Available online at www.sciencedirect.com
SCIENCE DIRECT®
Neurocomputing 58–60 (2004) 801–806
NEUROCOMPUTING
www.elsevier.com/locate/neucom

A unifying framework for natural image statistics: spatiotemporal activity bubbles

Aapo Hyvärinen[a,b,*], Jarmo Hurri[a], Jaakko Väyrynen[a]

[a] *Neural Networks Research Centre, Helsinki University of Technology, Finland*
[b] *Helsinki Institute for Information Technology, Basic Research Unit, Department of Computer Science, University of Helsinki, Finland*

Abstract

Recently, different models of the statistical structure of natural images (and sequences) have been proposed. Maximizing sparseness, or alternatively temporal coherence of linear filter outputs leads to the emergence of simple cell properties. Taking account of the basic dependencies of linear filter outputs enables modelling of complex cell and topographic properties as well. Here, we propose a unifying framework for all these statistical properties, based on the concept of spatiotemporal activity bubbles.

Keywords: Natural image statistics; Simple cells; Sparse coding; Independent component analysis; Temporal coherence

1. Introduction

Natural images are not white noise; they have some robust regularities. Previous research has built statistical models of natural images, and utilized them either for modelling the receptive fields of neurons in the visual cortex, or for developing new image processing methods. The following three properties seem to be the most important found so far: sparseness, temporal coherence, and topographic dependencies. This paper proposes a new framework for modelling the statistical structure of natural image sequences, combining these three properties. It leads to models where activation of the simple cells takes the form of "bubbles", which are regions of activity that are

* Corresponding author. Tel.: +358-9-191-44270; fax: +358-9-191-44441.
E-mail address: aapo.hyvarinen@helsinki.fi (A. Hyvärinen).
URL: http://www.cs.helsinki.fi/aapo.hyvarinen/

0925-2312/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2004.01.130

localized both in time and in space (space meaning the cortical surface). First, we will review some of the existing literature and known properties of natural images.

1.1. Sparseness

Outputs of linear filters that mimic simple cell receptive fields maximize sparseness [7]. Sparseness means that the random variable takes very small (absolute) values or very large values more often than a gaussian random variable would; to compensate, it takes values in between relatively more rarely. Thus the random variable is activated, i.e. significantly nonzero, only rarely. The probability density of the absolute value of a sparse random variable is often modelled as an exponential density, which has a higher peak at zero than a gaussian density.

Sparseness has nothing to do with the variance (scale) of the random variable. To measure the sparseness of a random variable s_i with zero mean, let us first normalize its scale so that the variance $E\{s_i^2\}$ equals some given constant. Then sparseness can be measured as the expectation $E\{G(s_i^2)\}$ of a suitable nonlinear function of the square. Typically, G is chosen to be convex, i.e. its second derivative is positive, e.g. $G(s_i^2) = (s_i^2)^2$. Convexity implies that this expectation is large when s_i^2 typically takes values that are either very close to 0 or very large, i.e. when s_i is sparse.

1.2. Temporal coherence

An alternative to sparseness is given by temporal coherence [2,9,10]. When the input consists of natural image *sequences*, i.e. video data, the simple-cell receptive fields optimize this criterion as well. Temporal coherence as defined in [2] is a nonlinear form of correlation, defined, for example, as the temporal correlation of the squared outputs. This means that the general activity level (variance) changes smoothly in time, although the actual cell outputs cannot be predicted.

It must be noted that ordinary *linear* correlation is *not* able to produce well-defined filters. Receptive fields maximizing linear correlation are more similar to Fourier components and lack the localization properties of simple-cell RFs [2].

1.3. Topographic dependencies

Consider a number of representational components s_i, $i = 1, \ldots n$, such as outputs of simple cells. Now, we consider their statistical dependencies, assuming that the joint distribution of the s_i is dictated by the natural image input. Again, we must consider *nonlinear* correlations like in the case of temporal coherence, since linear correlations are typically constrained to zero. In image data, the principal dependency between two simple-cell outputs seems to be captured by the correlation of their energies s_i^2, that is, the general activity levels or variances [8,3,4].

The dependencies of simple-cell outputs can be used to define a topographic organization. Let us assume that the s_i are arranged on a two-dimensional grid or lattice as is typical in topographic models. We have proposed a model [3,4] in which the energies are strongly positively correlated for neighboring cells. This means *simultaneous*

activation of neighboring cells; such simultaneous activation is implicit in much of the work in cortical topography.

1.4. Linear models of natural images

The statistical properties discussed above are usually utilized in the framework of a generative model. Denote by $I(x, y, t)$ the observed data whose components are pixel gray-scale values (point luminances) in an image patch at time point t. The models that we consider here express a monochrome image patch as a linear superposition of some features or basis vectors a_i:

$$I(x, y, t) = \sum_{i=1}^{n} a_i(x, y)s_i(t). \tag{1}$$

The $s_i(t)$ are stochastic coefficients, different from patch to patch. In a cortical interpretation, the s_i model the responses of (signed) simple cells, and the a_i are closely related to their classical receptive fields [7]. For simplicity, we consider only spatial receptive fields in this paper. Estimation of the model consists of determining the values of both s_i and a_i for all i, given a sufficient number of observed patches I_t.

In the most basic models, the s_i are assumed to be statistically independent, i.e. the value of s_j cannot be used to predict s_i for $i \neq j$. Then we can use either sparseness or temporal coherence to estimate the receptive fields [6]. If sparseness is used [7], the temporal structure of the data is ignored; indeed, the data does not need to have any temporal structure in the first place. The resulting model is called independent component analysis (ICA) [6], and it can be considered a nongaussian version of factor analysis. Temporal coherence leads to quite similar receptive fields [2]. When using topography, the s_i are not assumed to be independent anymore; instead, they have topographic dependencies as defined above. This leads to the topographic ICA model [3,4] which combines the properties of sparse components and topographic dependencies in a single model.

2. Activity bubbles as a unifying framework

2.1. Temporal bubbles

As discussed above, both maximization of the sparseness of linear filter outputs and the maximization of their temporal coherence lead to receptive fields that have the principal properties of simple cells. How is it possible that two quite different criteria give quite similar receptive fields? What is the connection between the two criteria?

To answer these questions, we propose a model of the linear filter outputs that combines the two properties. The model explains why both criteria give similar estimation results from natural images, and can be expected to give an improved model of the statistical structure of linear filter outputs. Here we only explain the basic idea of the model, see [5] for more details.

The new model is based on the concept of *sparse temporal activity bubbles*. (This will be extended to sparse *spatio*temporal activity bubbles below.) We assume that the simple cell output $s(t)$ is a product of an underlying latent signal $z(t)$ and a variance signal $v(t)$. Thus, we define

$$s(t) = v(t)z(t). \tag{2}$$

The underlying signal $z(t)$ does not need to have any special properties. In fact, we assume here, for simplicity, that $z(t)$ is gaussian white noise with unit variance. The interesting statistical properties of $s(t)$ are thus due to $v(t)$ alone.

The crucial assumptions are that $v(t)$ is *sparse* and has *temporal correlation*. To model such a signal, we assume that it is a low-pass filtered (smoothed) version of a very sparse signal possibly followed by a pointwise (scalar) function:

$$v(t) = f(\varphi(t) * u(t)) = f\left(\sum_{\tau} \varphi(\tau)u(t-\tau)\right), \tag{3}$$

where φ is a simple low-pass filter, such as the Gaussian kernel $\exp(-\tau^2/(2\sigma^2))$. The random process $u(t)$, which we call the bubble process, is obtained by sampling a very sparse nonnegative random variable independently at each time point, resulting in something similar to a point process with nonnegative values. The function f is a technical addition that has little influence on the basic principle, and in most cases we could take it just linear f.

The resulting signal $s(t)$ has both sparseness and temporal coherence, as shown in [5]. Thus, if one mixes linearly independent signals of this kind, the original signals can be separated by using either of these two properties [6]. In particular, if we consider the image sequences to be linear sums of spatial basis vectors as in Eq. (1), and assume that the signals $s_i(t)$ consist of temporal bubbles as defined above, it is natural that we obtain similar basis vectors with either criterion, since both of them are applicable on this kind of data [6]. We have also shown that temporal bubbles model characterizes the outputs of Gabor-like linear filters better that either sparseness or temporal coherence alone [5].

2.2. Spatiotemporal bubbles

Now, we will show how to *combine the three properties* discussed above: sparseness, topography, temporal coherence. Combination of sparseness and topography means that each input activates a limited number of spatially limited "blobs" on the topographic grid. If these regions are temporally coherent, they resemble activity bubbles as found in many earlier neural network models.

An activity bubble thus means the *activation of a spatially and temporally limited region*. This is illustrated in Fig. 1 for a one-dimensional map. Such an activity bubble corresponds to a basic element of visual input: A short (moving) luminance contour that is of a given orientation and frequency and inside a small spatiotemporal window. It is not quite the same as the spatial receptive field of a complex cell because the bubble has temporal characteristics.

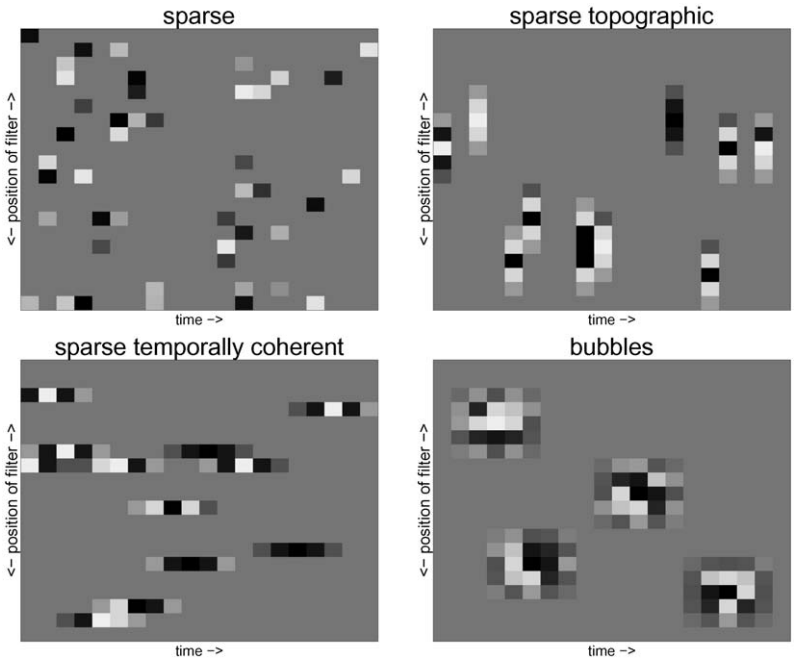

Fig. 1. The four types of representation. The plots show activities of simple cells as a function of time and the position of the cell on the topographic grid. For simplicity, the topography is here one-dimensional. In the basic sparse representation, the filters are independent. In the topographic representation, the activations of the filters are also spatially grouped. In the representation that has temporal coherence, they are temporally grouped. The bubble representation combines all these aspects, leading to spatio-temporal activity bubbles. Note that the two latter representation require that the data has a temporal structure, unlike basic sparse coding.

Based on earlier work [3,2], we can formulate generative models based on activity bubbles. We postulate a higher-order random process u that determines the variance at each point. This nonnegative, highly sparse random process obtains independent values at each point in time and space (space referring to the topographic grid). For simplicity, let us denote the location on the topography by a single index i. Then, the variances v of the observed variables are obtained by a spatiotemporal convolution followed by a pointwise nonlinearity:

$$v_i(t) = f\left(\sum_j h(i,j)[\varphi(t) * u_j(t)]\right), \tag{4}$$

where $h(i,j)$ is the neighborhood function that defines the spatial topography, and φ is a temporal smoothing kernel. The simple cell outputs are now obtained by multiplying simple gaussian white noise $z_i(t)$ by this variance signal:

$$s_i(t) = v_i(t) z_i(t). \tag{5}$$

Finally, the latent signals $s_i(t)$ are mixed linearly to give the image, as in Eq. (1). The three Eqs. (4), (5) and (1) define a statistical generative model for natural image sequences. The basis vectors estimated in the bubble model are quite similar to those obtained by topographic ICA [3]. The new contribution of the model is not in estimating a new basis, but in providing a better model of the statistical of the components s_i. For more information, see [5].

3. Discussion

Why would the visual system bother about such a sophisticated model of natural image statistics? First and foremost, bubble coding provides a suitable internal model of the structure of natural stimuli. If we consider visual processing in a Bayesian framework, it is paramount to obtain statistical models of the input that are as accurate as possible. Second, estimating the bubble process may be more interesting for higher areas than the activations of the single cells. In fact, temporal coherence has earlier been proposed as a principle for learning *invariant* features [1,10], and topographic ICA leads to emergence of features that are invariant to phase; in fact, they are very similar to complex cell responses. Thus, the bubbles are quite strongly invariant features. A further utility of temporal coherence may be that if the code is based on firing rates of neurons, temporal stability makes it easier to "read" the firing rates and reduces the Poisson noise that is inherent in such an operation.

To conclude, we have proposed a new framework for the low-level statistical structure of natural image sequences. This is based on the notion of spatio-temporal activity bubbles. This combines the properties of sparseness (the bubbles being sparse), topography (which corresponds to the spatial continuity of the bubbles), and temporal coherence (which corresponds to the temporal continuity of the bubbles).

References

[1] P. Földiák, Learning invariance from transformation sequences, Neural Comput. 3 (1991) 194–200.
[2] J. Hurri, A. Hyvärinen, Simple-cell-like receptive fields maximize temporal coherence in natural video, Neural Comput. 15 (3) (2003) 663–691.
[3] A. Hyvärinen, P.O. Hoyer, A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images, Vision Res. 41 (18) (2001) 2413–2423.
[4] A. Hyvärinen, P.O. Hoyer, M. Inki, Topographic independent component analysis, Neural Comput. 13 (7) (2001) 1527–1558.
[5] A. Hyvärinen, J. Hurri, J. Väyrynen, Bubbles: a unifying framework for low-level statistical properties of natural image sequences, J. Opt. Soc. Am. A 20 (7) (2003) 1237–1252.
[6] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley Interscience, New York, 2001.
[7] B.A. Olshausen, D.J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (1996) 607–609.
[8] O. Schwartz, E.P. Simoncelli, Natural signal statistics and sensory gain control, Nature Neurosci. 4 (8) (2001) 819–825.
[9] J. Stone, Learning visual parameters using spatiotemporal smoothness constraints, Neural Comput. 8 (7) (1996) 1463–1492.
[10] L. Wiskott, T.J. Sejnowski, Slow feature analysis: unsupervised learning of invariances, Neural Comput. 14 (4) (2002) 715–770.