# 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**

We present a two-layer dynamic generative model of the statistical structure of natural image sequences. The second layer of the model is a linear mapping from simple cell outputs to pixel values, as in most work on natural image statistics. The first layer models the dependencies of the activity levels (amplitudes or variances) of the simple cells, using a multivariate autoregressive model. The second layer shows emergence of basis vectors that are localized, oriented and have different scales, just like previous work. But our new model enables the first layer to learn connections between the simple cells that are similar to complex cell pooling: connections are strong among cells with similar location, frequency and orientation. In contrast to previous work in which one of the layers needed to be fixed in advance, the dynamic model enables us to estimate both of the layers simultaneously from natural data.

*Key words:* natural image sequences, primary visual cortex, generative models

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

## 1 Introduction

A central question in the study of sensory neural networks is how stimuli are represented or coded by neurons. One approach to studying the neural code is to examine how its properties are related to the statistics of natural stimuli. Within the past ten years, computational principles relating the properties of cells in the primary visual cortex to the statistics of natural stimuli have been proposed. The most influential of these theories have been sparse coding [1], independent component analysis [2,3], and temporal coherence [4]. The principle of temporal coherence is based on the idea that when processing temporal input, the representation changes as little as possible over time. In a recent paper [5] we have shown that a nonlinear form of temporal coherence is related to the structure of simple cell receptive fields. According to the results presented in [5], simple cell receptive fields are optimally temporally coherent in the sense that the *activity levels* of simple cells are stable over short time intervals. By activity level we mean the amplitude or energy of the output of a linear filter that models a simple cell.

In the measure of temporal activity coherence introduced in [5] there was no possibility of an interaction between the activity levels of different cells. In this paper, we introduce a model which includes inter-cell activity dependencies. This is accomplished by a two-layer generative model in which the activity levels are generated in an autoregressive manner. We will show that estimation of the model from natural image sequence data yields simple-cell-like receptive fields, and a completely unsupervised complex-cell-like pooling between the outputs of simple cells.

(Aapo Hyvärinen).

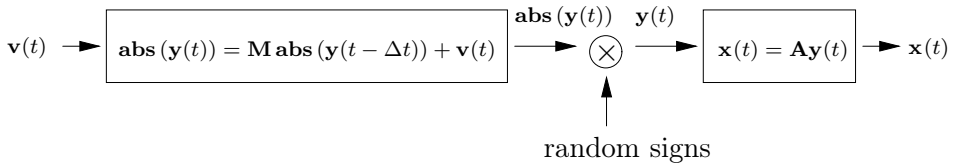


Fig. 1. The two layers of the generative model with spatiotemporal activity dependencies. Let $\mathbf{abs}\,(\mathbf{y}(t)) = [|y_1(t)| \cdots |y_K(t)|]^T$ denote the activity levels (amplitudes) of simple cell responses. In the first layer, the driving noise signal $\mathbf{v}(t)$ generates the activities of simple cells via a multivariate autoregressive model. Matrix $\mathbf{M}$ captures the spatiotemporal activity dependencies in the model. The signs of the responses are generated randomly between the first and second layer to yield signed responses $\mathbf{y}(t)$. In the second layer, natural image sequence $\mathbf{x}(t)$ is generated linearly from simple cell responses. In addition to the relations shown here, the generation of $\mathbf{v}(t)$ is affected by $\mathbf{M}\,\mathbf{abs}\,(\mathbf{y}(t-\Delta t))$ to ensure non-negativity of $\mathbf{abs}\,(\mathbf{y}(t))$. See text for details.

## 2 Definition of the model

The generative model of natural image sequences introduced in this paper has two layers (see Fig. 1). The first layer is a multivariate autoregressive model of the activity levels (amplitudes) of simple cell responses at time $t$ and time $t - \Delta t$. The signs of cell responses are generated by a latent signal between the first and second layer. The second layer is linear, and maps cell responses to image features.

We start the formal 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 patch taken from natural image sequences at time $t$. (Vectorization of image patches can be done by scanning images column-wise into vectors.) Let the vector $\mathbf{y}(t) = [y_1(t) \cdots y_K(t)]^T$ represent the outputs of $K$ simple cells.

The linear generative model for $\mathbf{x}(t)$ is similar to the one in [1]:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{y}(t). \tag{1}$$

Here $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_K]$ denotes a matrix which relates the image patch $\mathbf{x}(t)$ to the outputs of 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 obtain first is the mapping from $\mathbf{x}(t)$ to $\mathbf{y}(t)$, denoted by

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t). \tag{2}$$

Conceptually, the set of filters (vectors) $\mathbf{w}_1, ..., \mathbf{w}_K$ corresponds here 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 (2) 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 $\mathbf{A} = \mathbf{W}^T(\mathbf{W}\mathbf{W}^T)^{-1}$.

In contrast to basic sparse coding [1] or independent component analysis [3] we do *not* assume that the components of $\mathbf{y}(t)$ are independent. Instead, we assume that the activity levels (amplitudes) of the components of $\mathbf{y}(t)$ are correlated. We model these dependencies with a multivariate autoregressive model in the first layer of our model. Let $\mathbf{abs}\,(\mathbf{y}(t)) = [|y_1(t)| \cdots |y_K(t)|]^T$, and let $\mathbf{v}(t)$ denote a driving noise signal (the distribution of $\mathbf{v}(t)$ is constrained by the non-negativity of the process, and will be discussed in more detail below). Let $\mathbf{M}$ denote a $K \times K$ matrix, and let $\Delta t$ denote a time lag. Our model for the activities is a *constrained multidimensional first-order autoregressive process*, defined by

$$\mathbf{abs}\,(\mathbf{y}(t)) = \mathbf{M}\,\mathbf{abs}\,(\mathbf{y}(t - \Delta t)) + \mathbf{v}(t), \tag{3}$$

and unit energy constraints $\mathrm{E}_t\left\{y_k^2(t)\right\} = 1$, for $k = 1, ..., K$.

There are dependencies between the driving noise $\mathbf{v}(t)$ and $\mathbf{abs}\left(\mathbf{y}(t)\right)$, caused by the non-negativity of $\mathbf{abs}\left(\mathbf{y}(t)\right)$. Let $\mathbf{u}(t)$ denote a zero-mean random vector with components which are statistically independent of each other. We define $\mathbf{v}(t) = \mathbf{max}\left(-\mathbf{M}\,\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right), \mathbf{u}(t)\right)$, where, for vectors $\mathbf{a}$ and $\mathbf{b}$, $\mathbf{max}\left(\mathbf{a}, \mathbf{b}\right) = \left[\max(a_1, b_1) \ \cdots \ \max(a_n, b_n)\right]^T$. We assume that $\mathbf{u}(t)$ and $\mathbf{abs}\left(\mathbf{y}(t)\right)$ are uncorrelated. 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.

Note that the unit energy constraints and the uncorrelatedness of the outputs can be represented by a single matrix equation

$$\mathbf{W}\mathbf{C}_{\mathbf{x}(t)}\mathbf{W}^T = \mathbf{I}, \tag{4}$$

where $\mathbf{C}_{\mathbf{x}(t)} = \mathrm{E}_t\left\{\mathbf{x}(t)\mathbf{x}(t)^T\right\}$, and that they imply $\mathrm{E}_t\left\{\|\mathbf{y}(t)\|^2\right\} = K$. Therefore, because the sign generation mechanism also implies that each $y_k(t)$ has zero mean, the variances of the outputs will also be constant.

In equation (3), a large positive matrix element $\mathbf{M}(i, j)$, or $\mathbf{M}(j, i)$, indicates that there is a strong dependency between the activities of cells $i$ and $j$. Thinking in terms of grouping cells with large activity dependencies together, matrix $\mathbf{M}$ can be thought of as containing similarities (reciprocals of distances) between different cells. We will use this property in the experimental section to derive a spatial organization of the simple cells from $\mathbf{M}$.

## 3 Estimation of the model

To estimate the model defined above we need to estimate both $\mathbf{M}$ and $\mathbf{W}$ (the pseudoinverse of $\mathbf{A}$). In this section we first show how to estimate $\mathbf{M}$, given $\mathbf{W}$. Then we describe an objective function which can be used to estimate $\mathbf{W}$, given $\mathbf{M}$. Each iteration of the estimation algorithm consists of two steps. During the first step $\mathbf{M}$ is updated, and $\mathbf{W}$ is kept constant; during the second step these roles are reversed.

First, regarding the estimation of $\mathbf{M}$, consider a situation in which $\mathbf{W}$ has been fixed. It can be shown [6] that $\mathbf{M}$ can be estimated by using approximative method of moments, and that the estimate is given by

$$\begin{aligned}\mathbf{M} \approx \beta \mathrm{E}_t \Big\{ \left(\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right) \left(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T \Big\} \\ \times \mathrm{E}_t \Big\{ \left(\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right) \left(\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T \Big\}^{-1},\end{aligned} \tag{5}$$

where $\beta > 1$. Because this scaling does not affect the optima of (6), and because we are more interested in the relative magnitudes of elements of $\mathbf{M}$ than their absolute values, we can discard $\beta$ in the estimation process. See [6] for details.

The estimation of $\mathbf{W}$ is more complicated. A rigorous derivation of an objective function based on well-known estimation principles is very difficult because the statistics involved are non-Gaussian, and the processes have difficult interdependencies. Therefore, we derived an objective function heuristically, and verified its validity through simulations. The objective function is defined by

$$f(\mathbf{W}, \mathbf{M}) = \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbf{M}(i,j) \operatorname{cov}\left\{ \left|y_i(t)\right|, \left|y_j(t-\Delta t)\right| \right\}. \tag{6}$$

In the actual estimation algorithm, $\mathbf{W}$ is updated by employing a gradient projection approach to the optimization of (6) under constraint (4). The initial value of $\mathbf{W}$ is selected randomly. The fact that the algorithm described above is able to estimate models following the two-layer model has been verified through extensive simulations (details are given in [6]).

## 4 Experiments with natural image sequences

The data and preprocessing used in the experiments were very similar to those in [5], so we will describe them only shortly here, and refer the reader to [5] for details. The natural image sequences used in data collection consisted of 129 image sequences, which were a subset of natural image sequences used in [7]. The sampling rate in these sequences was 25 Hz. Initially 200,000 image sequences with a duration of 440 ms, and spatial size $16 \times 16$ pixels, were sampled from these sequences. The fairly long duration of these initial samples was necessary because of the temporal filtering used in preprocessing,

The preprocessing consisted of four steps: temporal decorrelation, subtraction of local mean, normalization, and dimensionality reduction. Temporal decorrelation, performed via temporal filtering, enhances temporal changes in the data, and differentiates our results from those obtained with static images [5]. It can also be motivated as a model of temporal processing at the lateral geniculate nucleus. After temporal decorrelation the spatial local mean (spatial DC component) was subtracted from each of the 400,000 image patches, and the patches were normalized to unit norm. This normalization can be considered as a form of contrast gain control. Finally, to reduce the effect of noise and aliasing artifacts, the dimensionality of the data was reduced to 160

using principal component analysis [3].

The estimation algorithm described in Section 3 was applied to the preprocessed natural image sequence data to obtain estimates for $\mathbf{M}$ and $\mathbf{A}$ (the pseudoinverse of $\mathbf{W}$). The extracted matrices $\mathbf{A}$ and $\mathbf{M}$ can be visualized simultaneously by using the interpretation of $\mathbf{M}$ as a similarity matrix (see Section 2), and by applying multidimensional scaling (MDS) to it (see [6] for details). The resulting spatial layout produced by the MDS procedure is shown in Fig. 2. First, regarding the properties of the resulting basis vectors, we can see that the basis vectors are localized, oriented, and have multiple scales. These are the most important defining criteria of simple cell receptive fields [8]. Second, regarding the similarity defined by the spatiotemporal dependency matrix $\mathbf{M}$, we can see that those basis vectors which are very close to each other seem to be mostly coding for similarly oriented features with the same frequencies at nearby spatial positions. This kind of grouping is characteristic of pooling of simple cell outputs at complex cell level, as well as of the topographic organization of the visual cortex [8]. In addition to the local topography described above, some global topography also emerges in the results: those basis vectors which code for horizontal features are on the left in Fig. 2, while those that code for vertical features are on the right.

## 5 Conclusion

We have described a two-layer dynamic generative model of image sequences, and an algorithm for estimating the model from sample data. Application of the algorithm to natural image sequences yields a set of linear filters, or basis vectors, which are similar to simple cell receptive fields, and connections

Fig. 2. Grouping similar to complex cell pooling of simple cell outputs emerges from spatiotemporal activity dependencies. Here we have plotted each of the basis vectors (columns of $\mathbf{A}$) at a 2D position obtained by applying multidimensional scaling to the similarity values defined by $\mathbf{M}$. As can be seen, nearby basis vectors seem to be mostly coding for similarly oriented features with similar frequencies at nearby spatial positions. In addition, some global topographic organization also emerges: those basis vectors which code for horizontal features are on the left in the figure, while those that code for vertical features are on the right. In this figure, some short distances have been extended in order to be able to show the basis vectors in a reasonable scale.

between the simple cells that are similar to the way in which simple cell outputs presumably are pooled at the complex cell level. The basis vectors are learned in one layer of the model, and the pooling property in the other. Both layers are learned simultaneously and in a completely unsupervised manner.

## References


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

[2] Anthony Bell and Terrence J. Sejnowski. The independent components of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[3] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

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

[5] Jarmo Hurri and Aapo Hyvärinen. Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation*, 2002. In press.

[6] Jarmo Hurri and Aapo Hyvärinen. A two-layer dynamic generative model of natural image sequences. Submitted, 2002.

[7] J. Hans van Hateren and Dan 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.

[8] Stephen E. Palmer. *Vision Science – Photons to Phenomenology*. The MIT Press, 1999.