SPARSE CODING OF TIME-VARYING NATURAL IMAGES

Bruno A. Olshausen

Center for Neuroscience and
Department of Psychology, UC Davis
Davis, California 95616
baolshausen@ucdavis.edu

ABSTRACT

We show how the principle of sparse coding may be applied to learn the forms of structure occurring in time-varying natural images. A sequence of images is described as a linear superposition of space-time functions, each of which is convolved with a time-varying coefficient signal. When a sparse, independent representation is sought over the coefficients, the basis functions that emerge are space-time inseparable functions that resemble the motion-selective receptive fields of cortical simple cells. Interestingly, the coefficients form a spike-like representation of moving images, and thus suggest an interpretation of spiking activity in the brain in terms of sparse coding in time.

1. INTRODUCTION

The images that fall upon the retina change over time due to motions of the eye and body, as well as the motions of objects in the world. Since static natural images contain many forms of spatial structure, these time-varying images should also contain interesting forms of structure across both space and time. In this paper, we attempt to model the forms of structure that occur in time-varying natural images in terms of a linear superposition sparse, independent events.

Previous work characterizing the second-order statistics of time-varying natural images—i.e., the space-time autocorrelation function, or the spatio-temporal power spectrum—has shown that they contain forms of structure that can be related to the well-known $1/f^2$ power spectrum of static natural images. Namely, Atick and Dong (1995) showed that there is an interesting dependence between spatial and temporal frequency that may be explained by the distribution of moving objects in the environment. They also showed that this spectral structure may be removed by the type of response properties seen in LGN neurons (Dong & Atick 1995).

Our goal in this paper is to see what may be learned from examining the higher-order statistical structure in time-varying natural images (i.e., beyond pairwise statistics), and how this may relate to the response properties of cortical neurons. In earlier work (Olshausen & Field, 1996; 1997), we have shown that the higher-order forms of structure in static images may be characterized in part by a linear generative model that attempts to describe images in terms of sparse, statistically independent components. That is, an image $I(x, y)$ is modeled as a linear superposition of basis functions, $\phi_i(x, y)$, multiplied by coefficients, a_i:

$$I(x, y) = \sum_i a_i \phi_i(x, y) \ . \qquad (1)$$

When a set of basis functions is sought such that the coefficients are as sparse and statistically independent as possible, averaged over many natural images, the basis functions that emerge are localized, oriented, and bandpass (selective to structure at different spatial scales). These properties are similar to the receptive fields of neurons in mammalian primary visual cortex, and thus suggest that the cortex has evolved according to a similar coding principle.

Recently, van Hateren and Ruderman (1998) applied the technique of independent components analysis (ICA) , which is intimately related to our sparse coding model, to natural images sequences. They showed that the receptive fields that emerge from ICA are localized in space and time and resemble the non-separable space-time receptive fields of cortical simple cells. Thus, it would seem that the spatiotemporal receptive field properties of cortical neurons also adhere to a statistically sensible coding strategy. However, a potential problem arises in directly relating their results to the visual cortex because the algorithm was applied in a blocked fashion to moving image sequences. Blocks of

Supported by NIMH R29-MH057921. We thank Hans van Hateren for making available his natural image and movie database (http://hlab.phys.rug.nl/archive.html).

size 12x12 pixels and 12 samples in time were extracted at random from a larger movie, and a set of basis functions was sought so as to maximize independence (by seeking extrema of kurtosis) among the coefficients averaged over many such blocks. The consequence of applying ICA in this manner is that there is no explicit representation of time among the coefficients, since an image block $I(x, y, t)$ is described via

$$I(x, y, t) = \sum_i a_i \, \phi_i(x, y, t) \; . \tag{2}$$

Thus, while the coefficients may be independent of each other over an ensemble of image blocks, there is nothing forcing them to be independent of themselves over time because there is no notion of time represented in the coefficients of this model. In addition, although the full dimensionality of the data space is 1728 (12^3), only 288 components were learned. This does not constitute a complete representation, and so it is difficult to draw any conclusions from how the basis functions choose to tile space and time.

Here, we analyze natural image sequences in terms of sparse, independent components, but without any blocking of the image stream. Time-varying images are modeled as a linear superposition of basis functions, but importantly with functions that are time-invariant (so that each function can be applied at any point in time). A complete set of basis functions is sought so as to yield a code that is as sparse and independent as possible across both time and space. We show that the basic form of the learned basis functions is very similar to that obtained with ICA, the main difference being that the coefficients are now explicitly modeled as time-varying functions. As we shall see, this leads us to an interpretation of neural spike trains in terms of a *sparse code in time* (Rieke et al., 1997).

2. MODEL

A time varying image, $I(x, y, t)$, is modeled as a linear superposition of basis functions, $\phi_i(x, y, \tau)$, where each basis function is localized in time but can be applied at any instant during the image sequence:

$$\begin{aligned} I(x, y, t) &= \sum_i \sum_{t'} a_i(t') \, \phi_i(x, y, t - t') + \nu(x, y, t) \\ &= \sum_i a_i(t) * \phi_i(x, y, t) + \nu(x, y, t) \end{aligned} \tag{3}$$

where $*$ denotes convolution over time. Thus, the time-varying coefficient, $a_i(t)$, tells us the amount by which basis function ϕ_i is multiplied to model the structure around time t in the moving image sequence. The noise $\nu(x, y, t)$ is used to model additional uncertainty not captured by this model. Importantly, we examine here the case where the image code is overcomplete, meaning that the number of coefficient signals $a_i(t)$ exceeds the dimensionality of the movie $I(x, y, t)$. The model is illustrated schematically in figure 1.

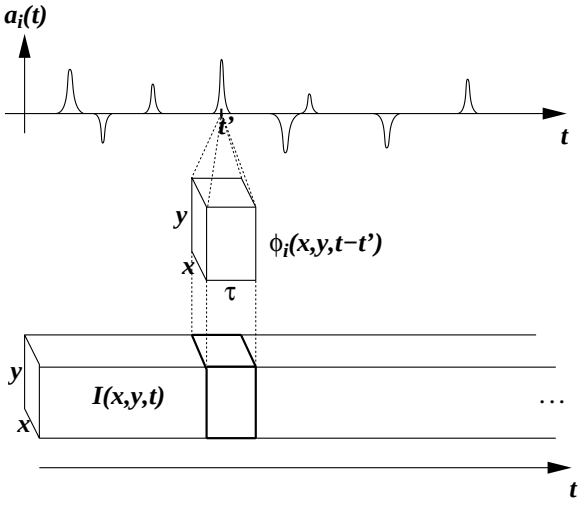

Figure 1: Image model. A movie $I(x, y, t)$ is modeled as a linear superposition of spatio-temporal basis functions, $\phi_i(x, y, \tau)$, each of which is localized in time but may be applied at any time within the movie sequence.

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

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} P(\mathbf{a}|\mathbf{I}, \theta) \tag{4}$$

$$= \arg\max_{\mathbf{a}} P(\mathbf{I}|\mathbf{a}, \theta) P(\mathbf{a}|\theta) \tag{5}$$

where θ denotes the model parameters. The image likelihood $P(\mathbf{I}|\mathbf{a}, \theta)$ is Gaussian (assuming Gaussian noise ν)

$$P(\mathbf{I}|\mathbf{a}, \theta) = \frac{1}{Z_{\lambda_N}} e^{-\frac{\lambda_N}{2} |I(x,y,t) - \sum_i a_i(t) * \phi_i(x,y,t)|^2} \tag{6}$$

and λ_N is the inverse of the noise variance. The prior probability distribution is specified to be factorial (i.e., statistical independence) over both coefficients and time, and the marginal distribution of each coefficient is assumed to be sparse

$$P(\mathbf{a}|\theta) = \prod_{i,t} P(a_i(t)) \tag{7}$$

$$P(a_i(t)) = \frac{1}{Z_S} e^{-\beta \, S(a_i(t)/\sigma)} \tag{8}$$

where S is a non-convex function appropriate for shaping the prior to be of sparse form (i.e., more peaked at zero and with heavy tails as compared to a Gaussian of the same variance, as shown in figure 2). Here we use $S(x) = \beta \log(1 + (x/\sigma)^2)$, where σ is a scaling parameter, and β controls the degree of sparseness.

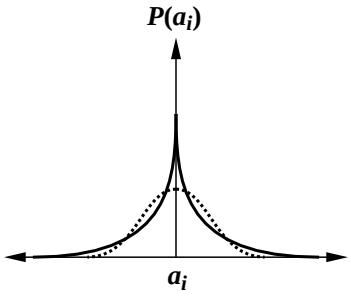

Figure 2: The prior probability distribution over the coefficients is peaked at zero with heavy tails as compared to a Gaussian of the same variance (overlaid as dashed line). Such a distribution would result from a sparse activity distribution over the coefficients.

Maximizing the the posterior distribution over the coefficients is equivalent to minimizing $-\log P(\mathbf{a}|\mathbf{I},\theta)$

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \left[\tfrac{\lambda_N}{2} |I(x,y,t) - \textstyle\sum_i a_i(t) * \phi_i(x,y,t)|^2 + \beta \textstyle\sum_i \sum_t S(a_i(t)/\sigma)\right] \quad (9)$$

which may be accomplished by gradient descent, yielding the following differential equation for determining the coefficients:

$$\dot{a}_i(t) \propto \lambda_N \sum_{x,y} \phi_i(x,y,t) \star e(x,y,t) - \beta/\sigma S'(a_i(t)/\sigma) \quad (10)$$

$$e(x,y,t) = I(x,y,t) - \sum_i a_i(t) * \phi_i(x,y,t) \quad (11)$$

where $\star$ denotes correlation over time. Note however that in order to be considered a causal system, $\phi(x,y,t)$ must be zero for $t > 0$. For the purposes of this paper though we shall overlook the issue of causality and focus on what may be learned from sparse coding of time-varying images per se.

3. LEARNING

The objective function for learning the basis functions is the average code length of the images under the model

$$\mathcal{L} = -\langle \log P(\mathbf{I}|\theta) \rangle \quad (12)$$

where

$$P(\mathbf{I}|\theta) = \int P(\mathbf{I}|\mathbf{a},\theta) P(\mathbf{a}|\theta) d\mathbf{a} \; . \quad (13)$$

$\mathcal{L}$ is minimized by gradient descent, yielding the following Hebbian update rule:

$$\Delta\phi_i(x,y,t) \propto \frac{\partial \mathcal{L}}{\partial \phi_i(x,y,t)} \quad (14)$$

$$= \left\langle \langle a_i(t) \star e(x,y,t) \rangle_{P(\mathbf{a}|\mathbf{I},\theta)} \right\rangle \; . \quad (15)$$

Thus, the basis functions are updated by an amount proportional to the correlation between the residual error $\mathbf{e}$ and the coefficients $\mathbf{a}$. Instead of sampling from the full posterior distribution, though, we utilize a simpler approximation in which a single sample is taken at the posterior maximum, and so we have

$$\Delta\mathbf{\Phi} \propto \langle \hat{a}_i(t) \star e(x,y,t) \rangle \; . \quad (16)$$

The price we pay for this approximation, though, 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 (9). This trivial solution is avoided by rescaling the basis functions after each learning step (16) so that their L2 norm, $g_i = |\phi_i|_{L2}$, maintains an appropriate level of variance on each corresponding coefficient a_i:

$$g_i^{new} = g_i^{old} \left[\frac{\langle a_i^2 \rangle}{\sigma^2}\right]^{\alpha} \; , \quad (17)$$

where σ is the scaling parameter used in the sparse cost function. 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).

4. RESULTS

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 images 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. 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 via eqs. 10 and 11. The statistics for learning were averaged over ten such subimage sequences and the basis functions were then updated according to equation 16. After several hours

of training on a 450Mhz Pentium, the solution reached equilibrium.

The results for a set of 96 basis functions, each 8×8 pixels and of length 5 in time, are shown in figure 3. These functions are similar to those obtained earlier with ICA in that they become localized in both space and time. All are direction selective, with the low spatial-frequency functions becoming selective to high velocities and the high spatial-frequency functions selective to low velocities. The entire set of basis functions spans the joint space of position, orientation, spatial-frequency, and velocity sufficiently to provide good reconstructions (figure 4).

In comparing the form of the basis functions to neural receptive fields, it is important to bear in mind that the basis functions of the model do not exactly tell us the same thing as the receptive field of the neuron. The typical model of a simple-cell receptive field is that its response is determined via linear projection (i.e., inner product) of the time-varying image onto the space-time receptive field function. That is, for a space-time receptive field function $\psi(x, y, \tau)$, the response of the neuron at time t, $r(t)$, would be computed from the image $I(x, y, t)$ via

$$r(t) = \sum_{x,y,t'} \psi(x, y, t' - t)\, I(x, y, t') \; . \qquad (18)$$

Note however that in our image model (3), the activity of a neuron at time t, $a_i(t)$ signifies how much of the corresponding basis function $\phi_i(x, y, \tau)$ is contained in the image, and the neuron's activity is determined by maximizing the posterior distribution over its state (9). The effect of this computation is to sparsify neural activity, as shown in figure 4. When the solution obtained by maximizing the posterior is compared to that obtained by simply convolving the basis functions with the image (in which they are acting simply as linear filters), it is clear that the former produces a much sparser, spike-like representation. Thus, the non-linearities inherent in the sparsification process produce a time-localized representation of the spatio-temporal events occurring in the signal, such as moving edges, changes in luminance, etc.

5. CONCLUSIONS

We have shown in this work how natural image sequences can be decomposed into a superposition of sparse, spatiotemporal events. An 8x8 pixel movie is re-represented as a stream of 96 analog signals that are sparse over both space (i.e., across the ensemble of coefficients) and time. 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 events. These brief events represent longer spatiotemporal events in the image via the basis functions, which resemble the space-time receptive fields of cortical simple cells. It is thus suggested that 1) the reason cortical simple-cells have the space-time receptive fields they do is because it allows for time-varying natural images to be represented in terms of a collection of sparse temporal events, 2) these sparse temporal events may manifest themselves in visual cortical neurons in the form of action potentials, or spikes, and 3) if so, spike trains themselves can be thought of as a form a sparse code in time (Rieke et al., 1997), which provides a more efficient representation of visual information.

Do V1 neurons actually work in this way? One possibility for testing this hypothesis is to measure the space-time receptive field of a V1 simple-cell via reverse-correlation using an unstructured image ensemble (e.g., white noise, or spots) and then use the measured receptive field to predict the response of the neuron to natural images. If a form of sparsification were occurring, then one would expect the actual response obtained with natural images to yield a sparser output distribution than expected from simply convolving the measured receptive field with the image. The reason for this is that the artificial image ensembles do not contain structures that match well any of the basis functions in the model. Thus, the representation does not get appreciably sparsified, and so the basis functions tend to reveal themselves in the measured receptive field (actually, it is the pseudo-inverse of the basis function matrix, but since most of the basis functions are near to orthogonal this resembles fairly well the basis functions). When natural images are used as the stimulus ensemble, then the features in these images match well with the basis functions (as they were designed to do), and so the sparsification non-linearity kicks in. Note that the sparsification effect is different from gain control in that it does not involve a simple universal gain adjustment on the population of responses, but rather attenuates low-valued coefficients more severely than high-valued coefficients. Which coefficients survive this self-inhibition depends on how well each basis function matches a given image feature relative to the others.

An important but unresolved issue in implementing this scheme is that of causality. As the coefficients are currently computed, they have the advantage of being able to look both backwards and forwards in time in order to determine their optimal state. But in a real physical system, signals can be determined only based on the past and present activity of themselves

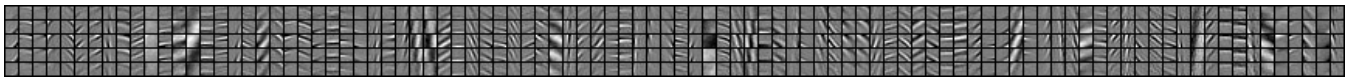

Figure 3: Learned basis functions. Each column is a different basis function, with time going downwards. An animation may be downloaded from `ftp://redwood.ucdavis.edu/pub/bfmovie.mpg`.

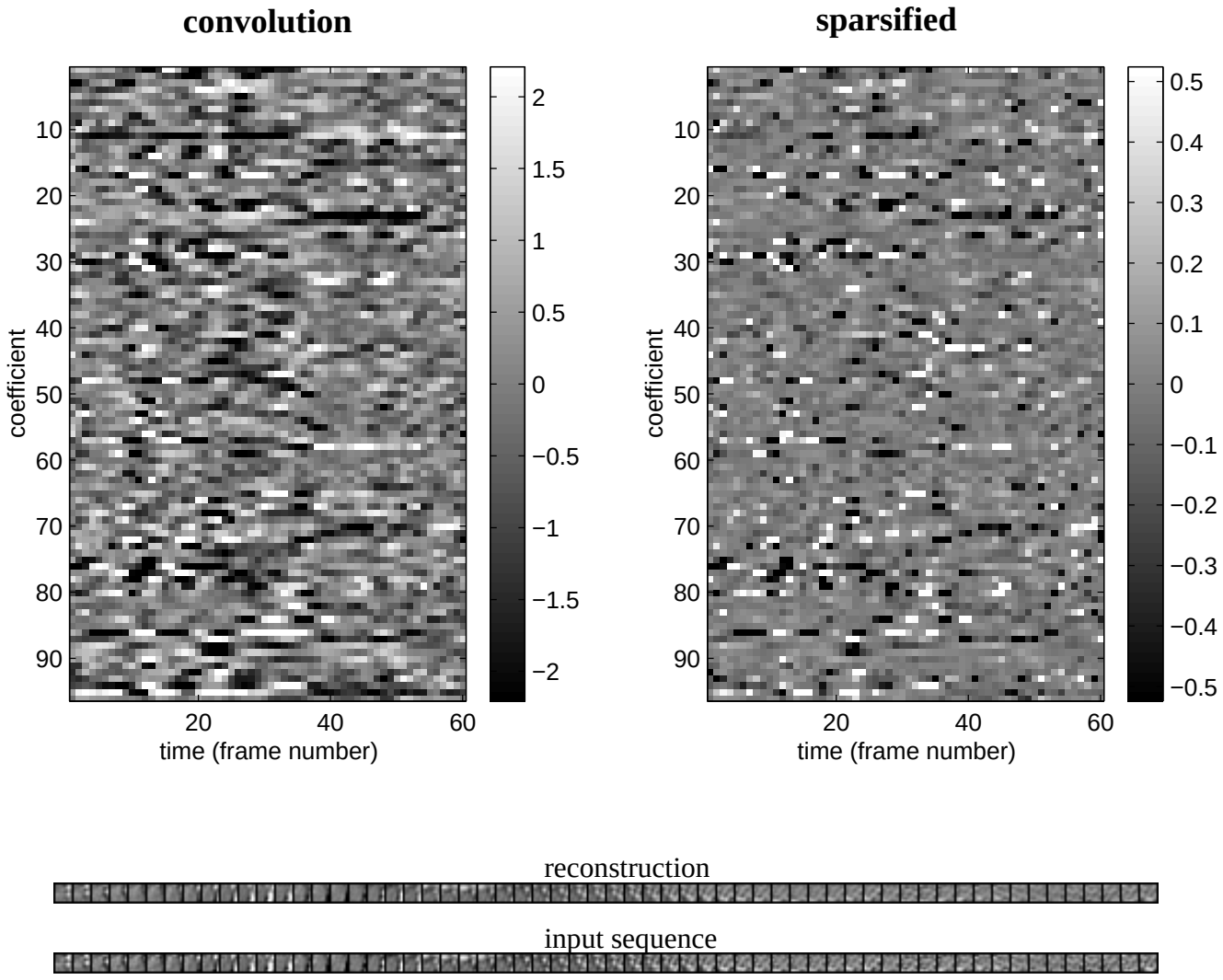

Figure 4: Coefficients computed via convolution (left) and sparsified (right) for the 60-frame image sequence shown (at bottom). The sparsified coefficients yield a much more time-localized representation of spatio-temporal events in the image sequence than does linear convolution.

and others. Thus, it will be necessary to modify the current model in order to be predictive about future events based upon present and past activity. This is the focus of current research.

6. REFERENCES

Atick JJ, Dong DW (1995) Statistics of natural time-varying images. *Network, 6*, 345-358.

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

Olshausen BA, Field DJ (1996). Natural image statistics and efficient coding. *Network, 7*, 333-339.

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

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

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.