IEEE International Conference on Image Processing. Sept. 14-17, 2003. Barcelona, Spain.

LEARNING SPARSE, OVERCOMPLETE REPRESENTATIONS OF TIME-VARYING NATURAL IMAGES

Bruno A. Olshausen

Redwood Neuroscience Institute
Menlo Park, CA 94025
and Center for Neuroscience, UC Davis
Davis, California 95616
baolshausen@ucdavis.edu

ABSTRACT

I show how to adapt an overcomplete dictionary of space-time functions so as to represent time-varying natural images with maximum sparsity. The basis functions are considered as part of a probabilistic model of image sequences, with a sparse prior imposed over the coefficients. Learning is accomplished by maximizing the log-likelihood of the model, using natural movies as training data. The basis functions that emerge are space-time inseparable functions that resemble the motion-selective receptive fields of simple-cells in mammalian visual cortex. When the coefficients are computed via matching-pursuit in space and time, one obtains a punctate, spike-like representation of continuous time-varying images. It is suggested that such a coding scheme may be at work in the visual cortex.

1. INTRODUCTION

Time-varying images present a challenge for efficient coding and compression, as one must consider how to best deal with the redundancies contained in natural images over both space and time. Many of the currently employed coding schemes are derived from rather casual observations about the structure of time-varying images. For example, MPEG relies upon estimating the motion from frame to frame and then coding the image displacement and residual error. But there are many different ways to do this—which is optimal for natural image sequences? And how do we even know that motion estimation is the right way to formulate the problem in the first place?

The approach taken here is to *learn* the best way to represent time-varying images by appealing to the principle of sparseness. That is, we would like to find a "vocabulary" for describing natural image sequences such that the number of words needed to describe what is going on at any point in time is small (although the number of words in the vocabulary itself may be quite large). The idea is that the words will be tailored to the common space-time structures occuring in natural image sequences, thus providing a natural and efficient way to represent time-varying images.

In earlier work [2,3], we used this approach in an attempt to account for the spatial receptive properties of neurons in the primary visual cortex of mammals. A small image patch, $I(x, y)$, is modeled as a linear superposition of basis functions, $\phi_i(x, y)$, multiplied by coefficients, a_i:

$$I(x, y) = \sum_i a_i \phi_i(x, y) \ . \tag{1}$$

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 (area V1), thus suggesting that the cortex has evolved according to a similar coding principle.

van Hateren and Ruderman [5] extended this idea to the time domain and showed that the basis functions that emerge have similar spatial properties and translate as a function of time, similar to the non-separable (direction selective) receptive fields of cortical simple cells. However, their image model relies upon blocking the image stream into a small number of frames and treating time simply as another dimension:

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

An image sequence is then represented by simply computing inner products between a set of biorthogonal functions and a block of image frames.

Here we model time-varying images without blocking by assuming time-invariance in the basis functions, so that

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

each function can be applied at all points in time. Importantly, the basis set is overcomplete, so that there are multiple ways to describe a given image sequence. When a sparse representation is selected via matching pursuit, then one obtains a recoding of the image in terms of sparse, punctate events in time, similar to neural spike trains. The suggestion is that the spike trains of V1 neurons themselves serve as a sparse code in time, and that V1 receptive fields have been adapted to represent images in this way.

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:

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

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 term $\nu(x, y, t)$ is used to model additional structure not well described 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.

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

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

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

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

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

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

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

$$P(a_i(t)) = \frac{1}{Z_S} e^{-S(a_i(t))} \quad (8)$$

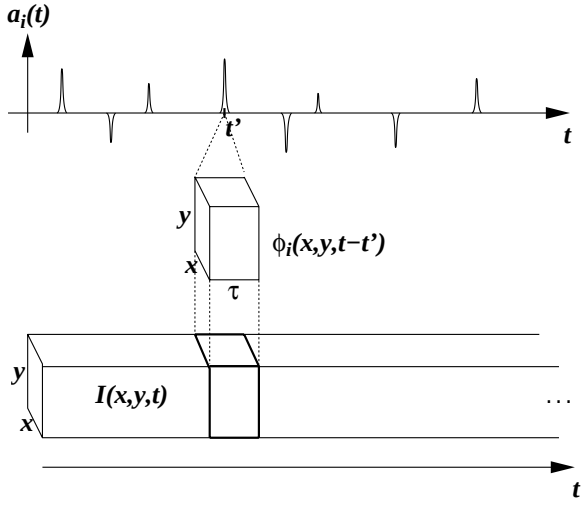

Fig. 1. Image model. A movie $I(x, y, t)$ is modeled as a linear superposition of spatio-temporal basis functions, $\phi_i(x, y, \tau)$, each of which is localized in time but may be applied at any time within the movie sequence.

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.

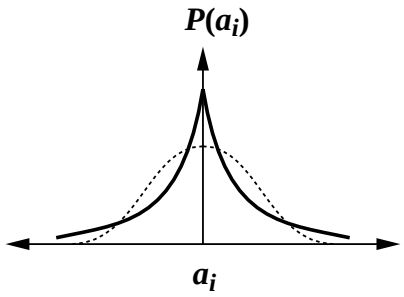

Fig. 2. The prior probability distribution over the coefficients is peaked at zero with heavy tails as compared to a Gaussian of the same variance (overlaid as dashed line). Such a distribution would result from a sparse activity distribution over the coefficients.

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

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

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

$$\begin{aligned} \dot{a}_i(t) &\propto \lambda_N \sum_{x,y} \phi_i(x,y,t) \star e(x,y,t) - S'(a_i(t)) \\ e(x,y,t) &= I(x,y,t) - \sum_i a_i(t) * \phi_i(x,y,t) \end{aligned} \tag{10}$$

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 now 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 log-likelihood of the model

$$\mathcal{L} = \langle \log P(\mathbf{I}|\theta) \rangle \tag{11}$$

where

$$P(\mathbf{I}|\theta) = \int P(\mathbf{I}|\mathbf{a},\theta) P(\mathbf{a}|\theta) d\mathbf{a} \; . \tag{12}$$

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

$$\Delta\phi_i(x,y,t) \propto \frac{\partial \mathcal{L}}{\partial \phi_i(x,y,t)} \tag{13}$$

$$= \left\langle \langle a_i(t) \star e(x,y,t) \rangle_{P(\mathbf{a}|\mathbf{I},\theta)} \right\rangle \; . \tag{14}$$

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

$$\Delta\mathbf{\Phi} \propto \langle \hat{a}_i(t) \star e(x,y,t) \rangle \; . \tag{15}$$

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 (15) so that their L2 norm, $g_i = |\phi_i|_{L2}$, maintains an appropriate level of variance on each corresponding coefficient a_i:

$$g_i^{new} = g_i^{old} \left[\frac{\langle a_i^2 \rangle}{\sigma^2} \right]^{\alpha} \; , \tag{16}$$

where σ is the scaling parameter used in the sparse cost function and α is the rate of adaptation.

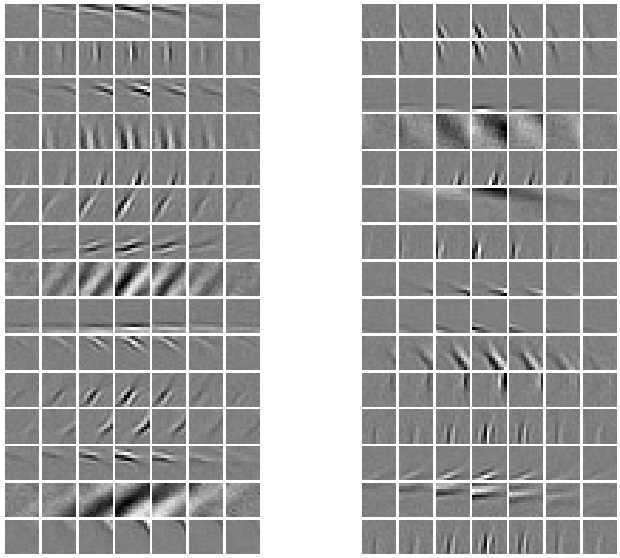

Fig. 3. Space-time basis functions learned from time-varying natural images. Shown are 30 basis functions randomly selected from the entire set of 200 functions learned, arranged into two columns. Each basis function is 12×12 pixels in space and 7 frames in time. Each row shows a different basis function, with time proceeding left to right. An animation may be downloaded from `http://redwood.ucdavis.edu/bruno/research/bfmovie.avi`.

4. RESULTS

The model was trained on moving image sequences obtained from a natural movie database [6]. 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 eq. 10. The statistics for learning were averaged over ten such subimage sequences and the basis functions were then updated according to equation 15. After several hours of training the solution reached equilibrium.

The results for a set of 200 basis functions, each 12×12 pixels and 7 frames in time, are shown in figure 3. These functions are similar to those obtained earlier with ICA [5]. All are direction selective, with the high spatial-frequency functions biased towards slow speeds, as expected. The entire set of basis functions spans the joint space of position, orientation, spatial-frequency, and velocity, as shown in figure 4.

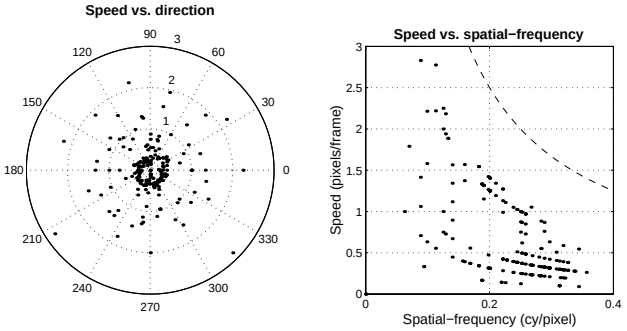

Fig. 4. Basis function tiling properties. Each data point denotes a different basis function. In the polar plot at left, radius denotes speed (in units of frames/sec) and angle denotes the direction in which the basis function translates. In the plot at right, the dashed line denotes the limit imposed by the Nyquist frequency (12.5 Hz). (The striated clustering is an artifact due to decimation in the spatiotemporal frequency domain.)

When the coefficients are computed via gradient descent (eq. 10), one obtains highly sparse representations of time-varying images. However, the coefficients never actually reach values exactly equal to zero, and so there is no clear distinction between active and inactive coefficients. This problem can be ameliorated by matching pursuit [1], yielding a representation that is clearly punctate in time, similar to neural spike trains (figure 5). Note that even though the image model itself is linear, the coding of images is highly nonlinear.

5. CONCLUSIONS

We have shown in this work how natural image sequences can be described in terms of a superposition of sparse, spatiotemporal events. A 12×12 pixel movie is re-represented as a stream of 200 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 *both* the receptive fields and the spiking nature of neural activity work hand in hand to achieve a sparse code in space and time [4], providing a more efficient representation of visual information.

An important but unresolved issue in implementing this scheme is that of causality. In the matching pursuit scheme,

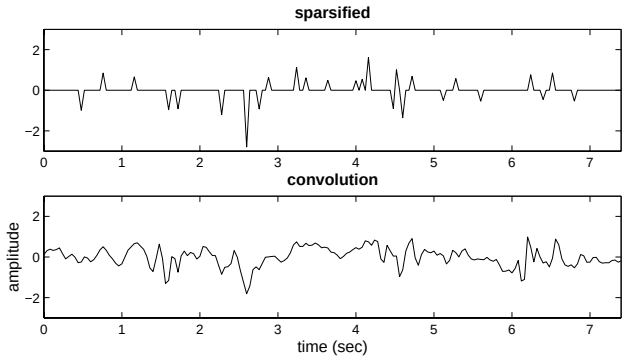

Fig. 5. Time-varying value of a single coefficient computed by sparsification (top) vs. convolving the basis functions with the image sequence (bottom) for a 7.4 second image sequence (25 f/s).

each coefficient has the advantage of being able to look both backwards and forwards in time in order to determine its optimal state. But in a real physical system, signals can be determined only based on the past and present activity of themselves and others. Thus, it will be necessary to modify the current model in to be predictive about future events based upon present and past activity in order to determine where to spike. This is the focus of current research.

6. REFERENCES

1. Mallat SG, Zhang Z (1993) Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing, 41*: 3397-3415.

2. Olshausen BA, Field DJ (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*, 607-609.

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

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

5. 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.

6. van Hateren (2002) Natural Stimuli Collection. `http://hlab.phys.rug.nl/vidlib/vid_db`