Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Physiology - Paris xxx (2003) xxx–xxx
Journal of Physiology Paris
www.elsevier.com/locate/jphysparis

Role of statistical symmetries in sensory coding: an optimal scale invariant code for vision

Antonio Turiel [a], Néstor Parga [b,*]

[a] *Air Project—INRIA, Domaine de Voluceau, BP 105, 78153 Le Chesnay Cedex, France*
[b] *Departamento de Física Teórica, Universidad Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain*

Abstract

The visual system is the most studied sensory pathway, which is partly because visual stimuli have rather intuitive properties. There are reasons to think that the underlying principle ruling coding, however, is the same for vision and any other type of sensory signal, namely the code has to satisfy some notion of optimality—understood as minimum redundancy or as maximum transmitted information. Given the huge variability of natural stimuli, it would seem that attaining an optimal code is almost impossible; however, regularities and symmetries in the stimuli can be used to simplify the task: symmetries allow predicting one part of a stimulus from another, that is, they imply a structured type of redundancy. Optimal coding can only be achieved once the intrinsic symmetries of natural scenes are understood and used to the best performance of the neural encoder. In this paper, we review the concepts of optimal coding and discuss the known redundancies and symmetries that visual scenes have. We discuss in depth the only approach which implements the three of them known so far: translational invariance, scale invariance and multiscaling. Not surprisingly, the resulting code possesses features observed in real visual systems in mammals.

Keywords: Sensory coding; Vision; Multiscaling; Edge detection; Wavelets; Learning

1. Introduction

1.1. Need for an efficient representation. Optimization criteria

The world is an extremely complex system for living organisms. Living in it requires to solve the problem of how to represent and process efficiently the stimuli that are continuously received through the sensory pathways. This is a formidable task; think for example in the complexity of the visual world. It is endowed with properties such as contrast, motion, color and depth that have to be detected, represented and processed in order to reach a description which is useful and meaningful for the organism. No doubt that our brain has found excellent strategies to do this, but how? The first question that arises is how to deal with the huge amount of redundancy present in natural signals. It has been proposed that the goal of the first stages of the sensory pathways (e.g. the retina and the very first layers behind) is to realize an efficient neural representation (or code) of the environment and that cells achieve this goal by detecting statistical regularities in the stimuli [3]. The knowledge of those regularities could then be used to build an efficient internal representations of the environment. For instance, if there are image features that tend to appear together, a cell responding quasi-optimally to them is rather likely to exist.

To go further with this analysis it is necessary to assume that the properties of real nervous systems result from the optimization of some cost function which characterizes the quality of the code. Some time ago [2,3] it was suggested that *information theory* [10] could provide appropriate tools. For example Barlow [3] insists on the need of building a neural representation that could be easily used in subsequent processing. This leads to the idea of *factorial code*: each output unit should be statistically independent from any other unit. Hence the network decorrelates independent features that are mixed in the input signal. This means that one should minimize the *redundancy* in the neural code, a fact that

[*] Corresponding author.
E-mail addresses: antonio.turiel@inria.fr (A. Turiel), parga@delta.ft.uam.es (N. Parga).

0928-4257/$ - see front matter © 2004 Published by Elsevier Ltd.
doi:10.1016/j.jphysparis.2004.01.007

can be quantified in terms of an information theoretic criterion. Another possible requirement is that the system should simply maximize the amount of information that the output conveys about the input signal. This suggests in a particular way for modelling how the transfer function of a given sensory neuron is adapted to the particular environment in which the animal lives [21]. This idea that the information has to be preserved has been also developed by Linsker [23] under the name of infomax principle in a model of the first layers of the visual system.

The maximization of information transfer (the *infomax* principle of Linsker), and its the redundancy reduction of Barlow are in fact related. The predictions of these two strategies, redundancy reduction and maximization of mutual information, appeared to be very similar and the question arises of under which conditions they lead to the same predictions. The equivalence of these two criteria seems plausible and a first indication in this direction was noticed in an analysis of a population of McCulloch and Pitts neurons acting as a neuron encoder [30]. A direct evaluation of the mutual information and the redundancy in such a network showed that *infomax* and *redundancy reduction* principles are, for this system, equivalent. The proof of the equivalence was finally done in [31] where it was shown that, in the low synaptic noise limit with nonlinear outputs, infomax implies redundancy reduction, when optimization is done over both the synaptic efficacies and the non-linear transfer functions. As a result, the optimal neural representation is a factorial code. An extension of this result to stochastic neurons was done in [33].

A steepest descent algorithm taking the mutual information as cost function to find the independent components of data (that is, to minimize their redundancy) was soon afterwards proposed by Bell and Sejnowski [5]. The equivalence of these criteria with yet another popular principle, maximum likelihood, was found in [32]. As it was shown in this work, the cost function provided by the mutual information is identical to the one derived several years before from a maximum likelihood approach [37].

These results can be summarized in the following way:

- The redundancy can be reduced by means of the infomax principle [31].
- If the goal is to maximize the information, both the weights and the transfer function should be optimized [31].
- The redundancy can also be reduced by means of maximum likelihood [37].
- Maximum likelihood can be used to maximize the mutual information, once the transfer function is identified with the cumulative of the prior [32].

1.2. Statistical invariances of natural scenes

Natural images are complex objects, quite random (the content of one scene is highly variant) but at the same time quite structured (scenes consist of objects more or less smoothly illuminated). A complete description of their statistics would allow a direct knowledge of the optimal code, but such a description is far from possible: even a tiny 16×16 grid of pixels, quantized to 8 bits for pixel (256 gray levels) means a probability space with 2^{768} events, that is, more than 1.5×10^{231} possibilities. Obtaining even a coarse estimation of the probability distribution is unrealistic, the number of needed examples exceeding any device recording capability. It is highly unlikely that the visual system makes use of such an exhaustive description of images. In the huge space of possible images, scenes from the real world just occupy a small fraction of the whole. This means that they possess regularities and it is very attractive to think that the visual system makes use of them to achieve a good representation of the visual world [3]. According to this view the visual system first learns the regularities and produces a code well adapted to the type of images that it uses to see.

What sort of regularities one should look for? One immediate answer to this question is to look for symmetries in natural images. Of course they should be statistical in character: may be one particular image does not verify the symmetry, but it will be observed on average over a large enough set of images. Translational invariance is the simplest statistical symmetry in images. It means that there is no center of the universe, no distinguished point in the visual field. Something which in a particular scene is at a given location will eventually be observed in any other location for another different scene (for instance, trees are not always to the right of the observer, but they can appear anywhere). Of course it could happen that the ensemble of images acquired by the observer is biased in some way (e.g. if the subject of the study suffers the effect of a scotoma [28] or another anomaly in vision, or an attentive gaze). We consider no bias, no attention is happening in this early stage, hence guaranteeing translational invariance. Image ensembles used in vision research are nevertheless biased: the most important bias is reflected in the non-isotropy of images, as the recording device is normally kept horizontal with respect to the ground. Such a bias does not affect translational invariance, as objects happen to appear also at different heights, due for instance to irregularities of the terrain.

A more involved but also essential statistical symmetry is scale invariance. It implies that any part in a particular image will eventually happen at a different relative size in another image (for instance, trees have no universal, fixed visual size, but their apparent size depends on the random distance to the observer: the clo-

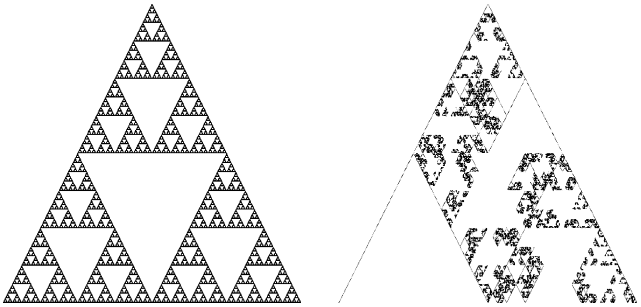

Fig. 1. Left: affine Sierpinski's gasket. The figure is generated by recursion, dividing an equilateral triangle in four half-height triangles and removing the central one. Right: probabilistic Sierpinski's gasket. It is generated similarly to the affine one, but at each step the triangle to be removed is decided at random.

ser, the larger). Scale invariance means that images are self-similar in a statistical sense. Self-similar objects are associated to the usual idea of fractal sets [15]: something which resembles to any small part of itself (as the Sierpinski's gasket shown in Fig. 1, left). In fact, an exact resemblance of the whole set to its parts is just a simple way of thinking fractals; a more realistic framework generalizes the concept of deterministic (affine) resemblance to that of statistical (probabilistic) resemblance (as that of the random Sierpinski's gasket shown in Fig. 1, right).

Natural images are scale invariant, which implies they are somehow fractals, in the sense of statistical self-resemblance. One of the salient features of fractal sets is power law behavior. Power law means that any statistical variable defined at a scale l (for instance, correlation of contrast for any pair de points a distance l apart) depends on the scale just as l^{γ}, with a certain exponent γ. This exponent is related to the fractal dimension of the set, in a way which depends on the precise definition of the considered variable. A well-known power law behavior of this type is that of the power spectrum of images [16] $S(\vec{f})$,

$$S(\vec{f}) \sim f^{-2+\delta} \tag{1}$$

where the exponent δ is usually small and depends on the particular ensemble of images considered. For a simple fractal system, one of this power law exponents suffices to define the scale invariance symmetry completely: it provides the dimension of the fractal (in the case of the power spectrum, the fractal would have dimension $2-\delta$). If scale invariance of natural images were described by just its power spectrum then images would be simple fractals. This assumption was done very frequently, specially in older studies of natural scenes in connection with the early stages of the visual system [1,16,52]. Also, models of natural images that attempt to explain the power law in Eq. (1) are usually simple fractals [40]. Evidence that scale invariance is also present in the non-gaussian statistics of natural scenes was observed in different studies [38,39,41,44–46], however the precise nature of scale invariance of natural scenes and its role in the development of visual systems has not been studied until very recently.

Natural images are not simple fractals, as they present different, not easily related power law exponents. This is because images are multifractals [44–46], so their scale invariant properties require a more complicated description. Multifractality is a beautiful structure, revealing important geometrical properties, but its discussion would exceed the limits of this paper; the reader is encouraged to read our paper on the subject in [46].

This geometrical property is related to a third symmetry of natural images, multiscaling. The existence of a new symmetry places stronger constraints in what a natural scene is, reducing their entropy. When the three symmetries are included in models of natural scenes [34,47] and the resulting model is used in combination with a plausible optimization principle [48,49], edge detectors are predicted in a rather natural way.

Symmetries of natural scenes have been considered in some theoretical modelling of simple cells [22], but this has not always been the case [6] and much of the insight gained from them is lost when they are not taken into account properly. In the following sections we will analyse its statistical meaning in conjunction with multiscaling; as we will see, this approach will lead us to extract relevant consequences for coding in the visual pathway.

1.3. Cells adapt to stimulus statistics

It is beyond the scope of this work to present a detailed account of cell adaptation to stimulus statistics. Here we just give a few examples of this question. If vision neurons have adapted to the regularities of natural scenes this should be evident in nature. There are indeed numerous examples of this adaptation: a first example is the orientation selective cells found by Hubel and Wiesel [20]; but after their discovery many other examples were found. One suggestive example out of many possible ones is that birds have horizontal edge detectors [29], which is probably due to the fact that the horizon is an important feature for them; in flies, adaptation of the transfer function to the statistics of images has been tested in [21], and the question has deserved further attention more recently [7].

Some forms of adaptation occurring in small time-scales could also be explained in terms of similar optimization principles as those used to explain adaptation through evolution [7,9,28,54].

Theoretical predictions for receptive fields have also been experimentally checked (see e.g. [52]). Experiments on cats suggest that some kind of optimization is taking place in LGN [12,43] to obtain a decorrelating code.

If scale invariance has been perceived as another regularity it should also be reflected in the properties of cells in the visual pathway. Although retinal and LGN cells do not seem to be specialized to detect features at a fixed scale, simple cells do [24].

The layout of the rest of this paper is as follows. In the next section we mention our main results on image statistics and the type of code suggested by them. We close Section 2 by stating the optimization criterion used to obtain the optimal filter. In Section 3 we describe the multiscaling property of natural scenes and show that a factorial code cannot be obtained by a linear filter. The optimal filter, an edge detector, is obtained as a consequence of the three symmetries of images in Section 4 and its important properties are described in Section 5. The self-consistency of the assumptions made to derive the optimal filter is analysed in the more numerical Section 6 and our conclusions are presented in the last section.

2. Towards an efficient code for natural visual stimuli

How can an efficient code for natural scenes be achieved? If scale invariance is such a prominent feature of the visual world special attention should be paid to the role that it plays in the construction of a non-redundant representation. There have been many suggestions on this regard.

Most of the effort in this direction concerns the scale invariant properties of the power spectrum of images. The most frequently invoked optimal criterion in this case has been decorrelation [1,52]. But this does not take into account other manifestations of scale invariance. In fact there is evidence that simple cells sample a visual stimulus by a joint representation of space and scale, instead of space and spatial frequency. This is reflected for instance in that the bandwidth is roughly constant in logarithmic units of the frequency; see for instance the discussions in [18,25]. The mathematical equivalent of this is to represent the image as a set of filters that explore the visual input at different positions and scales. The filters are obtained as scaled copies of a primitive one, conveniently placed at different locations, forming a *wavelet* family; the image can be reconstructed recombining the filters (wavelets) and activities (wavelet projections, also called wavelet coefficients) [13]. However a single wavelet family does not usually suffice to provide a complete description; hence several families, each associated to a mother wavelet, are generally used.

Wavelet expansions have been extensively used both in modeling the orientation selective cells and in the description of natural images, see e.g. [17,22]. However these and other studies do not address the problem of understanding the way scale invariance appears in natural scenes. They do not address either the question of how the wavelet exponents are related, a fact that can be explained if multiscaling is included in the description of natural scenes. Only once this question is ellucidated should the consequences that scale invariance—together with translation invariance and multiscaling—has on the visual system be investigated. This program has been done in a series of papers [34,44–46,48,51] from where the following conclusions have been reached:

- Scale invariance is not exhausted by the fact that the power spectrum of natural scenes is an algebraic function of the spatial frequency. Once images are decorrelated they are still very informative about the scene: edges are still present and make the image fully recognizable [4,16]. This implies that the higher order statistics cannot be neglected and, in particular, that non-gaussian scaling properties are also important.
- It turns out that images are extremely irregular: no matter the scale of observation the contrast has an intermittent behavior [46].
- Wavelet representations are not intrinsically efficient. It is well-known that wavelet coefficients have a persistence effect [11,26], a sharp transition in contrast of given spatial extension will give rise to large wavelet coefficients at all the scales finer that its size. Hence, if the activities were proportional to the wavelet coefficients, many cells would fire under the presentation of such a stimulus. A non-redundant representation could be obtained by cells that respond to *what is new* at the range of scales where it is most sensitive. In that case, it is not the presence of a change in contrast what is represented but the emergence of a new, finer structure in the image at the scale represented by the neuron.
- This non-redundant code is not given by the wavelet coefficients themselves but by a non-linear function of ratios between coefficients at two different scales. Possible non-linearities that arise naturally in the problem are: a power of the ratio of activities and the log-transfer of the wavelet filter. The first case corresponds to the divisive normalization [8,19], while the second could describe the saturation of simple cell response [24].
- Since in natural scenes the emergence of new details as the scale becomes finer are rare, the predicted code is sparse. Sparseness has not to be imposed (as it was in [36]) in this approach but it arises as a consequence of the sparseness of edges in natural scenes.
- The second order statistics can be dealt with as it is done in more classical approaches (e.g. [1]). This is because all the previous properties are valid regardless of the precise behavior of the (scale invariant) power spectrum. They hold no matter where in the sensory pathway decorrelation takes place. It is then

indifferent in this approach if decorrelation occurs fully in the LGN, as claimed in [12], or in V1, as has been suggested more recently in [57].

An important ingredient in this analysis is the following *optimization criterion* for scale invariant problems [48,49]: *the redundancy between the representation at different scales should be minimized*. Indeed, if simple cells work as feature detectors at different scales, it is then reasonable to start the search for a factorial code by asking independence across scales. Once this is achieved the next step towards attaining an efficient code is to reduce the spatial correlations within a fixed scale.

In the following sections we describe how these conclusions arise from the analysis of the statistics of local contrast changes in natural scenes and how the optimization principle just stated leads to the prediction that optimal filters are edge detectors.

3. In natural scenes scale invariance appears as multiscaling

In scale invariant systems quantities defined on a scale l have to behave as a power law of the scale, as this is the only function that does not require the existence of a privileged scale. In the simplest manifestation of this property the exponents for each possible quantity are trivially related. Natural scenes, however, happen to be an example of a more complex class of scale invariant problems. For instance, let us consider the moments of a contrast dependent random variable defined at the scale l. As we have just said each of those moments should behave as a power of the scale. In a simple system, the exponents depend linearly on the order of the moments; on the contrary, for natural images the corresponding exponents have a complex, non-linear dependence. In these problems, understanding scale invariance means to be able to predict the non-trivial relations between scale exponents. Natural scenes are an example of these problems [44,45]. Fortunately they obey multicaling, a property that makes the computation of the scale exponents feasible.

To explain the phenomenon of multiscaling better the first step is to define appropriate wavelet projections. We will discuss more about wavelets in Section 4; for the moment let us just say that a wavelet Ψ is a special function with a number of vanishing moments. Let $c(\vec{x})$ represent the luminosity recorded by the optical device at the point $\vec{x}$. We define the wavelet projection of $c(\vec{x})$ on Ψ at the scale l and position $\vec{x}$ as

$$\alpha_\Psi(l,\vec{x}) \equiv \int \mathrm{d}\vec{y}\, c(\vec{y}) \frac{1}{l^2} \Psi\left(\frac{\vec{x}-\vec{y}}{l}\right) \tag{2}$$

Expressed in a more intuitive way, the wavelet projection performs a zoom on the details of the function around $\vec{x}$ only at the scale l. In this way, it should be able to extract one or another of the different scaling sets in the system. In a statistical approach the p-moments of $\alpha_\Psi(l,\vec{x})$ are considered

$$\langle \alpha_\Psi^p(l,\cdot)\rangle \sim l^{\tau_p} \tag{3}$$

where the average (angular brackets) in the previous expression is taken over an ensemble of images and also over all the points, using translational invariance. Eq. (3) is known as statistical self-similarity (SS). It has been verified over large sets of very different natural images [34,44–47]. For a single fractal system, the SS exponents follow a linear relation with the order p: $\tau_p = (D-1)p + (2-D)$, where D is the dimension of the fractal set. However, it has been experimentally shown that the curve τ_p vs p deviates considerably from a straight line [44,45]. The SS exponents for natural images correspond to a more complicated multiscaling hierarchy.

An immediate consequence of multiscaling is that there exists a stochastic process, relating the wavelet coefficients at two different scales, which is *independent* across scales [44,45]. This can be seen more clearly by formulating the SS property in Eq. (3) in statistical terms. In fact that equation is equivalent to the following relation

$$\alpha_\Psi(l,\vec{x}) \doteq \eta(l/L)\alpha_\Psi(L,\vec{x}) \tag{4}$$

for any pair of scales $l < L$; the symbol $\doteq$ means that both sides of the equation are distributed in the same way. The variable $\eta(l/L)$ has important properties: (1) it is statistically independent of $\alpha_\Psi(L,\vec{x})$ and (2) it is an infinitely divisible process: given three scales $L > l' > l$, it satisfies $\eta_{l,L} \doteq \eta_{l,l'}\eta_{l',L}$. Besides, because of scale invariance, its distribution only depends on the ratio of scales $\frac{l}{L}$. This defines a multiplicative process [35]. Using the relation (4) one can predict the exponents τ_p, [44,45], which are independent of the particular wavelet chosen [46]. This is because the $\eta(l/L)$'s characterize an invariant property of images. It is clear from Eq. (4) that a linear filter cannot give a factorial code, the independent variables are the η's and some sort of non-linear operation is needed to extract them. In Section 5 we say more about how this can be done.

4. Optimal wavelet basis

4.1. Multi-feature wavelet basis

Before implementing the optimization criterion of independent resolution levels we need to introduce a few technical aspects on wavelets. In a wavelet expansion the signal is represented in successive levels of detail, from the coarsest (larger scales) to the finest details (smaller scales), which are obtained by resizing and translating some functions (wavelets) $\{\phi_r\}_{r=0}^{n-1}$. Apart from position and size each wavelet is characterized by other properties which here have been represented generically by the index r. Each ϕ_r is a feature detector of some type. This is mathematically necessary because, as we will see later, an extra dimension is required to achieve a complete description of the image ensemble. Vision serves as a guide to select the nature of the extra dimension: simple cells exhibit orientation selectivity [20] and it is then natural to identify r with an angular variable. However, for the time being it is not necessary to assign any particular meaning to it.

We have already seen that natural images have the properties of translational and scale invariances. We would like to implement a representation of images such that both invariances were already contained in it. One advantage of this is that the coding cost would be reduced, and only the particularities of the image would need to be encoded. Although no linear representation can possess both invariances, it is possible to build a representation which is translational invariant for some discretized translations, and scale invariant for some discretized changes in scale [22]. The resulting expansion belongs to the class of discrete wavelet expansions, which have been extensively used in the context of image processing and image compression [27].

The simplest way to discretize the scale is to consider a dyadic wavelet expansion. It is called dyadic because from one level of resolution to the next the scale is divided by a factor two. The largest scale is fixed as one, and then the jth scale is 2^{-j}. Assuming that the dispersion of the wavelet is of the order of the scale, it is possible to distinguish up to 2^j different blocks along each spatial dimension (2^{2j} blocks in our case, as images are bi-dimensional). A dyadic wavelet expansion for $c(\vec{x})$ corresponds to the following mathematical expression:

$$c(\vec{x}) = \sum_{r=0}^{n-1} \sum_{j=0}^{\infty} \sum_{\vec{k}\in(Z_{2^j})^2} \alpha_{rj\vec{k}} \phi_{rj\vec{k}}(\vec{x}) \tag{5}$$

where the $\alpha_{rj\vec{k}}$'s are the wavelet coefficients, the wavelets are normalized as $\int \mathrm{d}\vec{x}\, \phi_r^2(\vec{x}) = 1$ and

$$\phi_{rj\vec{k}}(\vec{x}) \equiv \phi_r(2^j\vec{x} - \vec{k}) \tag{6}$$

Not every collection of functions $\{\phi_r\}$ can be used to represent arbitrary signals $c(\vec{x})$, but they should meet some conditions to reach a compromise between localization and detail detection (i.e., the space and frequency dispersions are kept small enough. See [13] for technical details). For some particular wavelet families $\{\phi_r\}$, there exists an associated dual family $\{\tilde{\phi}_r\}$ expanding a dyadic wavelet basis $\{\tilde{\phi}_{rj\vec{k}}\}$ such that the coefficients $\alpha_{rj\vec{k}}$ can be retrieved by simple wavelet projection on $\tilde{\phi}_{rj\vec{k}}$, namely:

$$\alpha_{rj\vec{k}} = 2^{2j} \int \mathrm{d}\vec{x}\, c(\vec{x}) \tilde{\phi}_{rj\vec{k}}(\vec{x}) \tag{7}$$

Eq. (4) now reads

$$\alpha_{rj\vec{k}} \doteq \eta_{rj\vec{k}} \alpha_{r,j-1,\left[\frac{\vec{k}}{2}\right]} \tag{8}$$

This relation has a similar interpretation to Eq. (4): the variables $\eta_{rj\vec{k}}$ are independent from the $\alpha_{r,j-1,\left[\frac{\vec{k}}{2}\right]}$ and, given that for a dyadic expansion the ratio l/L is fixed, they have the same distribution for all feature types r, resolution levels j and spatial locations $\vec{k}$.

4.2. Optimality: independence of the resolution levels

In spite of the many virtues of wavelets not every wavelet can yield an efficient code. We will make use of a particular, optimal wavelet to construct an efficient code. Let us remark that the wavelet expansion in (5) is able to implement scale and translational invariances, *but not multiscaling*. Multiscaling will impose additional constraints (in this case, over the wavelet coefficients) leading to a more efficient coding, in terms of independent levels of resolution, as we will see.

In [48] we addressed the question of whether the multiplicative process also holds *point-by-point* that is, whether the $\eta_{rj\vec{k}}$'s computed from

$$\eta_{rj\vec{k}} = \alpha_{rj\vec{k}} / \alpha_{r,j-1,\left[\frac{\vec{k}}{2}\right]} \tag{9}$$

still define statistically independent variables at different scales. The answer is that this equality does not hold for arbitrary wavelets. But now the optimization criteria—the representation should minimize the redundancy across scales—can be applied to determine an optimal wavelet for which it is fulfilled. To be more precise, this means that one should search for a wavelet such that Eq. (9) is true. If one succeeds to find such a wavelet then the efficient image representation will be given not by the wavelet coefficients themselves but by the variables $\eta_{rj\vec{k}}$.

The requirement that Eq. (9) define a multiplicative process for any image, feature type, resolution and location is very strong. In fact, it completely determines a unique wavelet Ψ, the optimal average wavelet [48], which is a linear combination of the wavelets $\{\phi_r\}_{r=0}^{n-1}$ in the family.

4.3. The optimal wavelet

Multiscaling is assumed to occur in the form just described, for each feature detector ϕ_r. Let us also consider a weighted average of these detectors,

$$\Psi = \sum_{r=0}^{n-1} p_r \phi_r \tag{10}$$

for some unknown weights p_r such that $\sum_{r=0}^{n-1} p_r^2 = 1$. Given the contrasts of the N images in the dataset,

$\{c_i(\vec{x})\}_{i=1}^{N}$, the optimal average wavelet Ψ can be obtained from the average contrast $C(\vec{x}) = \frac{1}{N}\sum_i c_i(\vec{x})$ according to the following expression (both the average contrast and the wavelet are represented in Fourier space):

$$\widehat{\Psi}(\vec{f}) = \widehat{C}(\vec{f}) - g(\vec{f})\widehat{C}\left(\frac{\vec{f}}{2}\right) \tag{11}$$

up to a normalization constant. Here $g(\vec{f})$ is a purely geometrical factor. [1] For details of its derivation see [48].

Eq. (11) gives us a very simple procedure to obtain the average wavelet from the image dataset: it is given by the difference between the average contrast at two consecutive scales. Although we do not consider here the learning problem (the network and learning algorithm that learns the wavelet from visual stimuli), we notice that a remarkable feature of this expression is that Ψ can be learnt online: because of the linearity of the equation, each time that a new image arrives its contribution (again, given by the difference of the contrast of *this* new image at two consecutive scales) is just added to what has already been learnt throughout all the past experience.

As expected, the wavelet depends on the particular visual environment considered. We have mainly studied van Hateren's dataset [53]. The function obtained from a 1000-image training set is shown in Fig. 2. It is an edge detector and it has appeared as a consequence of the multiscaling properties of natural scenes [48].

The function Ψ generates the wavelet basis for the case of a single detector ($n = 1$). One can wonder if this is enough to describe natural images. It is not, as it was discussed in great detail in [51]. A representation with just one orientation does not allow for a good reconstruction of the image (see Fig. 5). As soon as we assume that more orientations are necessary we have to face the problem of how to obtain the optimal wavelets ϕ_r from the optimal weighted average Ψ. A possible way to do it is described in the next subsection.

4.4. Multiple orientations

In order to obtain the wavelet family $\{\phi_r\}$ it is necessary to make further assumptions. The simplest guess is that they are rotated versions of the same detector, and that they are all mutually orthogonal. In [51] the theoretical derivation to extract ϕ_r from Ψ is presented. There exist in general several possible choices for the first feature detector ϕ_0 (from which all the others are obtained by simple rotation); the simplest is the one which resembles the most to Ψ. As an experimental observation $\phi_0 = \Psi$ with a great accuracy, up to $n = 8$ different orientations. So, for this image ensemble, it is possible to take the average detector Ψ as the general feature detector; we will make use of this in the following.

Before discussing the evidence in favor of the assumptions made to obtain the optimal detector (see Section 6) we now present the main implications of the results.

5. Implications of the efficient representation

Scale invariance leads to a non-linear code: Scale invariance gives a non-linear efficient representation. Non-linearities of simple cells are indeed well-known (see e.g. [8,24]) The efficient code has a first, linear stage where the orientation is detected. As we have seen, this linear response *is not* an efficient code. It corresponds to the wavelet coefficients and these are very correlated through scales. The prediction however is that the *ratios* of coefficients at different scales are independent [44]. This is a non-linear operation.

The non-linearity as divisive normalization: There is still much freedom left about how to implement the non-linearity. If the ratio between coefficients at different scales are independent, any function of them is also independent. In particular this can accommodate divisive normalization, as proposed in [8,19]. Let us notice that also in this more phenomenological work a first linear stage where the feature (orientation) is detected is followed by the non-linearity. The observation that ratios of wavelet coefficients decrease statistical dependencies was also noticed empirically in [55] and its connection with divisive normalization was studied in [56].

Another possible way to deal with the non-linearity is to consider a log-transfer after the linear computation; this alternative would be in agreement with the logarithmic fit of simple cell responses reported in [24].

Spatial correlations: Our criterion of efficiency has been independence of the representations at different scales. This does not implies the absence of dependencies between cells coding features at the same scale. These dependencies are however of very short range [48,49]. A systematic study of these correlations (which indeed still carry a lot of information about the image) has not been done yet, here we just mention that these correlations may be processed either before the non-linearity or after it. The elimination of this type of redundancy could give rise to, e.g., bar detectors.

Network implementation: The previous discussions suggest a network architecture that implements the way the independent representation could be extracted [49]. A very sketchy architecture is shown in Fig. 3.

Relation between multiscaling and the power spectrum: Given the scale invariance of images their power spec-

[1] Its complete expression is $g(\vec{f}) = \frac{\Lambda(\vec{f})}{8\Lambda(\frac{\vec{f}}{2})}$, where $\Lambda(\vec{f}) = (1 - \mathrm{e}^{-2\pi i f_1}) \times (1 - \mathrm{e}^{-2\pi i f_2})$.

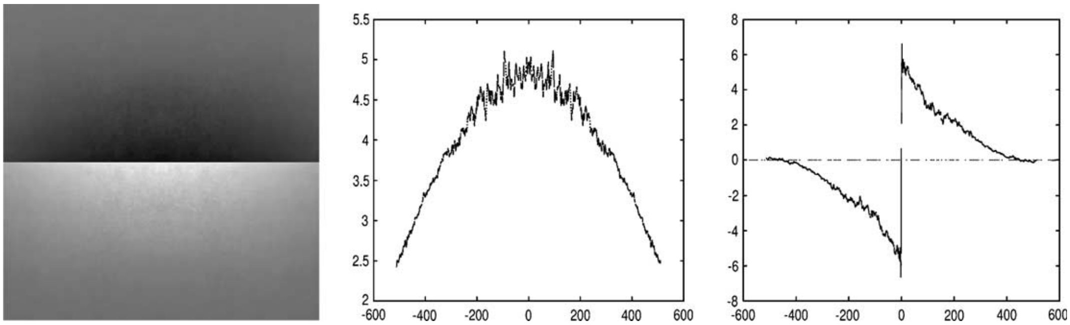

Fig. 2. Left: gray level representation of the optimal wavelet Ψ. Middle: horizontal cut. Right: vertical cut.

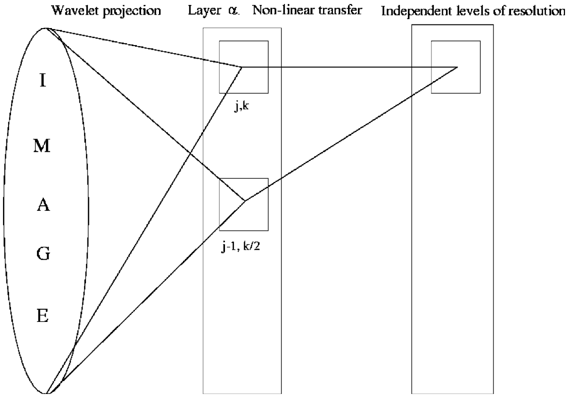

Fig. 3. A possible neural architecture to extract independent features as predicted by the scale invariant properties of images. The image is captured on the photoreceptor layer and it is then projected forward by the optimal wavelets to produce "layer α" cells. However the activity on the second layer is not necessarily given by the linear transformation since inhibitory interactions between these cells can implement the non-linearity, e.g., in the form of divisive normalization [8,19]. Alternatively, a logarithmic transfer could give independent responses of cells coding for features at different scales on a third layer. Only this latter case is shown here.

trum behaves as in Eq. (1). Are the multiscaling properties present in the higher order statistics compatible with the well-known power law behavior of the second order statistics? To answer this question we first notice that using the wavelet representation of the contrast, Eq. (5), to compute the power spectrum we have (for a translationally invariant ensemble)

$$S(\vec{f}) = \langle |\hat{c}|^2(\vec{f})\rangle = \sum_{j=0}^{\infty} 2^{-2j}\langle\eta^2\rangle^j \sum_{r=0}^{n-1} p_r^2 |\hat{\phi}_r|^2(2^{-j}\vec{f}) \qquad (12)$$

We now compute the modulus of the Fourier transform of the optimal wavelet, $|\widehat{\Psi}(\vec{f})|$. Its average over all orientations is represented in Fig. 4 together with a fit to a $1/f$ law. The agreement of the fit is very good.

It is immediate from Eq. (12) that a wavelet Ψ such that $|\widehat{\Psi}|(\vec{f}) \sim f^{-1}$ leads to the correct power spectrum (the correction exponent δ and the weak anisotropy come out from the uneven weightings p_r for the different orientations in the orientational wavelet expansion).

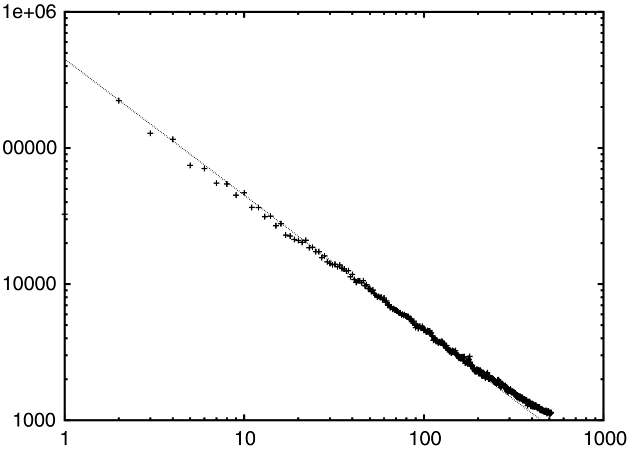

Fig. 4. Orientational average of $|\widehat{\Psi}|(\vec{f})$ in log–log scale and best fit with a k/f curve, k constant.

The power law behavior of the optimal wavelet is extremely important for the compatibility between second order scaling and multiscaling. If $|\widehat{\Psi}|$ is different from a power law then from Eq. (12) it follows that

$$S(2\vec{f}) \approx 2^{-2}\langle\eta^2\rangle S(\vec{f}) \qquad (13)$$

According to [46], $\langle\eta^2\rangle = 2^{-(2+\tau_2)}$ and $-1 < \tau_2 < 0$. We thus obtain $S(2\vec{f}) \approx 2^{-4-\tau_2}S(\vec{f})$ and in general $S(a\vec{f}) \approx a^{-4-\tau_2}S(\vec{f})$, that is, $S(\vec{f}) \sim f^{-4-\tau_2}$. Hence any wavelet such that $|\widehat{\Psi}| \neq f^{-1}$ would give rise to an incorrect exponent for the power spectrum.

Representation of images in the optimal basis: Fig. 5 shows the representation with $n = 1$, 2 and 3 of one particular image. The optimal number of orientations seems to be $n = 2$, although larger number of orientations could be used to introduce redundancy and stability in the presence of noise.

Image reconstruction: The scaling property for $|\widehat{\Psi}|$ also allows to establish a link between the dyadic representation and the reconstruction algorithm proposed in [50]. In that paper, the authors show that images can be reconstructed from the values of contrast changes over the borders (which are identified with the most singular manifold in the multifractal structure [46]). The

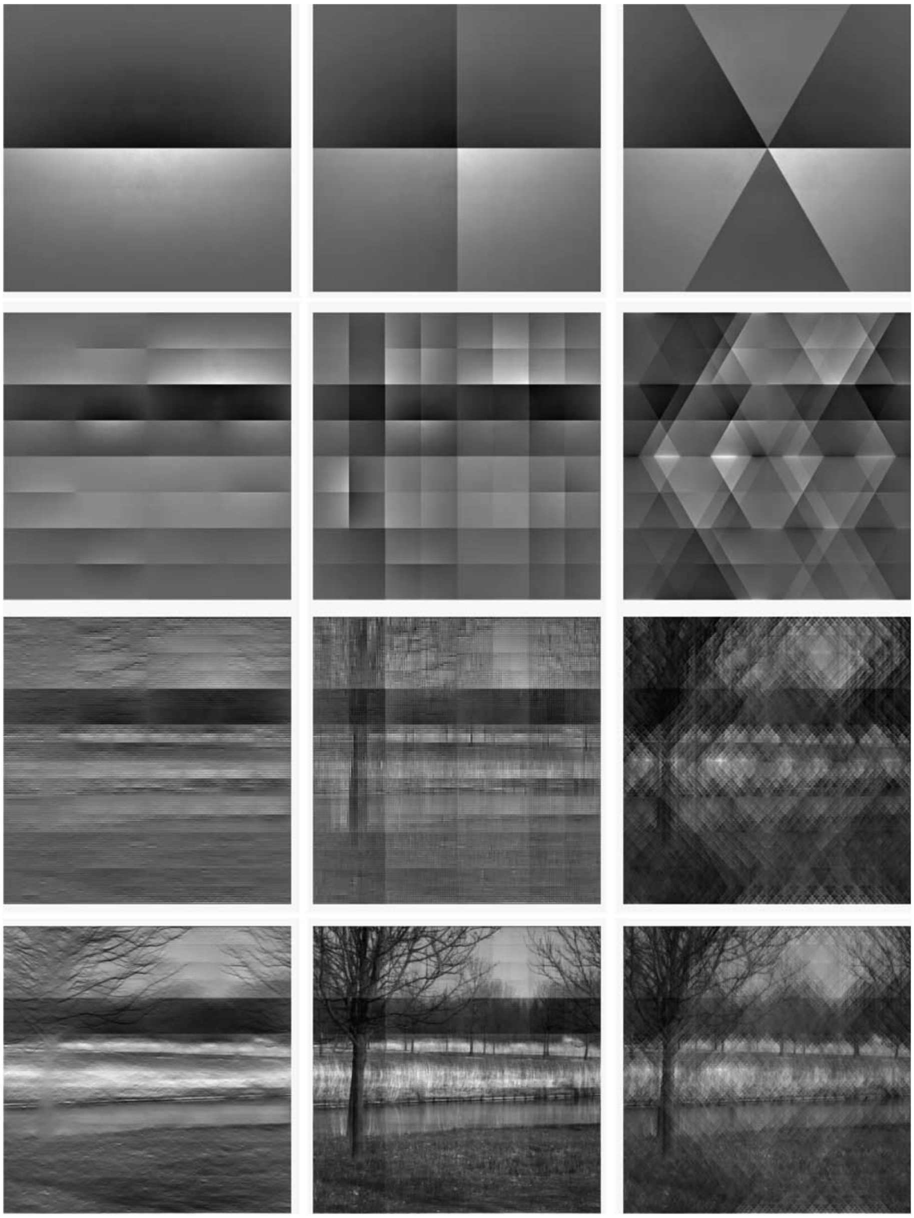

Fig. 5. $\sum_{rj\vec{k}} \alpha_{rj\vec{k}} \Psi_{rj\vec{k}}(\vec{x})$ for $j = 0$ (top), $j \leqslant 2$, $j \leqslant 6$ and $j \leqslant 8$ (bottom) for imk00640.imc (from van Hateren's dataset) with $n = 1$ (left), 2 (middle) and 3 (right) orientations.

reconstruction formula is essentially a diffusion of the values of the contrast along the edges according to a kernel which behaves as $1/f$ in Fourier space. As it can be seen in Fig. 5, the wavelet expansion works much in the same way: each resized, translated wavelet appearing in the sum in Eq. (5) is equivalent to a light-spreading edge element of that size and location, weighted with the appropriated coefficient $\alpha_{rj\vec{k}}$.

6. Consistency requirements

Several assumptions have been made, first to obtain the optimal average wavelet and after to extract the oriented edge detectors. Here we summarize the numerical evidence assessing the consistency of the model for the considered image ensemble.

The first property that should be checked is the orthogonality between feature detectors. In fact, a stronger statement can be proved: for some n's, the wavelet basis is self-dual ($\tilde{\phi}_{rj\vec{k}} = \phi_{rj\vec{k}}$), that is, it is an orthogonal wavelet basis. Then the wavelet coefficients $\alpha_{rj\vec{k}}$ can be obtained with a good approximation just projecting on $\phi_{rj\vec{k}}$. As an empirical measure of how accurately this property holds, consider the average orthogonality error defined as

$$\epsilon_{nj} \equiv \sum_{\vec{k}} \left| \left\langle \int \mathrm{d}\vec{x} \phi_0(\vec{x}) R_n \phi_{0j\vec{k}}(\vec{x}) \right\rangle \right| \qquad (14)$$

Table 1
Average orthogonality errors for $n = 1$, 2 and 3

j	0	1	2	3	4	5	6	7
ϵ_{1j}	1.000	0.106	0.023	0.006	0.004	0.003	0.003	0.002
ϵ_{2j}	0.001	0.013	0.005	0.001	0.001	0.000	0.000	0.000
ϵ_{3j}	0.085	0.028	0.037	0.025	0.013	0.007	0.003	0.002

where R_n is the rotation operator of angle π/n. [2] The values of the average error for $n = 1$, 2 and 3 are given in Table 1. [3]

The average orthogonality error gives thus a measure of the error committed by assuming that the wavelet basis is orthogonal. They should be zero for all n and j, except for $n = 1$ and $j = 0$.

For $n = 1$ (first row in the Table) it provides a measure of the self-duality of the wavelet. The result is good except for $j = 1$, where the error is about 10%. We think that this value is mainly due to finite size effects in the image and finite sampling.

We observe that for $n = 2$ orientations the wavelets are close to orthogonality; however, for $n = 3$ there is a small coupling for several scales j. Note however that it is just orthogonality among features which is essential in the derivation (i.e. $\epsilon_{n0} = 0$ for $n > 1$), the orthogonality of the whole wavelet basis being an interesting bonus. It is thus assumed that orthogonality holds for $n = 2$ and that it is just an approximation for $n = 3$. In order to obtain the wavelet coefficients $\alpha_{rj\vec{k}}$ in the experiments we assume $\tilde{\phi}_{rj\vec{k}} = \phi_{rj\vec{k}}$ and apply Eq. (7).

The second hypothesis to be tested is that of independence among scales. This was studied in [48,49]. Let us notice that the hypothesis only requires independence between $\eta_{rj\vec{k}}$ and $\alpha_{r,j-1,[\frac{\vec{k}}{2}]}$ at every scale j, location $\vec{k}$ and orientation n. In [51] this independence was checked by measuring the mutual information [10] between $\eta_{rj\vec{k}}$ and $\alpha_{r,j-1,[\frac{\vec{k}}{2}]}$ for a subensemble of 100 images, assuming translational invariance to increase sampling. The calculated mutual informations were smaller than 10^{-3} bits (compared to a maximum of 11 bits) at all scales j and $n = 2$.

7. Conclusions

In this paper we have first reviewed the concept of optimality and its connection with sensory coding. Optimality can be described according to different criteria, among which the main two are redundancy reduction and infomax. We have seen that, under appropriate conditions, to reduce the redundancy and to maximize the information transfer are equivalent, leading to the same concept of optimality. Then, we have seen that optimality can only be obtained when the regularities of the stimuli are well-understood and used to reduce the redundancies in the code.

We have discussed the known statistical regularities of natural images as sensory stimuli. We have seen that there are important statistical symmetries in natural images: translational invariance, scale invariance and multiscaling. Translational invariance means that there is no preferred point in the scenes. Scale invariance means that objects can appear with any apparent size. Multiscaling means that objects are hierarchically structured in different parts (edges, textures) which are scale-invariant but behave differently under changes of scale. Each one carries a different type of redundancy. We have discussed the different approaches proposed to deal with some of those symmetries. The main part of the paper deals with multiscaling treated within the wavelet approach, which is the only one which takes the three symmetries into account.

The multiscaling wavelet approach is based on a wavelet expansion. Wavelet expansions allow us to represent signals as the combination of acitivities (wavelet coefficients) extracted using filters (wavelets) focussing at different scales and at different positions. Wavelet representations of natural stimuli are optimal for implementing both translational and scale invariances, but not multiscaling. In the multiscaling wavelet approach, there exists an optimal wavelet implementing multiscaling and leading to a minimum redundancy representation. The code is produced by decomposing any image in independent levels of resolution. In contrast with standard wavelet representations, the emergence of an object or structure as the scale becomes finer is coded just once, at the scale of first detection. The effect of increased activity in the wavelet coefficients is removed from the finer scales.

The optimal wavelet has remarkable properties: edge detection in a finite number of orientations, independency of the power spectrum, extraction of independent levels of resolution, online learning. Besides, the code is sparse (without explicitly requiring that property) because edges are sparse in images. All these properties are

[2] The wavelet experimentally obtained is antisymmetric, so it must be rotated an angle of π/n instead of $2\pi/n$.

[3] In the three cases, all the possible overlaps among different scales and orientations are equivalent to those provided by ϵ_{1j} or ϵ_{nj} due to the invariance of the inner product under the action of the rotation operator R_n; for that reason we concentrate on that cases.

observed in the first levels of visual processing. The optimal wavelet is thus a good candidate for modelling visual information processing in the brain.

The multiscaling wavelet approach is not, however, the ultimate model for neural visual processing. The multiscaling wavelet code has still some redundancies (among orientations and spatial locations) which could be removed in order to diminish redundancy and improve the code. An opposite concern is that of overcomplete representations. Once a complete, optimal code were accessible by means of these techniques, overcomplete representations could be explored, which would be more likely to describe real biological coding—in particular some redundancy is required to process information in the presence of noise [1,14] and to insure stability of the representation against translations of the image [42].

Acknowledgements

A. Turiel is financially supported by a post-doctoral fellowship from INRIA. This work has been funded by a Spanish grant (BMF2000-0011). NP thanks the Institute Henri Poincare (Paris) for their kind hospitality.

References

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

[2] F. Attneave, Informational aspects of visual perception, Psychol. Rev. 61 (1954) 183–193.

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

[4] H.B. Barlow, What is the computational goal of the neocortex?, in: C. Koch, J. Davis (Eds.), Large Scale Neuronal Theories of the Brain, MIT Press, Cambridge, MA, 1994, pp. 1–22 (Chapter 1).

[5] A. Bell, T. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (1995) 1129–1159.

[6] A.J. Bell, T.J. Sejnowski, The independent components of natural scenes are edge filters, Vision Res. 37 (1997) 3327–3338.

[7] N. Brenner, W. Bialek, R. de Ruyter van Steveninck, Adaptive rescaling maximizes information transmission, Neuron 26 (2000) 695–702.

[8] M. Carandini, D.J. Heeger, J.A. Movshon, Linearity and normalization of simple cells of the macaque primary visual cortex, J. Neurosci. 17 (1997) 8621–8644.

[9] M. Carandini, H.B. Barlow, L.P. O'Keefe, A.B. Poirson, J.A. Movshon, Adaptation to contingencies in macaque primary visual cortex, Phil. Trans. Roy. Soc. B 352 (1997) 1149–1154.

[10] T.M. Cover, J.A. Thomas, Elements of Information Theory, John Wiley, New York, 1991.

[11] M.S. Crouse, R.D. Nowak, R.G. Baraniuk, Wavelet-based statistical signal processing using hidden markov models, IEEE Trans. Signal Process. 46 (1998) 886–902.

[12] Y. Dan, J.J. Atick, R.C. Reid, Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory, J. Neurosci. 16 (1996) 3351–3362.

[13] I. Daubechies, Ten Lectures on Wavelets, CBMS-NSF Series in Applied Mathematics, Capital City Press, Montpelier, Vermont, 1992.

[14] P.D. Giudice, A. Campa, N. Parga, J.-P. Nadal, Maximization of mutual information in a linear noisy network: a detailed study, Network: Comput. Neural Syst. 6 (1995) 449–468.

[15] K. Falconer, Fractal Geometry: Mathematical Foundations and Applications, John Wiley and sons, Chichester, 1990.

[16] D.J. Field, Relations between the statistics of natural images and the response properties of cortical cells, J. Opt. Soc. Am. 4 (1987) 2379–2394.

[17] D.J. Field, Scale-invariance and self-similar 'wavelet' transforms: an analysis of natural scenes and mammalian visual systems, in: M. Farge, J.C.R. Hunt, J.C. Vassilicos (Eds.), Wavelets, Fractals, and Fourier Transforms, Clarendon Press, Oxford, 1993, pp. 151–193.

[18] D.J. Field, What is the goal of sensory coding?, Neural Comput. 6 (1994) 559–601.

[19] D. Heeger, Normalization of cell responses in cat striate cortex, Visual Neurosci. 9 (1992) 181–198.

[20] D. Hubel, T. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, J. Physiol. 160 (1962) 154.

[21] S.B. Laughlin, A simple coding procedure enhances a neuron's information capacity, Z. Naturf. 36 (1981) 910–912.

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

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

[24] L. Maffei, A. Fiaorentini, The visual cortex as a spatial frequency analyser, Vision Res. 13 (1973) 1255–1267.

[25] S. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (1989) 674–693.

[26] S. Mallat, S. Zhong, Characterization of signals from multiscale edges, IEEE Trans. Pattern Anal. Mach. Intell. 14 (1992) 710–732.

[27] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, London, 1999.

[28] G. Mato, N. Parga, Dynamic changes in receptive fields induced by cortical reorganization, in: R. Baddeley, P. Hancock, P. Földiák (Eds.), Information Theory and the Brain, Cambridge University Press, Cambridge, 2000, pp. 122–138 (Chapter 7).

[29] H.R. Maturana, S. Frenk, Unidirectional movement and horizontal edge detectors in pigeon retina, Science 142 (1963) 977–979.

[30] J.-P. Nadal, N. Parga, Information processing by a perceptron in an unsupervised learning task, Network 4 (1993) 295–312.

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

[32] J.-P. Nadal, N. Parga, Redundancy reduction and independent component analysis: conditions on cumulants and adaptive approaches, Neural Comput. 9 (1997) 1421–1456.

[33] J.-P. Nadal, N. Brunel, N. Parga, Nonlinear feedforward networks with stochastic outputs: infomax implies redundancy reduction, Network: Comput. Neural Syst. 9 (1998) 1–11.

[34] A. Nevado, A. Turiel, N. Parga, Scene dependence of the non-gaussian scaling properties of natural images, Network 11 (2000) 131–152.

[35] E.A. Novikov, Infinitely divisible distributions in turbulence, Phys. Rev. E 50 (1994) R3303.

[36] B. Olshausen, D.J. Field, Sparse coding with an overcomplete basis set: a strategy employed by v1?, Vision Res. 37 (1997) 3311–3325.

[37] D.-T. Pham, P. Garrat, C. Jutten, Separation of a mixture of independent sources through a maximum likelihood approach, in: EUSIPCO, 1992, pp. 771–774.

[38] D.L. Ruderman, W. Bialek, Statistics of natural images: scaling in the woods, Phys. Rev. Lett. 73 (1994) 814.

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

[40] D.L. Ruderman, Origins of scaling in natural images, Vision Res. 37 (1997) 3385–3389.

[41] M. Sigman, G. Cecchi, C. Gilbert, M. Magnasco, On a common circle: natural scenes and gestalt rules, Proc. Natl. Acad. Sci. USA 98 (2001) 1935–1940.

[42] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, D.J. Heeger,, Shiftable multi-scale transforms [or "what's wrong with orthonormal wavelets"], IEEE Trans. Inform. Theory, Special Issue on Wavelets 38 (2) (1992) 587–607.

[43] W.A. Truccolo, D.W. Dong, Dynamic temporal decorrelation: an information-theoretic and biophysical model of the functional role of the lateral geniculate nucleus, Neurocomputing 38–40 (2001) 993–1001.

[44] A. Turiel, G. Mato, N. Parga, J.P. Nadal, Self-similarity properties of natural images, in: Proc. of NIPS'97, vol. 10, MIT Press, 1997, pp. 836–842.

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

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

[47] A. Turiel, N. Parga, D. Ruderman, T. Cronin, Multiscaling and information content of natural color images, Phys. Rev. E 62 (2000) 1138–1148.

[48] A. Turiel, N. Parga, Multifractal wavelet filter of natural images, Phys. Rev. Lett. 85 (2000) 3325–3328.

[49] A. Turiel, N. Parga, Wavelet based decomposition of natural images in independent resolution levels, in: P. Pajunen, J. Karhunen (Eds.), Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000), Helsinki, Finland, 2000, pp. 339–344.

[50] A. Turiel, A. del Pozo, Reconstructing images from their most singular fractal manifold, IEEE Trans. Im. Proc. 11 (2002) 345–350.

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

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

[53] 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. B 265 (1998) 359–366.

[54] M.J. Wainwright, Visual adaptation as optimal information transmission, Vision Res. 39 (1999) 3960–3974.

[55] M.J. Wainwright, E.P. Simoncelli, Scale mixtures of gaussians and the statistics of natural images, in: S.A. Solla, T.K. Leen, K.-R. Müller (Eds.), Adv. Neural Inform. Process. Syst., 12, MIT Press, 2000, pp. 855–861.

[56] M.J. Wainwright, O. Schwartz, E.P. Simoncelli, Natural image statistics and divisive normalization: modeling nonlinearity and adaptation in cortical neurons, 2001.

[57] X.-J. Wang, Y. Liu, M.V. Sanchez-Vives, D.A. McCormick, Adaptation and temporal decorrelation by single neurons in the primary visual cortex, J. Neurophysiol 89 (2003) 3279–3293.