Activity bubbles and natural image sequences

Aapo Hyvärinen
Neural Networks Research Centre, Helsinki Univ. of Technology, Finland

Abstract. Recently, different models of the statistical structure of natural images 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. In this paper, I propose a unifying framework for all these statistical properties, based on the concept of spatio-temporal activity bubbles.

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. This framework combines the above-mentioned three properties. It leads to models where the activation of the units (linear filters or simple cells) takes the form of "bubbles", which are regions of activity that are localized both in time and in space (space meaning the cortical surface or some other topographic grid).

1 Three properties of natural images

1.1 Sparseness

Recently, a lot of attention has been paid to one particular statistical property of natural images: sparseness, alternatively called supergaussianity or leptokurtosis [1, 8]. 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 non-zero, only rarely. This is illustrated in Fig. 1.

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. The exponential density is compared with the density of the absolute value of a gaussian variable in Fig. 2.

Sparseness is not dependent on 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 the 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.

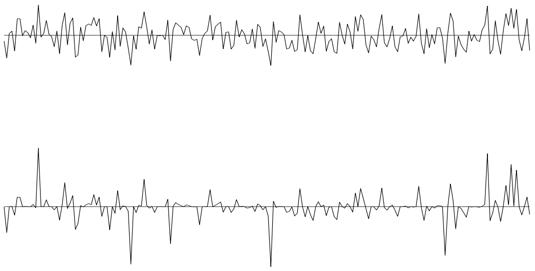

Figure 1: Illustration of sparseness. Random samples of a gaussian variable (top) and a sparse variable (bottom). The sparse variable is practically zero most of the time, occasionally taking very large values.

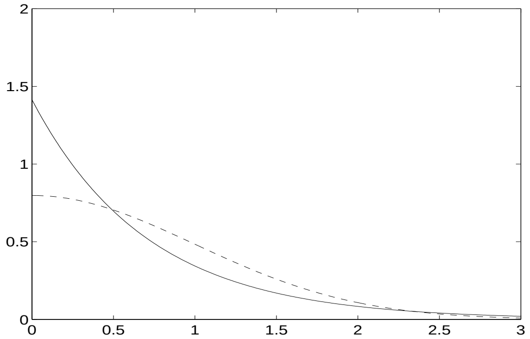

Figure 2: Illustration of a sparse probability density. Vertical axis: probability density. Horizontal axis, (absolute) value of random variable s_i. The sparse exponential density function is given by the solid curve. For comparison, the density of the absolute value of a gaussian random variable is given by the dash-dotted curve.

1.2 Temporal coherence

An alternative to sparseness is given by temporal coherence [2, 9]. 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 [4] is a nonlinear form of correlation. It can be defined, for example, as the temporal correlation of the squared outputs.

It must be noted that ordinary *linear* correlation is *not* able to produce well-defined filters. This is because the autocovariance (for a given time lag) of the sum $a_i s_i + a_j s_j$ of two independent signals is the sum of the autocovariances: $a_i^2 \text{autocov}(s_i) + a_j^2 \text{autocov}(s_j)$. Consider a case where the signals are uncorrelated, and have equal variances and autocovariances. Then, if the mixing coefficients fulfill $a_i^2 + a_j^2 = 1$, the mixture has the same variance and autocovariance as the original signals. Thus, maximization of autocorrelation does not properly define linear filters, and we have to use nonlinear autocorrelations [4].

1.3 Topographic dependencies

When using sparseness or temporal coherence, the outputs of linear filters are assumed independent. To go beyond this basic framework, we need to model their statistical dependencies. Consider a number of representational components $s_i, i = 1, ...n$, such as outputs of simple cells. Here, we analyze the pair-wise mutual informations $I(s_i, s_j)$, assuming that the joint distribution of the s_i is dictated by the natural image input. Note that again, we must consider *nonlinear* correlations such as those illustrated in Fig. 3; linear correlations do not provide enough information.

The dependencies of simple cell outputs have the important property of begin topographic. Let us assume that the s_i are arranged on a two-dimensional grid or lattice as is typical in topographic models. Topography is basically a property of the dependencies or pair-wise mutual informations. We say that the joint distribution of the s_i is topographic if the components s_i can be arranged on a topographic grid so that the neighbourhood function on that grid approximates the dependencies. The dependencies found in natural images can be used to model the cortical topography [5].

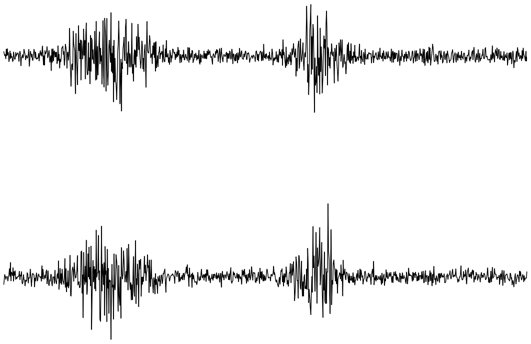

Figure 3: The dominant form of nonlinear dependency in linear filter outputs is energy correlation, illustrated here. The two signals in the figure are uncorrelated but they are not independent. In particular, their energies (squares) are correlated. The signals have thus strong simultaneous activation.

2 Linear models of natural images

The statistical properties discussed in the preceding section 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 expresses 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 (CRF's). 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. If sparseness is used [8], the temporal structure of the data is ignored; indeed, the data does not need ot have any temporal structure in the first place. The resulting model is called independent component analysis (ICA) [7], and it can be considered a nongaussian version of factor analysis. Temporal coherence leads to quite similar receptive fields [4]. In that case, the sparseness structure of the data is not utilized in the estimation.

When using topography, the s_i are not assumed to be independent anymore. Instead, they have topographic dependencies as defined in Section 1.3. This leads to the topographic ICA model [5, 6], which precisely combines the properties of sparse components and topographic dependencies in a single model. When topographic ICA is estimated from natural image data [5], the emerging topography is qualitatively very similar to the one observed in V1, and creates complex cell response properties as well.

3 Activity bubbles as a unifying framework

The idea in bubble coding is 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. 4 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 RF of a complex cell because the bubble has temporal characteristics.

Based on earlier work [5, 4], we can formulate generative models based on activity bubbles. Each simple cell is modelled as linear filter with adaptable weights, w_i. The output of the simple cell with index i, when input with an image patch I_t, is thus given by $s_i(t) = \langle w_i, I_t \rangle = \sum_{x,y} w_i(x,y) I(x,y,t)$.

As in [5], simple cell outputs are rectified by taking squares (energies), and these are fed to complex cells. To fix the pooling weights from simple cells to complex cells, we make here the assumption that complex cells only pool outputs of simple cells that are near-by on the topographic grid. Thus, the complex cell outputs are given by the locally pooled activations. The local activation at a position i on the grid for stimulus I_t, means a weighted sum of the energies of simple cells that are near-by in space. The pooling process into complex cell outputs can be expressed using a neighbourhood function $h(i,j)$ that is often used in topographic models. Basically, $h(i,j)$ tells how close to each other the filters i and j are on the topography, typically ranging from 0 to 1.

Here, however, we consider temporal pooling as well. We need to complement the neighbourhood function to obtain a spatio-temporal neighbourhood function as:

$$\tilde{h}(i,j,\tau) = h(i,j)\varphi(\tau) \tag{2}$$

where φ is a temporal smoothing kernel, for example, the gaussian kernel $\varphi(\tau) = \exp(-\tau^2/2)$. Thus, we define the output of a “bubble detector” at grid point i and time point t as

$$b_{it} = \sum_{\tau} \sum_{j=1}^{n} \tilde{h}(i,j,\tau) \langle w_j, I_{t-\tau} \rangle^2. \tag{3}$$

As an analogue to topographic ICA, we now define the likelihood of our model as:

$$\log L(I_1, \ldots, I_T;\ w_1, \ldots, w_n) = \sum_{t=1}^{T} \sum_{i=1}^{n} G(b_{it}). \tag{4}$$

The bubble pooling given by $\tilde{h}(i,j,\tau)$ is considered fixed, and only the first-layer weights w_j are estimated, so this likelihood is a function of the w_i only. The function G is typically convex to enforce sparseness of bubbles [5]. Note that alternatively, we could formulate a bubble model based on the autoregressive introduced model in [3].

4 Conclusion

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).

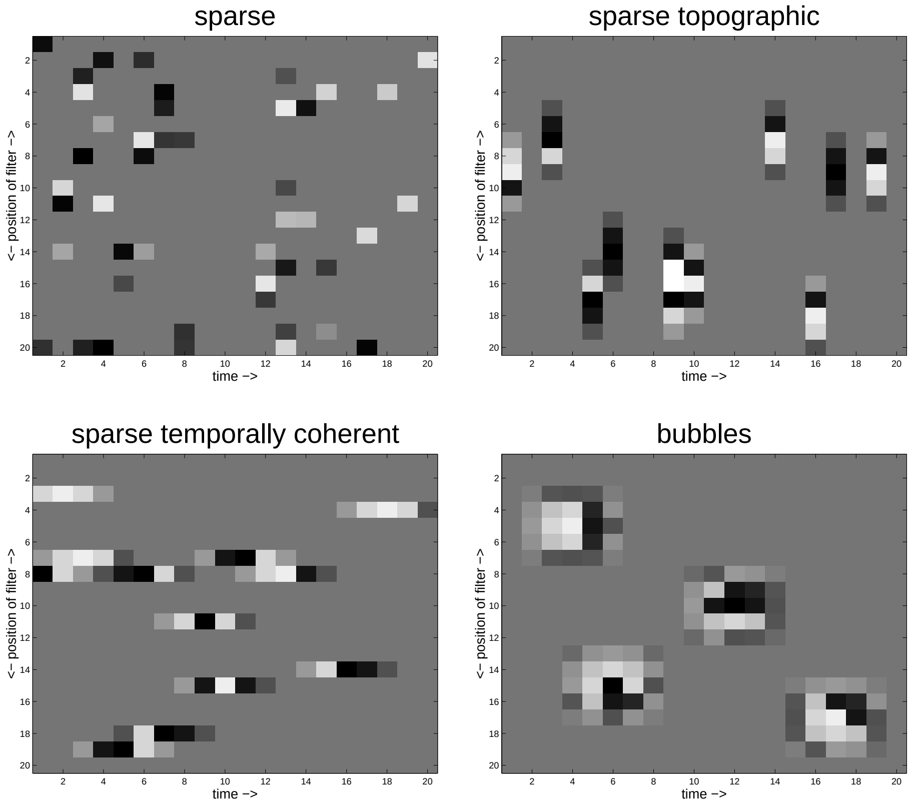

Figure 4: The four types of representation. The plots show activities of filters as a function of time and the position of the filter on the topographic grid. For simplicity, the topography is here one-dimensional. In the basic sparse representation, the filters are independent. In the topographic representation, 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 more or less require that the data has a temporal structure, unlike basic sparse coding.

References

[1] D.J. Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.

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

[3] J. Hurri and A. Hyvärinen. A novel temporal generative model of natural video as an internal model in early vision. In *Proc. First Int. Workshop on Generative Model Based Vision*, Copenhagen, Denmark. in press.

[4] J. Hurri and A. Hyvärinen. Receptive fields similar to simple cells maximize temporal coherence in natural video. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN2002)*, Madrid, Spain. in press.

[5] 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.

[6] A. Hyvärinen, P. O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558, 2001.

[7] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Interscience, 2001.

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

[9] L. Wiskott and T.J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.