Learning efficient internal representations from natural image collections

Antonio Turiel[1], José M. Delgado[2] and Néstor Parga

Departamento de Física Teórica, Universidad Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain.

Abstract

Learning in sensory systems takes place after a repeated exposure to the incoming signals and many ideas based in information theoretical principles have been proposed to explain the synaptic adaptation which improves the coding capabilities of sensory areas. In this paper we want to emphasize that a simple, natural learning rule can be derived from a careful treatment of image redundancies. The learning rule is used to split images into independent components which connect different resolution levels, in a non-linear way. The result shows the biological plausibility of this coding strategy not only in the visual pathway but also in other sensory modalities.

Keywords: Sensory coding, learning, feature detectors, natural images.

1 Introduction

Following early suggestions by Attneave (1) and by Barlow (2), many works have focussed on the use of information theoretical concepts to address the question of the efficiency of the neural code. Two of them are the minimization of the redundancy (2) and the maximization of the transmitted information (3). As shown in (4), the code which maximizes information transfer minimizes redundancy, that is, it extracts the independent components (5) of the signal. Several theoretical studies of the primary visual system have been done based on these ideas of information maximization and redundancy reduction (6; 7). The first non-trivial issue one has to deal with is the detection of the sources of redundancy in the stimuli. Second-order correlations have been extensively studied in this context (6; 7), but these are not the only source of redundancies (8). Statistical analysis of natural images points to the existence of important regularities in natural scenes related to their properties under scale tranformations (9; 10; 11).

[1]Present address: Departament de Física Fonamental, Universitat de Barcelona. Diagonal, 647. 08028 Barcelona, Spain

[2]Corresponding author: Tel.: +34-91 397 4884. Fax: +34-91 397 3936
e-mails: turiel@ffn.ub.es, Josemaria.Delgado@uam,es, Nestor.Parga@uam.es

One of these regularities is the persistency of image features across scales. This was studied in (10) and here we briefly review the main findings of that work (Sect.2). Eliminating this form of redundancy from the code leads to a prediction of feature detectors and efficient coding (12; 13). In Sect.3 we show that the filter resulting from our study has an intuitive interpretation in terms of learning. In the last section these results are discussed.

2 Feature persistency across scales

Here we deal with a particular kind of redundancy present in natural images. As an image is zoomed from coarse to finer scales details initially not seen suddenly become relevant. Once a feature (a spatial modulation of contrast) is detected at a given scale it will frequently be also present at finer scales (14). This persistency property of image features implies a redundancy that should be eliminated in order to obtain an efficient internal code.

For illustrative purposes let us consider the simple case where a single feature is present in a set of natural images, although it can appear at different scales and image positions. [3] The spatial dependence of this feature is described by a function $\Psi(\vec{x})$. Scenes can then be represented by placing the feature at a discrete set of scales and positions on the image. The coarsest scale is taken as one (the linear size of the image) and the jth scale as 2^{-j} ($j = 0, \ldots, \infty$). At a fixed scale j, where the feature has an extension of the order of the scale, there are up to 2^j distinct positions along each spatial dimension (one at the coarsest scale, four at the next finer scale, and so on) where the feature can be placed. We denote these scaled and shifted versions of the feature by $\Psi_{j\vec{k}}(\vec{x}) \equiv \Psi(2^{-j}\vec{x} - \vec{k})$ (the two components of $\vec{k}$ are the integers $0, \ldots, 2^j - 1$). Notice that this construction derives from a compromise between scale and translational invariances, as no representation can fulfil both at the same time (15).
The contrast $c(\vec{x})$ (luminosity minus its mean) of an image can then be expressed as

$$c(\vec{x}) \;=\; \sum_{j=0}^{\infty} \sum_{\vec{k} \in \mathcal{Z}^2_{2^j}} \alpha_{j\vec{k}}\, \Psi_{j\vec{k}}(\vec{x}) \tag{1}$$

The persistency property implies that the coefficients α at two different scales j and j' are statistically related. Knowledge of this relation would allow us to find a more efficient representation of images. This is because this regularity of the visual world could be stored in the wiring of the network instead of having to be observed at the arrival of every stimulus.

This problem was solved in (10), and its application to expansions such as eq. (1) was discussed in (12). The key point is that the feature coefficients at two different scales are related multiplicatively through another variable η which is *statistically independent* of the feature coefficient at the coarser scales (*i.e.*, with smaller j). For simplicity, we consider consecutive scales (the general case was discussed in (10)). Then, *persistency* is formulated as:

$$\alpha_{j\vec{k}} \;\doteq \eta_{j\vec{k}}\; \alpha_{j-1,\left[\frac{\vec{k}}{2}\right]} \tag{2}$$

where $\doteq$ means equality in the distributional sense. Eq. (2) implies that *statistically* the values of feature coefficients at a given scale are propagated to the next scale multiplicatively.

[3] For a generalization to several features see (13; 8).

However, we cannot say that large values at a coarse scale are followed by large values at finer scales and the corresponing location, because the equality is just distributional (*i.e.*, persistency is statistical). We will see how to implement persistency as a geometrical feature.

Figure 1 shows the numerical evidence that the multiplicative process, eq.(2) is indeed a property of natural images. The existence of a multiplicative process is *extremely robust*, in particular it holds for a large class of feature functions $\Psi(\vec{x})$. The left pannel in fig.(1) was computed using the Haar function, while for the right pannel a different function, optimal for image coding in a sense to be discussed in the next section, was used. In both cases the rightmost curve is the distribution of the feature coefficients at a coarse scale, obtained numerically on a set of natural images. The other two curves, overlapping each other to a large extent in both pannels, are the *experimentally* obtained distribution of feature coefficients at a finer scale and the *prediction* of the same distribution obtained with eq.(2). Previous works (10) already shown the validity of the multplicative process in a more general formulation.

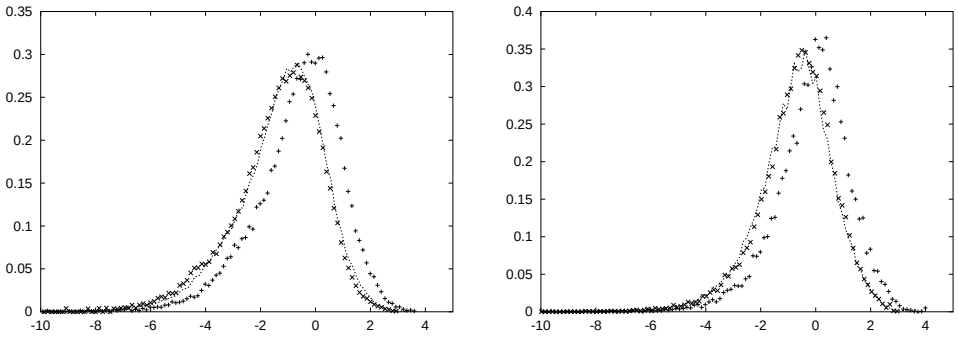

Figure 1: Experimental verification of the multiplicative process using two different feature functions. **Left**: Haar's function; **Right**: the optimal feature detector obtained from the learning rule, eq. (7). It was found using a set of ten images from Hans van Hateren's web database (see (16) for details). Starting from the histogram (+) of the wavelet coefficients $\alpha_{j,k}$ at scale $j = 5$, assuming translational invariance, and using a log-Poisson distribution (10) with parameters: $\Delta = 0.33$, $\beta = 0.66$ and $s = 1$ (11), we obtain a prediction for the distribution of $\alpha_{j,k}$ at scale $j = 6$ (dashed line). This has to be compared with the direct evaluation of the distribution of $\alpha_{j,k}$ at scale $j = 6$ (x).

3 The optimal filter: Learning rule

We have seen how the feature coefficients at different scales are related, according to the multiplicative process defined by eq. (2). As shown in Figure 1, we have verified the experimental validity of that model for a large class of filters and multiplicative processes in the class of log-Poisson distributions (10). However, having an efficient representation of natural images requires more than that: we would like to code an image by decomposing it in independent resolution levels. This means that the variables $\eta_{j\vec{k}}$ relating feature coefficients at two consecutive scales for a fixed image and position, in the way:

$$\alpha_{j\vec{k}} \;=\; \eta_{j\vec{k}}\, \alpha_{j-1[\frac{\vec{k}}{2}]} \tag{3}$$

should be independent of the feature coefficient at the coarser scale, $\alpha_{j-1\vec{k}}$, and have the same distribution for all resolution levels j and spatial locations $\vec{k}$. In other words, it is necessary that eq.(2) holds *point-by-point*, and not only distributionally.
This requirement is very restrictive and it cannot hold for any feature function Ψ. In fact, as it was shown in (12), this condition determines an optimal feature detector uniquely. Here we present a brief and simple derivation of this learning rule. From eq. (3) it follows that:

$$\alpha_{j\vec{k}} \;=\; \prod_{i=0}^{j-1} \eta_{j-i[\frac{\vec{k}}{2^i}]}\, \alpha_{0\vec{0}} \tag{4}$$

Now, taking eq. (4) into account and averaging eq.(1) over a large learning set of natural images (averages are indicated by angular brackets) we have

$$\langle c(\vec{x})\rangle \;=\; \langle \alpha_{0\vec{0}}\rangle \left(\Psi(\vec{x}) \;+\; \sum_{j=1}^{\infty} \sum_{\vec{k}\in\mathcal{Z}^2_{2^j}} \langle\eta\rangle^j \; \Psi_{j\vec{k}}(\vec{x}) \right) \tag{5}$$

A similar equation can be written for the average contrast at the next finer scale

$$\langle c(\vec{2x})\rangle \;=\; \frac{\langle \alpha_{0\vec{0}}\rangle}{\langle\eta\rangle} \left(\sum_{j=1}^{\infty} \sum_{\vec{k}\in\mathcal{Z}^2_{2^j-1}} \langle\eta\rangle^j \; \Psi_{j\vec{k}}(\vec{x}) \right) \tag{6}$$

The sum in eq. (6) is very similar to the second summand in eq. (5), the only difference is that the first runs over one fourth of the indices $\vec{k}$ of the second. So, translating the sum in eq. (6) to the four base points needed to fully expand the second summand in eq. (5), we can extract the filter Ψ by a simple substraction, that is,

$$\Psi(\vec{x}) \;=\; \frac{1}{\langle \alpha_{0\vec{0}}\rangle} \left(\langle c(\vec{x})\rangle - \frac{1}{2} \sum_{\vec{k}\in\mathcal{Z}^2_2} \langle c(2\vec{x}-\vec{k})\rangle \right) \tag{7}$$

where we used that $\langle|\eta|\rangle = \frac{1}{2}$ because image statistics is invariant under translations (11). Eq. (7) tells us that the feature function $\Psi(\vec{x})$ is obtained by averaging the images in the dataset and correcting this contribution from double-coding in detectors at the closest finer scale. The contribution of one scene is represented in Figure 2, the four terms of the correction are given by the same image rescaled by a factor one-half and centered at the four indicated positions. Eq. (7) defines a learning rule based on a correction to the simple Hebb's rule of the first term, in which neurons associated to finer scales act inhibitorily over neurons in the coarser scale, suppressing their reponse to features which were already detected at smaller sizes and forcing them to concentrate in the truly novel arriving features. So, the elliminaiton of redundancy between scales leads to an optimal filter deduced by means of a Hebbian-like learning rule.

Figure 2: Contribution to the optimal feature detector Ψ of a single image. **Left**: the selected image. **Middle**: the inhibitory contribution from cells at the next finer scale. **Right**: the net contribution, $\Psi_1(\vec{x})$, of the image.

4 Discussion

Let us now discuss some of the most relevant properties of this feature detector. The feature $\Psi(\vec{x})$ can be learned online by just accumulating the contributions from the incoming visual stimuli, as the expression eq.(7) is linear in $c(\vec{x})$. The optimal feature detector obtained after the observation of a large set of natural scenes is shown in Figure 3; it is an horizontal edge, which implies that images consist of, and are represented by, edges.

Eq.(7) assumes that the $\eta_{j\vec{k}}$'s are independent of the α's at scales coarser than j. This hypothesis can be verified, first evaluating $\alpha_{j\vec{k}}$ and $\alpha_{j-1,[\vec{k}/2]}$ (with the feature function in Figure 3), then computing the $\eta_{j\vec{k}}$'s as ratios of those coefficients, eq. (3). Using information theoretical measures one finds that the independence assumption holds very well (13).

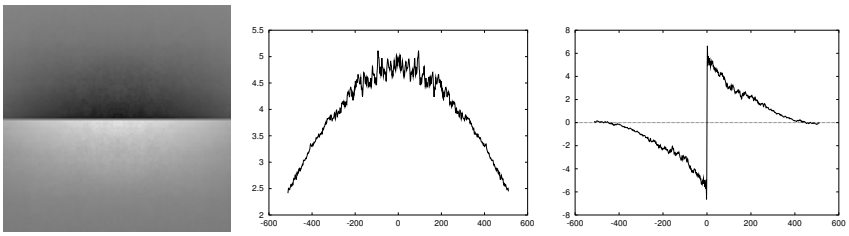

Figure 3: **Left**: Gray level representation of the filter learnt from 4000 images (white: positive values, black: negative values); **Middle**: Horizontal cut; **Right**: vertical cut

As a mathematical remark, it should be noticed that eq.(1) is a *wavelet expansion*, and the feature function $\Psi(\vec{x})$ is a special type of function, a *mother wavelet*. A wavelet is capable to expand any function $c(\vec{x})$ by a linear superposition of shifted and rescaled versions of itself. Our presentation shows that simple principles, derived from *observational properties* of the statistics of signals, can give rise to coding schemes which are very efficient in both coding and processing. At the same time, the algorithms proposed seem to be strongly connected with the way in which biological systems act. We think that this methodology is not exclusive of the visual system, and could be used to understand other sensory modalities.

Acknowledgements

A. Turiel is supported by a research contract from *Generalitat de Catalunya* (RED 2002). This work was funded by a Spanish grant BMF2000-0011.

References

[1] F. Attneave, Informational aspects of visual perception, Psychological Review 61 (1954) 183–193.

[2] H. B. Barlow, Possible principles underlying the transformation of sensory messages, in: W. Rosenblith (Ed.), Sensory Communication, M.I.T. Press, Cambridge MA, 1961, p. 217.

[3] R. Linsker, Self-organization in a perceptual network, Computer 21 (1988) 105.

[4] J.-P. Nadal, N. Parga, Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer, Network: Computation in Neural Systems 5 (1994) 565–581.

[5] P. Comon, Independent component analysis, a new concept?, Signal Processing 24 (1994) 287–314.

[6] J. J. Atick, Could information theory provide an ecological theory of sensory processing?, Network: Comput. Neural Syst. 3 (1992) 213–251.

[7] J. H. van Hateren, Theoretical predictions of spatiotemporal receptive fields of fly lmcs, and experimental validation, J. Comp. Physiology A 171 (1992) 157–170.

[8] A. Turiel, N. Parga, Role of statistical symmetries in sensory coding: an optimal scale invariant code for vision, accepted in Journal of Physiology (Paris) (2003).

[9] D. L. Ruderman, The statistics of natural images, Network 5 (1994) 517–548.

[10] A. Turiel, G. Mato, N. Parga, J. P. Nadal, The self-similarity properties of natural images resemble those of turbulent flows, Physical Review Letters 80 (1998) 1098–1101.

[11] A. Turiel, N. Parga, The multi-fractal structure of contrast changes in natural images: from sharp edges to textures, Neural Computation 12 (2000) 763–793.

[12] A. Turiel, N. Parga, Multifractal wavelet filter of natural images, Physical Review Letters 85 (2000) 3325–3328.

[13] A. Turiel, J.-P. Nadal, N. Parga, Orientational minimal redundancy wavelets: from edge detection to perception, Vision Research 43 (9) (2003) 1061–1079.

[14] S. Mallat, S. Zhong, Characterization of signals from multiscale edges, IEEE Trans. on Pattern Analysis and Machine Intelligence 14 (1992) 710–732.

[15] Z. Li, J. J. Atick, Towards a theory of the striate cortex, Neural Computation 6 (1994) 127–146.

[16] J. H. van Hateren, A. van der Schaaf, Independent component filters of natural images compared with simple cells in primary visual cortex, Proc. R. Soc. Lond. B265 (1998) 359–366.