \tau)$ must be zero for $t > 0$. For now though we shall overlook the issue of causality, and in the discussion we shall consider some ways of dealing with this issue.

This model differs from the ICA (independent components analysis) model for time-varying images proposed earlier by van Hateren and Ruderman (1998) in an important respect: namely, the basis functions are applied to the image sequence in a shift-invariant manner, rather than in a blocked fashion. In van Hateren and Ruderman's ICA model, training data is obtained by extracting blocks of size 12x12 pixels and 12 samples in time from a larger movie, and a set of basis functions were sought that maximize independence among the coefficients (by seeking extrema of kurtosis) averaged over many such blocks. An image block is described via

$$I(x, y, t) = \sum_i a_i \, \phi_i(x, y, t) \ . \tag{13.21}$$

and the coefficients are computed by multiplying the rows of the pseudo-inverse of $\mathbf{\Phi}$ with each block extracted from the image stream (akin to convolution). By contrast, our model assumes shift-invariance among the basis functions—i.e., a basis function may be applied to describe structure occurring at any point in time in the image sequence. In addition, since the basis set is overcomplete, the coefficients may be sparsified, giving rise to a non-linear, spike-like representation that is qualitatively different from that obtained via linear convolution (see "Results from natural movie sequences").

Learning

The objective function for adapting the basis functions is again the code length $\mathcal{L}$,

$$\mathcal{L} = -\langle \log P(\mathbf{I}|\theta) \rangle \tag{13.22}$$

$$P(\mathbf{I}|\theta) = \int P(\mathbf{I}|\mathbf{a},\theta) P(\mathbf{a}|\theta) d\mathbf{a} \tag{13.23}$$

where now the image likelihood and prior are defined as

$$P(\mathbf{I}|\mathbf{a},\theta) = \frac{1}{Z_{\lambda_N}} e^{-\frac{\lambda_N}{2}|I(x,y,t) - \sum_i a_i(t) * \phi_i(x,y,t)|^2} \tag{13.24}$$

$$P(\mathbf{a}|\theta) = \prod_{i,t} \frac{1}{Z_S} e^{-S(a_i(t))} \tag{13.25}$$

and θ refers to the model parameters ϕ_i, λ_N, and $S()$.

By using the same approximation to the true gradient of $\mathcal{L}$ discussed in the previous section, the update rule for the basis functions is then

$$\Delta\phi_i(x,y,\tau) \propto a_i(\tau) \star e(x,y,\tau) \tag{13.26}$$

Thus, the basis functions are adapted over space and time by Hebbian learning between the time-varying residual image and the time-varying coefficient activities.

Results from natural movie sequences

The model was trained on moving image sequences obtained from Hans van Hateren's natural movie database (`http://hlab.phys.rug.nl/vidlib/vid_db`). The movies were first whitened by a filter that was derived from the inverse spatio-temporal amplitude spectrum, and lowpass filtered with a cutoff at 80% of the Nyquist frequency in space and time (see also Dong & Atick, 1995, for a similar whitening procedure). Training was done in batch mode by loading a 128×128 pixel, 64 frame sequence into memory and randomly extracting a spatial subimage of the same temporal length. The coefficients were fitted to this sequence by maximizing the posterior distribution via eqs. (13.19) and (13.20). The statistics for learning were averaged over ten such subimage sequences and the basis functions were then updated according to (13.26), again subject to rescaling (13.15). After several hours of training on a 450Mhz Pentium, the solution reached equilibrium.

The results for a set of 96 basis functions, each 8x8 pixels and of length 5 in time, are shown in figure 13.6. Spatially, they share many of the same characteristics of the basis functions obtained previously with static images (figure 13.3). The main dif-

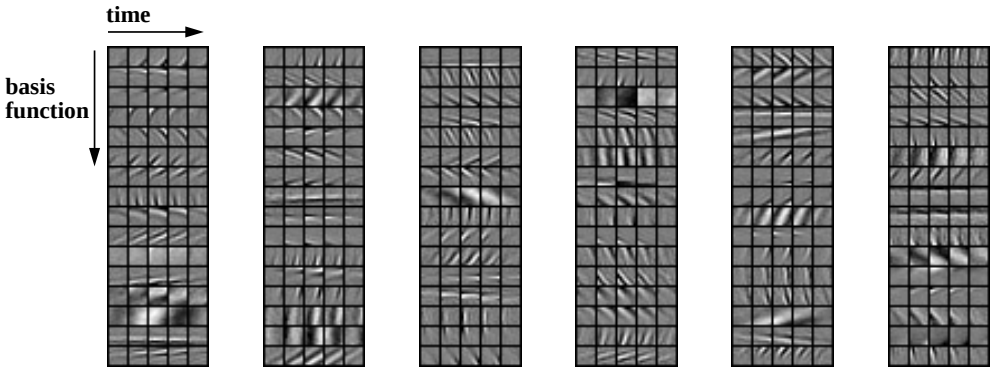

Figure 13.6. Space-time basis functions learned from time-varying natural images. Shown are a set of 96 basis functions arranged into six columns of 16 each. Each basis function is 8×8 pixels in space and 5 frames in time. Each row shows a different basis function, with time proceeding left to right. The translating character of the functions is best viewed as a movie, which may be viewed at `http://redwood.ucdavis.edu/bruno/bfmovie/bfmovie.html`.

ference is that they now also have a temporal characteristic, such that they tend to *translate* over time. Thus, the vast majority of the basis functions are direction selective (i.e., their coefficients will respond only to edges moving in one direction), with the high spatial-frequency functions biased toward lower velocities. These properties are typical of the space-time receptive fields of V1 simple-cells (Jones & Palmer, 1989; DeAngelis et al., 1995), and also of those obtained previously with ICA (van Hateren & Ruderman, 1998).

Because the outputs of the model are sparsified over both space and time, the model yields a qualitatively different behavior than linear convolution, as in ICA. Figure 13.7 illustrates this difference by comparing the time-varying coefficients obtained by maximizing the posterior to those obtained by straightforward convolution (similar to the linear prediction discussed in the previous section). The difference is striking in that the sparsified representation is characterized by highly localized, punctate events. Although still analog, it bears a strong resemblance to the the spiking nature of neural activity. At present though, this comparison is merely qualitative.

Discussion

We have shown in this chapter how both the spatial and temporal response properties of neurons may be understood in terms of a probabilistic model which attempts to describe images in terms of sparse, independent events. When the model is adapted to time-varying natural images, the basis functions converge upon a set of

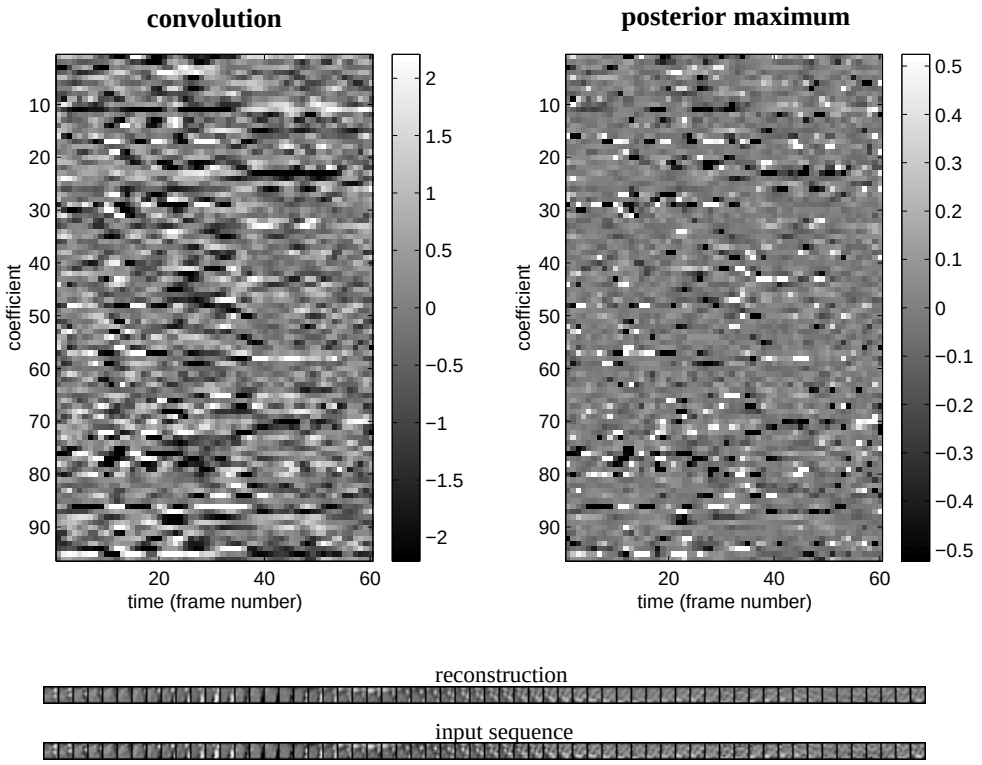

Figure 13.7. Coefficients computed by convolving the basis functions with the image sequence (*left*) vs. posterior maximization (*right*) for a 60 frame image sequence (*bottom*).

space-time functions which are spatially Gabor-like and translate with time. Moreover, the sparsified representation has a spike-like character, in that the coefficient signals are mostly zero and tend to concentrate their non-zero activity into brief, punctate events. These brief events represent longer spatiotemporal events in the image via the basis functions. The results suggest, then, that both the *receptive fields* and *spiking activity* of V1 neurons may be explained in terms of a single principle, that of sparse coding in time.

The interpretation of neural spike trains as a sparse code in time is not new. Most recently, Bialek and colleagues have shown that sensory neurons in the fly visual system, frog auditory system, and the cricket cercal system, essentially employ about one spike per "correlation time" to encode time-varying signals in their environment (Rieke et al., 1997). In fact, the image model proposed here is identical to their linear stimulus reconstruction framework used for measuring the mutual information between neural activity and sensory signals. The main contribution of this paper, beyond this previous body of work, is in showing that the particular spatiotemporal

receptive field structures of V1 neurons may actually be *derived* from such sparse, spike-like representations of natural images.

This work also shares much in common with Lewicki's shift-invariant model of auditory signals, discussed in the preceding chapter in this book. The main difference is that Lewicki's model utilizes a much higher degree of overcompleteness, which allows for a more precise alignment of the basis functions with features occurring in natural sounds. Presumably, increasing the degree of overcompleteness in our model would yield even higher degrees of sparsity and basis functions that are even more specialized for the spatio-temporal features occurring in images. But learning becomes problematic in this case because of the difficulties inherent in properly maximizing or sampling from the posterior distribution over the coefficients. The development of efficient methods for sampling from the posterior is thus an important goal of future work.

Another important yet unresolved issue in implementing the model is how to deal with causality. Currently, the coefficients are computed by taking into account information both in the past and in the future in order to determine their optimal state. But obviously any physical implementation would require that the outputs be computed based only on past information. The fact that the basis functions become two-sided in time (i.e., non-zero values for both negative and positive time) indicates that a coefficient at time t_0 is making a statement about the image structure expected in the future ($t > t_0$). This fact could possibly be exploited in order to make the model predictive. That is, by committing to respond at the present time, based only on what has happened in the past, a unit will be making a prediction about what is to happen a short time in the future. An additional challenge in learning, then, is to adapt an appropriate decision function for determining when a unit should become active, so that each unit serves as a good predictor of future image structure in addition to being sparse.

Acknowledgements

This work benefited from discussions with Mike Lewicki and was was supported by NIMH grant R29-MH57921. I am also indebted to Hans van Hateren for making his natural movie database freely available.

References

[1] Albrecht DG, Hamilton DB (1982) Striate cortex of monkey and cat: Contrast response function. *Journal of Neurophysiology, 48*: 217-237.

[2] Atick JJ, Redlich AN (1992) What does the retina know about natural scenes?

Neural Computation, 4: 196-210.

[3] Barlow HB (1961) Possible principles underlying the transformations of sensory messages. In: *Sensory Communication*, W.A. Rosenblith, ed., MIT Press, pp. 217-234.

[4] Barlow HB (1989) Unsupervised learning, *Neural Computation, 1*: 295-311.

[5] Baum EB, Moody J, Wilczek F (1988) Internal representations for associative memory, *Biological Cybernetics, 59*: 217-228.

[6] Bell AJ, Sejnowski TJ (1997) The independent components of natural images are edge filters, *Vision Research, 37*: 3327-3338.

[7] DeAngelis GC, Ohzawa I, Freeman RD (1995) Receptive-field dynamics in the central visual pathways. *Trends in Neurosciences, 18(10)*, 451-458.

[8] Dong DW, Atick JJ (1995) Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus, *Network: Computation in Neural Systems, 6*: 159-178.

[9] Field DJ (1994) What is the goal of sensory coding? *Neural Computation, 6*: 559-601.

[10] Foldiak P (1995) Sparse coding in the primate cortex, In: *The Handbook of Brain Theory and Neural Networks*, Arbib MA, ed, MIT Press, pp. 895-989.

[11] Lewicki MS, Olshausen BA (1999) Probabilistic framework for the adaptation and comparison of image codes, *J. Opt. Soc. of Am., A, 16(7)*: 1587-1601.

[12] Lewicki MS, Sejnowski TJ (2000) Learning overcomplete representations. *Neural Computation, 12*:337-365.

[13] McLean J, Palmer LA (1989) Contribution of linear spatiotemporal receptive field structure to velocity selectivity of simple cells in area 17 of cat. *Vision Research, 29(6)*:675-9.

[14] Mumford D (1994) Neuronal architectures for pattern-theoretic problems, In: *Large Scale Neuronal Theories of the Brain*, Koch C, Davis, JL, eds., MIT Press, pp. 125-152.

[15] Olshausen BA, Field DJ (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research, 37*, 3311-3325.

[16] Olshausen BA, Millman KJ (2000). Learning sparse codes with a mixture-of-Gaussians prior. In: *Advances in Neural Information Processing Systems, 12*, S.A. Solla, T.K. Leen, K.R. Muller, eds. MIT Press, pp. 841-847.

[17] Rieke F, Warland D, de Ruyter van Stevenick R, Bialek W (1997) *Spikes: Exploring the Neural Code*. MIT Press.

[18] Simoncelli EP, Freeman WT, Adelson EH, Heeger DJ (1992) Shiftable multiscale transforms, *IEEE Transactions on Information Theory, 38(2)*: 587-607.

[19] Tadmor Y, Tolhurst DJ (1989) The effect of threshold on the relationship between the receptive field profile and the spatial-frequency tuning curve in sim-

ple cells of the cat's striate cortex, *Visual Neuroscience, 3*: 445-454.

[20] van Hateren JH, van der Schaaff A (1998) Independent component filters of natural images compared with simple cells in primary visual cortex, *Proc. Royal Soc. Lond. B, 265*: 359-366.

[21] van Hateren JH, Ruderman DL (1998) Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. Proc.R.Soc.Lond. B, 265:2315-2320.