A Two-Layer Temporal Generative Model of Natural Video Exhibits Complex-Cell-Like Pooling of Simple Cell Outputs

Jarmo Hurri and Aapo Hyvärinen

Neural Networks Research Centre

Helsinki University of Technology, P.O.Box 9800, 02015 HUT, Finland

Abstract

A recently developed branch of computational neuroscience examines how the statistics of sensory input data are related to the structure and functional properties of corresponding sensory neural networks. In this branch it is assumed that the networks are tuned to the properties of input data, that is, that they have learned efficient internal representations of their environment. In this paper we present a hypothetical internal representation for natural video. We define a two-layer generative model for natural video, based on temporal relationships between simple cell outputs. Preliminary results of estimating the parameters of the model from natural data suggest that the learned temporal interactions between cell outputs are similar to complex cell pooling of simple cell outputs.

Email addresses: `jarmo.hurri@hut.fi` (Jarmo Hurri), `aapo.hyvarinen@hut.fi` (Aapo Hyvärinen).

Preprint submitted to Elsevier Science *8 February 2002*

1 Introduction

The functional role of simple and complex cells has puzzled scientists since their response properties and structure of their receptive fields was first mapped in the 1950s (see, e.g., [7]). The current view of sensory neural networks emphasizes learning and the relationship between the structure of the cells and the statistical properties of the information they process (see, e.g., [6,9]). We have previously shown [2] that a principle called *temporal coherence* [1,5,10] leads to the emergence of simple cell type receptive fields from *natural video*. Thus, temporal coherence provides an alternative to sparse coding [6] and independent component analysis [4,9] as a statistical computational principle behind the structure of the receptive fields. Temporal coherence is based on the idea that when processing temporal input, the representation changes as little as possible over time. The measure of temporal coherence used in [2] was *single-cell* temporal *response strength* correlation. That is, a receptive field (linear filter) was found by maximizing the correlation of response strengths of filter output at successive time points.

In this paper we extend the idea of temporal coherence to temporal dependencies between the responses of different cells. We first define a two-layer generative model for natural video. At the heart of this model is an autoregressive model of cell response strengths which captures inter-cell temporal dependencies. After defining the model we formulate a computable criterion for its estimation. We then show that the estimation of the model from natural data results in the emergence of receptive fields which resemble simple cells, and whose interactions are similar to complex-cell pooling properties. The most important contribution of this paper is this unsupervised pooling

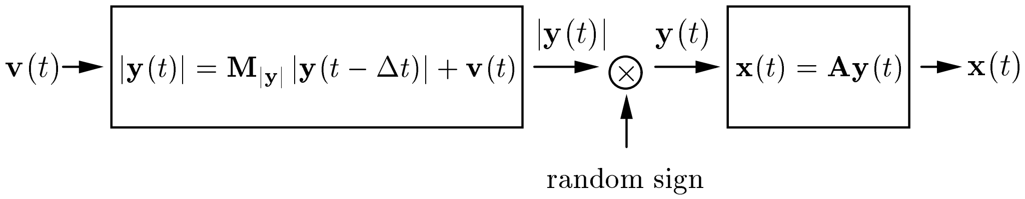
$\mathbf{v}(t)$
$|\mathbf{y}(t)| = \mathbf{M}_{|\mathbf{y}|}\,|\mathbf{y}(t-\Delta t)| + \mathbf{v}(t)$
$|\mathbf{y}(t)|$
$\otimes$
random sign
$\mathbf{y}(t)$
$\mathbf{x}(t) = \mathbf{A}\mathbf{y}(t)$
$\mathbf{x}(t)$

Fig. 1. The generative model. In the first layer the driving noise signal $\mathbf{v}(t)$ generates the amplitudes of simple cell responses, $|\mathbf{y}(t)|$, via an autoregressive model. The signs of simple cell responses are generated randomly between the first and second layer to yield cell responses $\mathbf{y}(t)$. In the second layer natural video $\mathbf{x}(t)$ is generated linearly from simple cell responses.

property of the model. This separates our model from other advanced self-organizing models of early vision, such as [3,5], in which this pooling must still be enforced somehow.

2 A two-layer generative model of natural video

The motivation behind the formulation of a generative model of natural perceptual input is the idea that the brain may utilize efficient internal models of the input it processes. The generative model of natural video introduced in this paper has two layers (see Fig. 1). The first layer is a multivariate autoregressive model relating a latent driving noise to the strengths (amplitudes) of simple cell responses at time t and time $t - \Delta t$. The signs of cell responses are generated by a second latent signal between the first and second layer. The second layer is linear, and maps cell responses to image features.

We start the description of the model with the second, linear layer. We restrict ourselves to linear spatial models of simple cells. Let vector $\mathbf{x}(t)$ denote an image taken from natural video at time t. (A vectorization of image patches can be done by scanning images column-wise into vectors.) Let $\mathbf{y}(t) = [y_1(t) \cdots y_K(t)]^T$ represent the outputs of K simple cells. The linear generative model for $\mathbf{x}(t)$ is similar to the one in [3,6]:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{y}(t).$$

Here $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_K]$ denotes a matrix that relates the image patch $\mathbf{x}(t)$ to the activities of the simple cells, so that each column $\mathbf{a}_k$, $k = 1, ..., K$, gives the feature that is coded by the corresponding simple cell. When the parameters of the model are estimated, what we actually obtain first is the mapping from $\mathbf{x}(t)$ to $\mathbf{y}(t)$, denoted by

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t), \tag{1}$$

Conceptually here the set of filters (vectors) $\mathbf{w}_1, ..., \mathbf{w}_K$ corresponds to the receptive fields of simple cells, and $\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_K]^T$ denotes a matrix with all the filters as rows. The dimension of $\mathbf{x}(t)$ is typically larger than the dimension of $\mathbf{y}(t)$, so that (1) is generally not invertible but an underdetermined set of linear equations. A one-to-one correspondence between $\mathbf{W}$ and $\mathbf{A}$ can be established by computing the pseudoinverse solution[1] $\mathbf{A} = \mathbf{W}^T(\mathbf{W}\mathbf{W}^T)^{-1}$.

In contrast to sparse coding [6] or independent component analysis [4] we do *not* assume that the components of $\mathbf{y}(t)$ are independent. Instead, we model the dependencies between these components with a multivariate autoregressive model in the first layer of our model. Let $|\mathbf{y}(t)|$ denote taking component-wise absolute values of $\mathbf{y}(t)$, and let $\mathbf{v}(t)$ denote a driving noise signal. Let

[1] When the solution is computed with the pseudoinverse, the solved $\mathbf{x}(t)$ is orthogonal to the nullspace of $\mathbf{W}$, $\mathcal{N}(\mathbf{W}) = \{\mathbf{b} \,|\, \mathbf{W}\mathbf{b} = \mathbf{0}\}$. In other words, that part of $\mathbf{x}(t)$ which would be ignored by the linear mapping in equation (1) is set to $\mathbf{0}$.

$\mathbf{M}_{|\mathbf{y}|}$ denote a square matrix with dimension K (the reason for the subscript notation is explained below). Our model is a *constrained multidimensional first-order autoregressive process*, specified by the following equation

$$|\mathbf{y}(t)| = \mathbf{M}_{|\mathbf{y}|} \, |\mathbf{y}(t - \Delta t)| + \mathbf{v}(t), \tag{2}$$

and unit variance and decorrelation constraints $\mathrm{E}_t \{y_{k_1}(t) y_{k_2}(t)\} = \delta(k_1 - k_2)$ for $k_1 = 1, ..., K$ and $k_2 = 1, ..., K$. Note that the constraints are equivalent to $\mathbf{W}\mathbf{C}_{\mathbf{x}(t)}\mathbf{W}^T = \mathbf{I}$, where $\mathbf{C}_{\mathbf{x}(t)} = \mathrm{E}_t \left\{\mathbf{x}(t)\mathbf{x}(t)^T\right\}$, and that they imply $\mathrm{E}_t \left\{\|\mathbf{y}(t)\|^2\right\} = K$. We also assume that the magnitudes of filter outputs and the driving noise are uncorrelated, that is, that $\mathrm{E}_t \left\{\mathbf{v}(t) \, |\mathbf{y}(t - \Delta t)|^T\right\} = \mathbf{0}$.[2] To make the generative model complete a mechanism for generating the signs of cell responses $\mathbf{y}(t)$ must be included. We specify that the *signs are generated randomly* with equal probability for plus or minus after the strengths of the responses have been generated.

It can be shown (see Appendix A) that least mean squares estimation of the model yields objective function

$$f_{|\mathbf{y}|}(\mathbf{W}) = \mathrm{E}_t \left\{ \left(|\mathbf{y}(t)| - \frac{1}{2}\mathbf{M}_{|\mathbf{y}|} \, |\mathbf{y}(t - \Delta t)| \right)^T \mathbf{M}_{|\mathbf{y}|} \, |\mathbf{y}(t - \Delta t)| \right\}, \tag{3}$$

[2] Note that $\mathrm{E}_t \left\{\mathbf{v}(t) \, |\mathbf{y}(t - \Delta t)|^T\right\} = \mathrm{cov}\left\{\mathbf{v}(t), |\mathbf{y}(t - \Delta t)|\right\} + \mathrm{E}_t\left\{\mathbf{v}(t)\right\} \mathrm{E}_t\left\{|\mathbf{y}(t - \Delta t)|\right\}^T$. If $\mathrm{cov}\left\{\mathbf{v}(t), |\mathbf{y}(t - \Delta t)|\right\} = \mathbf{0}$, which is a reasonable assumption, uncorrelatedness holds only if $\mathrm{E}_t\left\{\mathbf{v}(t)\right\} = \mathbf{0}$, because $\mathrm{E}_t\left\{|\mathbf{y}(t - \Delta t)|\right\}$ is strictly positive. By (2) $\mathrm{E}_t\left\{\mathbf{v}(t)\right\} = \left(\mathbf{I} - \mathbf{M}_{|\mathbf{y}|}\right) \mathrm{E}_t\left\{|\mathbf{y}(t)|\right\}$. It is easy to show that this expression is non-zero if $K = 1$. However, analysis of estimated models shows that it is very close to zero with larger values of K, so our assumption of uncorrelatedness is approximately correct.

where $\mathbf{M}_{|\mathbf{y}|} = \mathrm{E}_t\left\{|\mathbf{y}(t)|\,|\mathbf{y}(t-\Delta t)|^T\right\}\mathbf{C}^{-1}_{|\mathbf{y}(t)|}$. A problem with optimizing $f_{|\mathbf{y}|}(\cdot)$ is that the absolute value function is not differentiable at zero. This causes severe algorithmic problems with gradient-based algorithms, because these algorithms assume that the gradient changes smoothly. Therefore we replace the absolute value function with a smoothed version, $\mathbf{g}(\mathbf{y}(t)) = \mathbf{ln\,cosh}(\mathbf{y}(t))$, which maps each component $y_k(t)$ of $\mathbf{y}(t)$ to $g(y_k(t)) = \ln\cosh y_k(t)$. The objective function becomes

$$f_{\mathbf{g}(\mathbf{y})}(\mathbf{W}) = \mathrm{E}_t\left\{\left(\mathbf{g}(\mathbf{y}(t)) - \frac{1}{2}\mathbf{M}_{\mathbf{g}(\mathbf{y})}\mathbf{g}(\mathbf{y}(t-\Delta t))\right)^T \mathbf{M}_{\mathbf{g}(\mathbf{y})}\mathbf{g}(\mathbf{y}(t-\Delta t))\right\},$$

with $\mathbf{M}_{\mathbf{g}(\mathbf{y})} = \mathrm{E}_t\left\{\mathbf{g}(\mathbf{y}(t))\,\mathbf{g}(\mathbf{y}(t-\Delta t))^T\right\}\mathbf{C}^{-1}_{\mathbf{g}(\mathbf{y}(t))}$. The optimization of this objective under decorrelation and unit variance constraints can be done with a gradient projection method employing symmetric orthogonalization [2].

3 Experiment with natural video data

The natural image sequences used in our experiment were a subset of those used in [8]. Some video clips were discarded to reduce the effect of human-made objects and artifacts [2]. The preprocessed data set consisted of 200,000 pairs of consecutive 11 × 11 image windows at the same spatial position, but 40 ms apart from each other. Preprocessing consisted of temporal decorrelation, subtraction of local mean and normalization (see [2] for a description of the effect of temporal decorrelation). For purposes of computational efficiency the spatial dimensionality of the data was reduced to 80 with principal component analysis [4] – this still retains 97% of signal energy.

Optimization of the objective function is computationally very intensive, so in these preliminary results we have computed only a small number of filters.

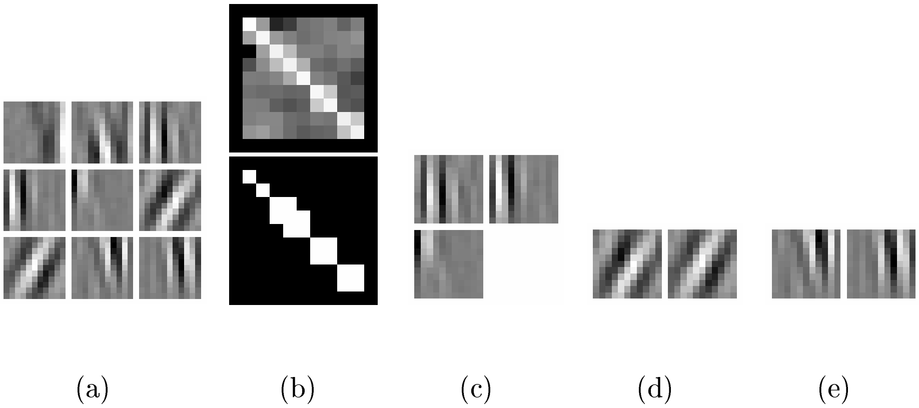

Fig. 2. The results of estimating the model from natural video when the number of filters $K = 9$. (a) A set of receptive fields (linear filters) found when the multivariate autoregressive model is estimated from natural video data. The ordering of the filters is irrelevant for the objective function – here the filters have been reordered to illustrate inter-cell dependencies (see text for details). (b) Two plots of matrix $\mathbf{M}_{\mathbf{g}(\mathbf{y})}$ for the receptive field set. Top: a plot of the logarithms of the elements of $\mathbf{M}_{\mathbf{g}(\mathbf{y})}$. Bottom: a thresholded, binary plot of those elements of $\mathbf{M}_{\mathbf{g}(\mathbf{y})}$ larger than 0.1. (c)–(e) Three multi-filter activity groups which correspond to the larger areas in the thresholded plot of $\mathbf{M}_{\mathbf{g}(\mathbf{y})}$.

Figure 2(a) shows a set of filters obtained when $K = 9$. The filters resemble simple cell receptive fields in that they are Gabor functions, or line/edge detectors which have distinctive on/off regions.

The objective function (2) is insensitive to a different ordering of components of $\mathbf{y}(t)$ / rows of $\mathbf{W}$, if accompanied by a corresponding rearrangement of the elements of $\mathbf{M}_{\mathbf{g}(\mathbf{y})}$. The filters in Fig. 2(a) have been ordered according to strengths of their interactions as follows. The first filter is the one that corresponds to the largest diagonal element of $\mathbf{M}_{\mathbf{g}(\mathbf{y})}$. Once the kth filter has been selected to be the filter with original index j, the index of the $(k+1)$th filter is chosen to be $\arg\max_i \left(\mathbf{M}_{\mathbf{g}(\mathbf{y})}(j,i) + \mathbf{M}_{\mathbf{g}(\mathbf{y})}(i,j)\right)$. After the filters have been ordered, the elements of $\mathbf{M}_{\mathbf{g}(\mathbf{y})}$ are rearranged accordingly.

A plot of the logarithms of elements of $\mathbf{M}_{\mathbf{g}(\mathbf{y})}$ for the filter set of Fig. 2(a) is shown at the top of Fig. 2(b). The diagonal values are large, indicating that for all the filters, activities at time $t - \Delta t$ and time t are highly correlated. This is in concordance with our previous results [2]. Looking at the nondiagonal elements of $\mathbf{M}_{\mathbf{g}(\mathbf{y})}$ we can observe some grouping emerging in the set of filters. This can be seen clearly when looking only at those values of $\mathbf{M}_{\mathbf{g}(\mathbf{y})}$ which are larger than 0.1 (Fig. 2(b), bottom). The set of filters can be divided into five groups, which we shall call *activity groups*: the first and the second filter form their own groups, then we have a group with three filters, and finally two groups with two filters in each. The common features of receptive fields in each of these groups is that they have the same orientation and frequency, and are close to each other in terms of spatial position. The same common features are typical for simple cell receptive fields that act as input to a single complex cell [7] (although some complex cells receive their input directly from the lateral geniculate nucleus). Thus these preliminary results suggest that the autoregressive model pools simple cell responses in groups similar to those of complex cell input groups. It should be emphasized that the pooling effect, including the actual connection weights, emerges in a completely unsupervised manner. This differentiates our work from other research done in this field [3,5].

4 Discussion

It is intuitively understandable that the model gives results like those in Fig. 2. Translation is the most common short-time transformation of lines and edges in natural video [2]. Cells with receptive fields like those in Fig. 2(a) respond strongly at successive time points in case of small translations [2]. Cells with similar orientation at nearby spatial locations also respond strongly in the case of a small translation, implying large values for corresponding elements of $\mathbf{M}_{\mathbf{g}(\mathbf{y})}$.

To conclude, we have described a two-layer generative model of natural video, and shown preliminary results of estimating the model from natural data. The results suggest that the model yields unsupervised pooling of simple cell receptive fields with the same orientation and nearby spatial location. The same pooling property is a distinctive feature of complex cells.

A Least mean squares estimation of the model

First we derive an expression for for $\mathbf{M}_{|\mathbf{y}|}$. The covariance between $|\mathbf{y}(t)|$ and $|\mathbf{y}(t-\Delta t)|$ is given by $\mathrm{E}_t\left\{|\mathbf{y}(t)|\,|\mathbf{y}(t-\Delta t)|^T\right\} = \mathrm{E}_t\left\{\mathbf{M}_{|\mathbf{y}|}\,|\mathbf{y}(t-\Delta t)|\,|\mathbf{y}(t-\Delta t)|^T\right\} + \underbrace{\mathrm{E}_t\left\{\mathbf{v}(t)\,|\mathbf{y}(t-\Delta t)|^T\right\}}_{=\mathbf{0}} = \mathbf{M}_{|\mathbf{y}|}\mathbf{C}_{|\mathbf{y}(t)|}$. So $\mathbf{M}_{|\mathbf{y}|} = \mathrm{E}_t\left\{|\mathbf{y}(t)|\,|\mathbf{y}(t-\Delta t)|^T\right\}\mathbf{C}^{-1}_{|\mathbf{y}(t)|}$. Least mean square (LMS) estimation gives $\mathrm{E}_t\left\{\|\mathbf{v}(t)\|^2\right\} = \mathrm{E}_t\left\{\left\||\mathbf{y}(t)| - \mathbf{M}_{|\mathbf{y}|}\,|\mathbf{y}(t-\Delta t)|\right\|^2\right\} = K - 2\mathrm{E}_t\left\{\left(|\mathbf{y}(t)| - \frac{1}{2}\mathbf{M}_{|\mathbf{y}|}\,|\mathbf{y}(t-\Delta t)|\right)^T\mathbf{M}_{|\mathbf{y}|}\,|\mathbf{y}(t-\Delta t)|\right\}$. Therefore LMS estimation is equivalent to maximizing (3).

References

[1] P. Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200, 1991.

[2] J. Hurri and A. Hyvärinen. Simple-cell-like receptive fields maximize temporal

coherence in natural video. Submitted manuscript, electronic version available at http://www.cis.hut.fi/jarmo/publications/.

[3] A. Hyvärinen and P. O. Hoyer. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, 41(18):2413–2423, 2001.

[4] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley & Sons, 2001.

[5] C. Kayser, W. Einhäuser, O. Dümmer, P. König, and K. Körding. Extracting slow subspaces from natural videos leads to complex cells. In G. Dorffner, H. Bischof, and K. Hornik, editors, *Artificial Neural Networks – ICANN 2001*, pages 1075–1080. Springer, 2001.

[6] B. A. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[7] S. E. Palmer. *Vision Science — Photons to Phenomenology.* The MIT Press, 1999.

[8] J. H. van Hateren and D. L. Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London B*, 265(1412):2315–2320, 1998.

[9] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London B*, 265(1394):359–366, 1998.

[10] L. Wiskott and T. J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. To appear in *Neural Computation*.