Towards a Theory of the Striate Cortex[1]

Published in Neural Computation, Vol 6, number 1, January 1994, p127-146

Zhaoping Li and Joseph J. Atick
The Rockefeller University
1230 York Avenue
New York, NY 10021, USA

Abstract

We explore the hypothesis that linear cortical neurons are concerned with building a particular type of representation of the visual world — one which not only preserves the information and the efficiency achieved by the retina, but in addition preserves spatial relationships in the input — both in the plane of vision and in the depth dimension. Focusing on the *linear* cortical cells, we classify all transforms having these properties. They are given by representations of the scaling and translation group, and turn out to be labeled by rational numbers '$(p+q)/p$' (p, q integers). Any given (p, q) predicts a set of receptive fields which come at different spatial locations and scales (sizes) with a bandwidth of $\log_2[(p+q)/p]$ octaves, and, most interestingly, with a diversity of 'q' cell varieties. The bandwidth affects the trade-off between preservation of planar and depth relations, and, we think, should be selected to match structures in natural scenes. For bandwidths between 1 and 2 octaves, which are the ones we feel provide the best matching, we find for each scale a minimum of two distinct cell types that reside next to each other and in phase quadrature, i.e., differ by 90^o in the phases of their receptive fields, as are found in the cortex, they resemble the "even-symmetric" and "odd-symmetric" simple cells in special cases. An interesting consequence of the representations presented here is that the pattern of activation in the cells in response to a translation or scaling of an object remains the same but merely shifts its locus from one group of cells to another. This work also provides a new understanding of color coding changes from the retina to the cortex.

[1]Work supported in part by a grant from the Seaver Institute.

1. Introduction

What is the purpose of the signal processing performed by neurons in the visual pathway? Are there first principles that predict the computations of these neurons? Recently there has been some progress in answering these questions for neurons in the early stages of the visual pathway. In Atick and Redlich (1990,1992) a quantitative theory, based on the principle of redundancy reduction, was proposed. It hypothesizes that the main goal of retinal transformations is to eliminate redundancy in input signals, particularly that due to pairwise correlations among pixels — second-order statistics.[2] The predictions of the theory agree well with experimental data on processing of retinal ganglion cells (Atick and Redlich 1992, Atick et al 1992).

Given the successes of this theory, it is natural to ask whether redundancy reduction is a computational strategy continued into the striate cortex. One possibility is that cortical neurons are concerned with eliminating higher-order redundancy, which is due to higher-order statistics. We think this is unlikely. To see why, we recall the facts that make redundancy reduction compelling when applied to the retina, and see that these facts are not as relevant for the cortex.

First, the retina has a clear bottleneck problem: the amount of visual data falling on the retina per second is enormous, of the order of tens of megabytes, while the retinal output has to fit into an optic nerve of a dynamic range significantly smaller than that of the input. Thus, the retina must compress the signal, and it can do so without significant loss of information by reducing redundancy. In contrast, after the signal is past the optic nerve, there is no identifiable bottleneck that requires continued redundancy reduction beyond the retina.

Second, even if there were pressure to reduce data[3], eliminating higher-order statistics does not help. The reason is that higher-order statistics do not contribute significantly to the entropy of images, and hence no significant compression can be achieved by eliminating them (for reviews of information theory see Shannon and Weaver 1949; Atick 1992). The dominant redundancy comes from pairwise correlations.[4]

There is another intrinsic difference between higher and second-order statistics that suggests their different treatment by the visual pathway. Fig. 1 shows image A and another image B which was obtained by randomizing the phases of the fourier coefficients of A. B thus has the same second-order statistics as A but no higher order ones. Contrary to A, B has no clear forms or structures (cf. Field 1989). This suggests that second-order statistics are useless, while higher-order ones are essential, for defining forms and for dis-

[2]Since retinal neurons receive noisy signals it is necessary to formulate the redundancy reduction hypothesis carefully taking noise into account. In Atick and Redlich (1990,1992) a generalized notion of redundancy was defined, whose minimization leads to elimination of pairwise correlations and to noise smoothing.

[3]For example there could be a computational bottleneck such as an attentional bottleneck occuring deep into the cortex — perhaps in the link between $V4$ and IT (Van Essen et al 1991).

[4]This fact is well known in the television industry (see $e.g.$ Schreiber 1956). This is why practical compression schemes for television signals never take into account more than pairwise correlations, and even then, typically nearest neighbor correlations. This fact was also verified for several scanned natural images in our laboratory by N. Redlich and by Z. Li.

criminating between images. Actually, eliminating the former highlights the higher-order statistics which should be used to extract form signals from "noise." [5]

So what is the cortex then trying to do? Ultimately, of course, the cortex is concerned with object and pattern recognition. One promising direction could be to use statistical regularities of images to discover matched filters which lead to better representations for pattern recognition. Research in this direction is currently underway. However, there is another important problem that a perceptual system has to face before the recognition task. This is the problem of *segmentation*, or equivalently, the problem of grouping features according to a hypothesis of which objects they belong to. It is a complex problem, which may turn out not to be solvable independently from the recognition problem. However, since objects are usually localized in space, we think an essential ingredient for its successful solution is a representation of the visual world where *spatial relationships*, both in the plane of vision and in the depth dimension, are preserved as much as possible.

In this paper we hypothesize that the purpose of early cortical processing is to produce a representation that 1. preserves information, 2. is free of second-order statistics, and 3. preserves spatial relationships. The first two objectives are fully achieved by the retina so we merely require that they be maintained by cortical neurons. We think the third objective is attempted in the retina ($e.g.$ retinotopic and scale invariant sampling); however, it is only completed in the cortex where more computational and organizational resources are available.

Here, we focus on the cortical transforms performed by the relatively linear cells, the first two requirements immediately limit the class of transforms that linear cells can perform on the retinal signals to the class of unitary matrices [6], $\mathbf{U}$ with $\mathbf{U} \cdot \mathbf{U}^\dagger = \mathbb{1}$. So the principle for deriving cortical cell kernels reduces to finding the $\mathbf{U}$ that best preserves spatial relationships. Actually, preserving planar and depth relationships simultaneously requires a trade off between the two (section two). This implies that there is a family of $\mathbf{U}$'s, one for every possible trade off. Each $\mathbf{U}$ is labelled by the bandwidth of the resulting cell filters and forms a representation of the scaling and translation group (section three). We show that the requirement of unitarity limits the allowed choices of bandwidths, and for each choice predicts the needed cell diversity. The bandwidth that should ultimately be selected is the one that best matches structures in natural scenes. For bandwidths around 1.6 octaves, which are the ones we feel are most relevant for natural scenes, the predicted cell kernels and cell diversity resemble those observed in the cortex.

The resulting cell kernels also possess an interesting *object constancy* property: when an object in the visual field is translated in the plane or perpendicular to the plane of vision, the pattern of activation it evokes in the cells remains intrinsically the same but shifts its locus from one group of cells to another, leaving the same total number of cells activated. The importance of such representations for pattern recognition has been stressed repeat-

[5]Extracting signal from noise can achieve by far more significant data reduction than trying to eliminate higher-order correlations.

[6]In this paper we use the term "unitary" instead of "orthogonal" since we find it more convenient to use complex basis ($e.g.$ e^{ifx} instead of $\cos(fx)$). $\mathbf{U}^\dagger \equiv \mathbf{U}^{*T}$, where $*$ denotes complex conjugate. For real matrices, unitary means orthogonal.

edly by many people before and recently by Olshausen et al (1992). Furthermore, this work provides a new understanding of color coding change from the single opponency in the retina to the double opponency in the cortex.

2. Manifesting Spatial Relationships

In this section we examine the family of decorrelating maps and see how they differ in the degree with which they preserve spatial relationships. We start with the input, represented by the activities of photoreceptors in the retina, $\{S(\underset{\sim}{x}_n)\}$ where $\underset{\sim}{x}_n$ labels the spatial location of the n'th photoreceptor in a two-dimensional ($2D$) grid. For simplicity we take the grid to be uniform. To focus on the relevant issues without the notational complexity of $2D$, we first examine the one-dimensional ($1D$) problem and then generalize the analysis to $2D$ in section four. The autocorrelator of the signals $\{S(x_n)\}$ is

$$R_{nm} \equiv < S(x_n)S(x_m) >, \tag{1}$$

where brackets denote ensemble average. To eliminate this particular redundancy one has to decorrelate the output and then apply the appropriate gain control to fit the signals into a limited dynamic range. This can be achieved by a linear transformation

$$O_j = \sum_{n=1}^{N} K_{jn} S(x_n), \tag{2}$$

where $j = 1, \cdots, N$ and the kernel K_{jn} is the product of two matrices. Using bold-face to denote matrices:

$$\mathbf{K} = \mathbf{V} \cdot \mathbf{M}. \tag{3}$$

M_{jn} is the rotation to the principal components of $\mathbf{R}$: $(\mathbf{M} \cdot \mathbf{R} \cdot \mathbf{M}^T)_{ij} = \lambda_i \delta_{ij}$, where $\{\lambda_i\}$ are the eigenvalues of $\mathbf{R}$. While $\mathbf{V}$ is the gain control which is a diagonal matrix with elements $V_{ii} = \frac{1}{\sqrt{\lambda_i}}$. Thus the output has the property

$$< O_i O_j >= (\mathbf{K} \cdot \mathbf{R} \cdot \mathbf{K}^T)_{ij} = \delta_{ij}. \tag{4}$$

An important fact to note is that redefining $\mathbf{K}$ by $\mathbf{K}' = \mathbf{U} \cdot \mathbf{K}$ where $\mathbf{U}$ is a unitary matrix ($\mathbf{U} \cdot \mathbf{U}^\dagger = 1$) does not alter the decorrelation property (4). (Actually $\mathbf{U}$ should be an orthogonal matrix for real O_i, but since we will for convenience use complex variables, unitary $\mathbf{U}$ is appropriate.) Therefore, there is a whole family of equally efficient representations parametrized by $\{\mathbf{U}\}$. Any member is denoted by $\mathbf{K_U}$

$$\mathbf{K_U} = \mathbf{U} \cdot (\mathbf{V} \cdot \mathbf{M}) \equiv \mathbf{U} \cdot \mathbf{K}^{(p)}, \tag{5}$$

where $\mathbf{K}^{(p)} \equiv \mathbf{V} \cdot \mathbf{M}$ is the transformation to the principal components. Without compromising efficiency, this non-uniqueness allows one to look for a specific $\mathbf{U}$ that leads to $\mathbf{K_U}$ with other desirable properties such as manifest spatial relationships.[7]

To see this let us exhibit the transformation $\mathbf{K}^{(p)}$ more explicitly. For natural signals, the autocorrelator is translationally invariant, in the sense that $R_{nm} = R(n-m)$. One can

[7]It should be noted that this non-uniqueness in receptive field properties is due to the fact that the

then define the autocorrelator by its fourier transform or its power spectrum, which in $2D$ is $R(\underline{\mathrm{f}}) \sim 1/|\underline{\mathrm{f}}|^2$, where $\underline{\mathrm{f}}$ is the $2D$ spatial frequency (Field 1987, Ruderman 1992). For illustration purposes we take in this section the analogous $1D$ "scale invariant" spectrum, namely $R(f) \sim 1/f$. In the $2D$ analysis of section four we use the measured spectrum $\sim 1/|\underline{\mathrm{f}}|^2$.

For a translationally invariant autocorrelator, the transformation to principal components is a fourier transform. This means, the principle components of natural scenes or the row vectors of the matrix $\mathbf{M}$ are sine waves of different frequencies

$$M_{jn} = \frac{1}{\sqrt{N}} e^{-i f_j x_n}, \tag{6}$$

where

$$\begin{aligned} j &= (0, 1, 2, \ldots, N-1) \\ f_j &= \begin{cases} \frac{2\pi}{N}\frac{j+1}{2} & \text{if } j \text{ is odd} \\ -\frac{2\pi}{N}\frac{j}{2} & \text{if } j \text{ is even} \end{cases} \end{aligned}$$

While the gain control matrix $\mathbf{V}$ is[8] $V_{jj} = 1/\sqrt{R(f_j)} = \sqrt{|f_j|}$. The total transform then becomes

$$K^{(p)}_{jn} = \frac{1}{\sqrt{N}} \cdot \sqrt{|f_j|} e^{-i f_j x_n}. \tag{7}$$

This performs a fourier transform and at the same time normalizes the output such that the power is equalized among frequency components $< O_i^2 >=$ const., $i.e.$ output is whitened. One undesirable feature of the transformation $\mathbf{K}^{(p)}$ is that it does not preserve spatial relationships in the plane. As an object is translated in the field of view the locus of response $\{O_i\}$ will not simply translate. Also two objects separated in the input do not activate two separate groups of cells in the output. Typically all cells respond to a mixture of features of all objects in the visual field. Segmentation is thus not easily achievable in this representation.

Mathematically, we say that the output $\{O_i\}$ preserves planar spatial relationships in the input if

$$O_i[S] = O_{i-m}[S'] \quad \text{when} \quad S'(x_n) = S(x_{n+m}), \tag{8}$$

principle used is decorrelation. If one insists on minimization of pixel entropy (which for gaussian signals is equivalent to decorrelation) this symmetry formally does not exist for ensembles of non-gaussian signals. In other words some choice of $\mathbf{U}$ may be selected over others. However for the ensemble of 40 images that we have considered we found that the pixel entropy varied only by few percent for different $\mathbf{U}$'s. This is consistent with the idea that natural scenes are dominated by second-order statistics which do not select any particular $\mathbf{U}$. In other systems it is possible that higher order statistics do select a special $\mathbf{U}$, see for example (Hopfield 1991). For another point of view see (Linsker 1992).

[8] In June 2010, a typo was discovered in this expression for V_{jj} and in the equation below. The expression $\sqrt{|f_j|}$ was mistakenly written as $1/\sqrt{|f_j|}$, and this typo was also in the published version in *Neural Computation* and in the manuscript posted on my webpage before June 2010. I apologize for any inconvenience caused by this typo.

where $O_i[S] \equiv \sum_{n=1}^{N} K_{in} S(x_n)$. In other words, a translation in the input merely shifts the output from one group of cells to another. Implicitly, preserving planar spatial relationship also requires, and we will therefore enforce, that the cell receptive fields be local, so a spatially localized object evokes activities only in a local cell group, which shifts its location when the object moves and is seperated from another cell group evoked by another spatially disjoint object in the image plane. Technically speaking an $\{O_i\}$ that satisfies (8) is said to form a representation of the discrete "translation group".

Insisting on (8) picks up a unique choice of $\mathbf{U}$. In fact in this case $\mathbf{U}$ is given by

$$U_{nj} = M_{nj}^* = \frac{1}{\sqrt{N}}\, e^{i f_j x_n}, \tag{9}$$

which is just the inverse fourier transform. The resulting transformation $\mathbf{K}^{(t)} = \mathbf{U} \cdot \mathbf{V} \cdot \mathbf{U}^{\dagger}$ gives translationally invariant center-surround cell kernels $K_{nm} = K(n-m) = \sum_j U_{nj} V_{jj} U_{mj}^* \propto \sum_f cos(f(x_n - x_m))/\sqrt{R(f)}$. In two dimensions, taking into account optical properties of the eye medium and the noise, these kernels were shown to account well for properties of retinal ganglion cells (Atick and Redlich 1992).

Although the representation defined by $\mathbf{K}^{(t)}$ is ideal for preserving spatial relationships in the plane, it completely destroys spatial relations **in scale or** depth dimension. The change in the patterns of activation in $\{O_i\}$ in response to a change in the object distance is very complicated. To preserve depth relations the output should form a representation of another group, the so called "scaling group". This is because when an object recedes or approaches, the image it projects goes from $S(x)$ to $S(\lambda x)$ for some scale factor λ. The requirement of object invariance under scaling dictates that

$$O_i[S(\lambda x)] = O_{i+l}[S(x)] \tag{10}$$

for some shift l depending on λ. It is not difficult to see that $\mathbf{K}^{(t)}$, which satisfies (8) all the way down to the smallest possible translation, violates this condition. Actually satisfying (8) and (10) for the smallest possible translation and scale changes simultaneously is not possible. A compromise between them has to be found.

The problem of finding the kernels that lead to $\{O_i\}$ with the best compromise between (8) and (10) is equivalent to the mathematical problem of constructing simultaneous representations of the translation and scaling group, which is what we do next.

3. Representations of Translation and Scaling group

To satisfy (8) and (10) the cells must carry two different labels. One is a spatial position label 'n' and the other is a scale label 'a'. The idea is that under translations of the input the output translates over the 'n' index while under scaling by some scale factor λ the output shifts over the 'a' index. Such cell groups can be obtained from $\mathbf{O} = \mathbf{U} \cdot \mathbf{K}^{(p)}$ using a $\mathbf{U}$ that is block diagonal:

$$\mathbf{U} = \begin{pmatrix} \boxed{U^0} & & & \\ & \boxed{U^1} & & \\ & & \boxed{U^2} & \\ & & & \ddots \end{pmatrix}$$

Each submatrix $\mathbf{U}^a$ has dimension N^a and gives rise to N^a cells with outputs O^a_n located at lattice points $x^a_n = (N/N^a)n$ for $n = 1, 2, \ldots, N^a$. Since the block matrices $\mathbf{U}^a$ act on $\mathbf{K}^{(p)}$ which are the fourier modes of the inputs, the resulting cells in any given block a filter the inputs through a limited and exclusive frequency band with frequencies f_j for $\sum_{a'<a} N^{a'} \le j < \sum_{a' \le a} N^{a'}$. Since $N^a < N$ these cells sample more sparsely on the original visual field. Notice, the cells from different blocks a are spatially mingled with each other and their total number add up to $N = \sum_a N^a$. The hope is to have translation invariance within each block and scale invariance between blocks, i.e.,

$$\begin{aligned} O^a_n[S] &= O^a_{n+\delta n}[S'] \quad \text{for } S'(x) = S(x+\delta x) \text{ and } \delta x = (N/N^a)\delta n && (11) \\ O^a_n[S] &= O^{a+1}_n[S'] \quad \text{for } S'(x) = S(\lambda x) && (12) \end{aligned}$$

Each block 'a' thus represents a particular scale, the translation invariance within that scale can be achieved with a resolution $\delta x \propto N/N^a$, inversely proportional to N^a. Larger blocks or larger N^a thus give better translation invariance, and the single block matrix $\mathbf{U} = \mathbf{U}^0 = \mathbf{M}^\dagger$ achieves this symmetry to the highest possible resolution. On the other hand, a higher resolution in scaling invariance calls for a smaller $\lambda > 1$. As we will see below, $(\lambda - 1) \propto N^a/f^a$, where f^a is the smallest frequency sampled by the a^{th} block. Hence a better scaling invariance requires smaller block sizes N^a. A trade-off between better translation and scaling invariance reduces to choosing the scaling factor λ, or the bandwidth depending on it. This will become clearer as we now follow the detailed construction of U. The unitarity condition now requires having $\mathbf{U}^a(\mathbf{U}^a)^\dagger = 1$ for each a, resulting in output cells uncorrelated within each scale and between scales.

To construct $\mathbf{U}^a$, one notices that the requirement of translation invariance is equivalent to having identical receptive fields, except for a spatial shift of the centers, within each scale a. It forces $\mathbf{U}^a_{nj} \propto e^{if_j x^a_n}$. For a general λ, it turns out that the constraint $\mathbf{U}^a(\mathbf{U}^a)^\dagger = 1$ for $a > 0$ cannot be satisfied if one insists on only one cell or receptive field type within the scale. However if one allows the existence of several, say 'q', cell types within the scale, $\mathbf{U}^a(\mathbf{U}^a)^\dagger = 1$ is again possible. In this case, each cell is identical to (or is the off-cell type of) the one that is q lattice spaces away in the same scale lattice (i.e. $x^a_n \to x^a_{n+q}$). The most general choice for real receptive fields is then

$$U^a_{nj} = \begin{cases} \frac{1}{\sqrt{N^a}} e^{i(-\phi^a\ n + f_j x^a_n + \theta)} & \text{if } f_j > 0 \\ \\ \frac{1}{\sqrt{N^a}} e^{-i(-\phi^a\ n + |f_j| x^a_n + \theta)} & \text{if } f_j < 0 \end{cases} \qquad (13)$$

where θ is an arbitrary phase which can be thought of as zero for simplicity at the moment, and

$$\phi^a = \frac{p}{q}\pi, \tag{14}$$

for two relatively prime integers p and q. This means the number of cell types in any given scaling block will be q. The frequencies sampled by this cell group are $f_j = \pm\frac{2\pi}{N}j$ for $j^a < j \leq j^{a+1}$. Including both the positive and the negative frequencies, the total number of frequencies sampled, and, since U^a is a square matrix, the total number of cells in this scale, is $N^a = 2(j^{a+1} - j^a)$.

The constraint of unitarity for $a > 0$ leads to the equation

$$\sum_{j=2j^a+1}^{2j^{a+1}} U^a_{nj}(U^a_{n'j})^* = \sum_{j=j^a+1}^{j^{a+1}} e^{i\Delta n(\frac{2\pi}{N^a}j - \phi^a)} + c.c. = 0, \tag{15}$$

whose solution is

$$\phi^a = \frac{2j^a + 1}{j^{a+1} - j^a} \cdot \frac{\pi}{2}. \tag{16}$$

The condition $\phi^a = \frac{p}{q}\pi$ then leads to the non-trivial consequence

$$j^{a+1} = \frac{(q+p)}{p}j^a + \frac{q}{2p}. \tag{17}$$

In a discrete system, the only acceptable solutions are those where $q/2p$ is an integer. For example the choice of $q = 2$ and $p = 1$ leads to the scaling $j^{a+1} = 3j^a + 1$. This is the most interesting solution as discussed below. Mathematically speaking, in the continuum limit a large class of solutions exists, since in that limit one takes $j^a \to \infty$ and $N \to \infty$ such that $f^a = (2\pi/N)j^a$ remains finite, then we are simply lead to $f^{a+1} = f^a(q+p)/p$ for any q and p. Thus representations of the scaling and translation group are possible for all rational scaling factors $\lambda = (q+p)/p$. The bandwidth, B_{oct}, of the corresponding cells is $\log_2[(q+p)/p]$.

Interesting consequences follow from the relationship between cell bandwidth and diversity:

$$\begin{aligned} B_{oct} &= \log_2[\frac{q+p}{p}] \\ \text{Cell types} &= q \end{aligned}$$

For example a bandwidth of one octave or a scaling factor $(q+p)/p = 2$ needs only one cell type in each scale, when $q = p = 1$. If it turns out to be necessary to have B_{oct} greater than 1 octave, then at least two classes of cells are needed to faithfully represent information in each scale, with $q = 2$ and $p = 1$ giving scaling factor of 3 or B_{oct} close to 1.6 octaves.

It is interesting to compare our solutions to the so called "wavelets" which, constructed in the mathematical literature, also form representations of the translation and scaling group. In the standard construction of Grossman and Morlet (1984) and Meyer (1985), the representations could be made orthonormal ($i.e.$ unitary in the case of real matrices) only for limited choice of scaling factors given by $1+1/m$ where $m \geq 1$ is an integer.

Such constructions need only one filter type in each scale and give scale factors no larger than 2 (equivalently the largest bandwidth is 1 octave — e.g., the well-known Haar basis wavelets (Daubechies 1988)). This agrees with what we derived above for the special case of $q = 1$ where $B_{oct} = \log_2(1 + 1/p)$. However, allowing $q > 1$ gives more bandwidth choices in our construction. For example, $q = 2$ gives $B_{oct} = \log_2(1 + 2/p)$, however no larger than 1.6 octaves, and $q = 3$ gives $\log_2(1 + 3/p)$, no larger than 2 octaves, etc. These results also agree with the recent theorem of Auscher (1992) who proved that multiscale representations can exist for scalings by any rational number k/l, provided $k - l$ filter types are allowed in each scale. Our conclusion above yields exactly the same result by redefining $k = p + q$ and $l = q$. We arrived at our conclusion independently through the explicit construction presented above.[9]

The connection between the number of cell types and the bandwidth that is possible to achieve is significant. We believe the bandwidth needed by cortical cells is determined by properties of natural images. Its value should be the best compromise between planar and depth resolution preservation for the distribution of structures in natural scenes. Actually, Field (1987, 1989) examined the issue of best bandwidth for filters that modelled cortical cells and found that bandwidths between 1 and 2 octaves best matched natural scene structures. Our results here show that cortical cells cannot achieve bandwidths more than one octave without having more than one cell type.

Next we show what the predicted cell kernels look like. For generality, we give the expression for the kernels in the continuum limit for any scale factor $\lambda = \frac{q+p}{p}$ or equivalently with any allowed bandwidth — although the ones we think are most relevant to the cortex are the discrete $p = 1, q = 2$ kernels. The cell kernels are given by $\{K^a(x_n^a - x), a > 0\}$ and $\{K^0(x_n^0 - x)\}$. For any given $a > 0$, the kernels sample the frequency in the range $f \in (f^a, \lambda f^a) = (f^a, f^{a+1})$. For $a = 0$, K^0 samples only frequencies $f \in (0, f^1)$, and $\mathbf{U}^0$ is given by $\mathbf{U}^0 = \mathbf{M}^\dagger$ in eqn. 9 with N replaced by N^0. Including both the positive and negative frequencies the predicted kernels are

$$\begin{aligned} K^a(x_n^a - x) &= \frac{1}{\sqrt{N^a}} \int_{f^a}^{f^{a+1}} df\, \sqrt{f}\, e^{i(f(x_n^a - x) + \frac{p}{q}\pi n + \theta)} + c.c. \\ &= \frac{2}{\sqrt{N^a}} \int_{f^a}^{f^{a+1}} df\, \sqrt{f}\, cos(f(x_n^a - x) + \frac{p}{q}\pi n + \theta) \qquad (18) \\ K^0(x_n^a - x) &= \frac{2}{\sqrt{N^0}} \int_0^{f^1} df \sqrt{f}\, cos(f(x_n^a - x)). \qquad (19) \end{aligned}$$

For any given p and q the kernels for $a > 0$ come in q varieties. Even and odd varieties are immediately apparent when one sets $q = 2$, $p = 1$, and $\theta = 0$ ($K^a(x_n^a - x)$ are even or odd functions of $x_n^a - x$ for even or odd n). In Fig. 2 we exhibit the even and odd kernels in two adjacent scales and their spectra. The $a = 0$ kernels, where $\theta = 0$ is chosen, are similar to the center-surround retinal ganglion cells (however they are larger in size), and hence we need not to exhibit them here. In general, though, θ can take any value, and

[9] We thank Ingrid Daubechies for pointing out the result of P. Auscher to us.

the neighboring cells will simply differ by a 90^o phase shift, or in quadrature, without necessarily having even or odd symmetry in their receptive field shapes.

From (18), it is easy to show that the kernels for $a > 0$ satisfy the following recursive relations

$$K^a(x_n^a - \lambda x) = \frac{1}{\lambda} K^{a+1}(x_n^{a+1} - x) \quad (20)$$

$$K^a(x_n^a - (x + x_q^a)) = K^a(x_{n-q}^a - x) \quad (21)$$

To prove these one needs to use the following facts, $f^{a+1} = \lambda f^a$, $N^{a+1} = \lambda N^a$, and $\lambda x_n^{a+1} = x_{\lambda n}^{a+1} = x_n^a$. ((21) also applies for K^0.) The above relations imply that, except shifted in space, each cell has the same receptive field as its q^{th} neighbor within the same scale block, $e.g.$, when $q = 2$ in the example above, all the even (or odd) cells are identical. Furthermore, except for the lowest scale $a = 0$, the n^{th} cell in all scales has the same receptive field except for a factor of λ expansion in size and a λ reduction in amplitude. Actually, since $x_n^a \neq x_n^{a+1}$, these cells are located at different spatial locations.

Now it is straight forward to see that the translation invariance (11) for $\delta n = q$ and scale invariance (12) are the direct consequence of the translation and scaling relationships (21) and (20), respectively, between the receptive fields. This is exactly our goal of object constancy. Notice that the scaling constancy would not have been possible if the whitening factor $\sqrt{f}$ was not there in eqn. (18). These results can be extended to $2D$ where the whitening factor is $1/\sqrt{R(\underline{f})} = |\underline{f}|$ as we will see next.

4. Extension to $2D$ and color vision: oriented filters and color opponent cells

The extension to two dimensions of the above construction is not difficult but involves a new subtlety. In this case, the constraint of unitarity on the matrices $\mathbf{U}^a, a > 0$ is hard to satisfy even if we allow for the phase factor ϕ which lead ultimately to different classes of cells. This constraint is considered in more detail in the appendix, here we only state the conclusions of that analysis.

What one finds is that to ensure unitarity of $\mathbf{U}^a$, one needs to allow for cell diversity of a different kind — cells in the a'th scale need to be further broken down into different types or orientations, each sampling from a limited region of the frequency space in that scale. Three examples of acceptable unitary breakings are shown in Fig. $3A$, B, C. In A (B) filters are broken into two classes in any scale $a > 0$ — in addition to the q-cell diversity discussed in $1D$. One filter type is a lowpass-bandpass in the x-y direction and the other is a bandpass-lowpass in the x-y direction which are denoted by 'lb' and 'bl'. In C there are three classes of filters, 'lb', 'bl' and finally a class of filters which are bandpass in both x and y, 'bb'. The 'lb' and 'bl' filters are oriented while the 'bb' ones are not [10].

[10]One notices that this extension to 2D requires a choice of orientations such as the x-y axes, breaking the rotational symmetry. Furthermore, it is natural to ask if the object constancy by translations and scalings should be extended to the object rotations in the image plane — requiring the cells be representations of the rotation group. At this point, it is not clear whether the rotational invariance is necessary (noting that we usually tilt our heads to read a tilted book or fail to recognize a face upside down), and whether the rotational invariance can be incorporated simultanously with the translation and scaling ones without increasing the number of cells. We will leave this outside the paper.

Figs. 4, A and B, show the five cell types one encounters for the breaking in $3B$ and the nine cell types for the breaking in Fig. $3C$, respectively, for a choice of scaling factor 3.

Finally, the object constancy eqns. (11), (12) still hold since (20) and (21) extend to $2D$ as

$$\begin{aligned} K^a(\underline{x}^a_{\underline{n}} - \lambda \underline{x}) &= \frac{1}{\lambda^2} K^{a+1}(\underline{x}^{a+1}_n - \underline{x}), \\ K^a(\underline{x}^a_{\underline{n}} - (\underline{x} + \underline{x}^a_{\underline{q}})) &= K^a(\underline{x}^a_{\underline{n}-\underline{q}} - \underline{x}). \end{aligned}$$

These relationships are understood to hold between cells belonging to the same frequency sampling category ('lb', 'bl', or 'bb'). The factor of $1/\lambda^2$, comes because the whitening factor in $2D$ is $1/\sqrt{R(|\underline{f}|)} = |\underline{f}|$.

From (18) and (19), it is clear that the cortical kernels $K^a(x) \propto \int_{f_a}^{f_{a+1}} df(1/\sqrt{R(f)})cos(fx+\phi_a)$ differ from the retinal kernel $K(x) \propto \int_0^{f_{max}} df(1/\sqrt{R(f)})cos(fx+\phi)$ only by the range of the frequency integration or selectivity. The cortical receptive fields are lowpass or bandpass versions of the retinal ones. One immediate consequence of this is that most cortical cells, especially the lowpass ones like those in the Cytochrome Oxidase Blob cells, have larger receptive fields than the retinal ones. Second, when considering color vision, the power spectrums $R_l(f)$ and $R_c(f)$ for the luminance and chrominance channels respectively, differ in their magnitudes. In reality when noises are considered, the receptive field filters are not simply $1/\sqrt{R(f)}$, which would have simply resulted in identical receptive field forms for luminance and chrominance except for their different strengths, but instead, the filter for luminance is more of a bandpass and the filter for chrominance a relatively lowpass. Since the retinal cells carry luminance and chrominance information simultanously by multiplexing the signals from both channels, the resulting retinal cells are of red-center-green-surround (or green-center-red-surround) types (Atick et al 1992). This is because at low spatial frequencies, the chrominance filter dominates, while at higher spatial frequencies, the luminance one dominates. As we argued above, the cortical cells simply lowpass or bandpass the signals from the retinal cells, thus the lowpass version will carry mostly the chrominance signals while the bandpass or highpass ones the luminance signals. This is indeed observed in the cortex (Livingstone and Hubel 1984, Ts'o and Gilbert 1988) where the large (lowpass) blob cells are more color selective, while the smaller (higher-pass) non-blob cells, which are also more orientation selective by our results above, are less color sensitive. Furthermore, since the luminance signals are negiligible at low frequencies, hence when one only considers the linear cell properties, the color sensitive blob cells are double-opponent (e.g. red-excitatory-green-inhibitory center and the red-inhibitory-green-excitatory surround) or color-opponent-center-only (type II), depending on the noise levels. This is apparent when one tries to spatially lowpass the signals from a group of single-opponent retinal cells (fig. 5).

5. Discussion: Comparison with other work

The types of cells that we arrive at in constructing unitary representations of the translation and scaling group (see Figs. $2, 4$) are similar to simple cells in cat and monkey striate cortex. The analysis also predicts an interesting relationship between bandwidths of cells and their diversity as was discussed in sections three and four. One consequence of that relationship is that for cells to achieve a representation of the world with sampling bandwidth between 1 and 2 octaves there must be at least two cell types adjacent to each other and differ by 90^o in their receptive field phases (Fig. 2). This bandwidth range is the range of measured bandwidths of simple cells ($e.g.$ Kulikowski and Bishop 1981; Andrews and Pollen 1979) and also, we think, is best suited for matching structures in natural scenes (cf. Field 1987,1989). This analysis thus explains the presence of phase quadrature (e.g., paired even-odd simple cells) observed in the cortex, (Pollen and Ronner 1981): such cell diversity are needed to build a faithful multiscale representation of the visual world.

The analysis also requires breaking orientation symmetry. Here we do not wish to advocate scaling symmetry as an explanation for the existence of oriented cells in the cortex. It may be that orientation symmetry is broken for a more fundamental reason and that scaling symmetry takes advantage of that. Either way, orientation symmetry breaking is an important ingredient in building these multiscale representations.

In the past, there has been a sizeable body of work on trying to model simple cells in terms of "Gabor" and "log Gabor" filters (Kulikowski et al 1982; Daugman 1985, Field 1987, 1989). Such filters are qualitatively close to those derived here, and they describe some of the properties of simple cells well. Our work differs from previous work in many ways. The two most important differences are the following. First, the filters here are derived by unitary transforms on retinal filters which reduce redundancy in inputs by whitening. By selecting the unitary transformation that manifests spatial-scale relationships in signals, one arrives at a representation that exhibits object constancy — the output response to an input $S(x)$ and its planar and depth translated version (i.e., $S(x) \rightarrow S(\lambda(x + \delta x))$ are related by

$$O^a_n[S(x)] = O^{a+1}_{n+\delta n}[S(\lambda(x + \delta x))]. \tag{22}$$

Hence a visual object moved in space simply shifts the outputs from one group of cells to another. Second, we find a direct linkage between cell bandwidth and diversity. Such linkage does not appear in previous works where orthonormality or unitarity was not required.

More recently there has also been a lot of work on orthonormal multiscale representations of the scaling and the translation group, alternatively known as wavelets (Meyer 1985, Daubechies 1988, Mallat 1989). The relationship of our work to wavelets was discussed in section three. Here we should add that in this paper we provide explicit construction of these representations for any rational scaling factor. Furthermore, our filters satisfy $K^a(\lambda x) = \frac{1}{\lambda^d} K^{a+1}(x)$ where d is the dimension of the space, e.g., $d = 1$ or 2, while those in the wavelet construction satisfy $K^a(\lambda x) = \frac{1}{\lambda^{d/2}} K^{a+1}(x)$. This difference stems from the fact that our filters are the convolution of the whitening filter *and* the standard-type wavelet. The whitening filter — given by $\sim 1/\sqrt{R(f)}$ where $R(f)$ is the scale invariant power spectrum of natural scenes — is what ultimately leads to the object constancy property which is absent from the standard-type wavelets.

The question at this stage is whether we could identify the pieces in our mathematical construction with classes of cells in the cortex. First, there is the class of lowpass cells $a = 0$, which have large receptive fields, and no orientation tuning (actually since their kernels have a whitening factor, they are not completely lowpass but an incomplete bandpass–weak surround). We think a good candidate for these cells are the cells in the Cytochrome Oxidase Blob areas in the cortex. When we add color to our analysis, this class will come out to be color opponent[11]. These cells, a lowpass version of the single opponent retinal cells, turn out to be double opponent or color-opponent-center-only (see Fig. 5) from this mathematical construction, in agreement with observations. Second, the representation requires several orientation classes in every choice of higher scale, they are not as likely to be color selective and, within each orientation and scale, there are two types of cells — in phase quadrature (e.g., even and odd symmetric) — if the bandwidth of the cells is greater than one octave. These have kernels similar to simple cells'. Also, in some choices of division of the two dimensional frequency space into bands (see Fig. 4) one encounters cells that are very different from simple cells. These cells come from the bandpass region in both the x and y directions (the 'bb' region in Fig. $3C$) and as such possess relatively small receptive fields in space. It is amusing to note their resemblance to the type of cells that Van Essen discovered in $V4$ (private communication).

It is important at this stage to look in detail for evidence that cortical neurons are building a multiscale, translationally invariant representation of the input along the lines described in this paper. However, in looking for those we must allow for the possibility that these representations are formed in an active process starting as early as the striate cortex, as was proposed recently by (Olshausen et al 1992). We also must keep in mind that to perform detailed comparison with real cortical filters, our filters have to modified to take noise into account.

Acknowledgments We would like to thank D. Field, C. Gilbert and N. Redlich for useful discussions, and the Seaver Institute for its support.

Appendix

In this appendix we examine the condition of unitarity on the matrix $\mathbf{U}^a$. The matrix elements of $\mathbf{U}^a$ in the scale $a > 0$ are generalized from the $1D$ case simply as

$$U^a_{\underline{\mathrm{n}}\underline{\mathrm{j}}} = e^{i(\underline{\phi}\underline{\mathrm{n}} + \underline{\mathrm{f}}_j \underline{\mathrm{x}}^a_n)} \tag{23}$$

where $\underline{\mathrm{n}} = (n_x, n_y)$, $\underline{\mathrm{x}}^a_n = (x^a_{n_x}, x^a_{n_y})$, $\underline{\mathrm{j}} = (j_x, j_y)$, $\underline{\mathrm{f}}_j = (f_{j_x}, f_{j_y})$, and $\underline{\phi} = (\frac{f_{jx}}{|f_{jx}|}\phi_x, \frac{f_{jy}}{|f_{jy}|}\phi_y)$. A priori the cells in $\mathbf{U}^a$ sample from the frequency region inside the big solid box but outside the dashed box in Fig. 3. The critical fact that makes the $2D$ case different from $1D$ is that there are $(N^a)^2 = 4(j^{a+1})^2 - 4(j^a)^2$ cells in the a'th class, while the total number of cells is $(N)^2$, then $(x^a_{n_x}, x^a_{n_y}) = (\frac{N}{N^a} n_x, \frac{N}{N^a} n_y)$.

[11]It is easy to see why: since they are roughly lowpass – large receptive fields – they have higher signal to noise in space and hence they can afford to have a low signal to noise in color. While opponent cells in space have low signal to noise and hence they need to integrate in color to improve their signal to noise (see Atick et al 1992)

The unitarity requirement $\mathbf{U}^a(\mathbf{U}^a)^\dagger = 1$ $(a > 0)$ can be shown to be equivalent to

$$\cos\left[(\frac{j^{a+1}+j^a+1}{N^a}\pi + \phi_x)\Delta n_x\right]\sin\left[\frac{j^{a+1}-j^a}{N^a}\pi\Delta n_x\right] = 0 \tag{24}$$

where Δn_x is any integer $\neq 0$. A similar condition in the y direction should also hold. To satisfy (24) one can only hope that the cosine factor is zero for odd Δn_x and the sine factor is zero for the rest. This is impossible in $2D$ although possible in $1D$. To see this difference, note that in $1D$, $N^a = 2(j^{a+1} - j^a)$ and the argument of the sine is $\Delta n\pi/2$ which leads to vanishing sine for even Δn. One then makes cosine term zero for odd Δn by choosing ϕ such that $\frac{j^{a+1}+j^a+1}{N^a}\pi + \phi_x = \pm\frac{\pi}{2}$. This is exactly how equation (16) is reached. In $2D$, $N^a = 2\sqrt{(j^{a+1})^2 - (j^a)^2}$, and hence the sine term is $\sin[\Delta n_x(\sqrt{(j^{a+1}-j^a)/(j^{a+1}+j^a})\pi/2] \neq 0$ for even Δn_x. Although we cannot prove that the negative result in $2D$ is not caused by the fact that we have a Euclidean grid, we think it not possible to construct the representation even when using a radially symmetric lattice.

To ensure unitarity of $\mathbf{U}^a$, we need to allow for cell diversity of a different kind — cells in a^{th} scale need to be further broken down into different types or orientations, each type sampling from, a limited region of the frequency space as shown for example in Fig. 3.

References

[1] Andrews, B. W. and Pollen, D. A. 1979. Relationship between spatial frequency selectivity and receptive field profile of simple cells. *J. Physiol.* (London), **287**, 163–176.

[2] Atick, J. J. 1992. Could Information theory provide an ecological theory of sensory processing? *Network* **3**, 213–251.

[3] Atick, J. J. and Redlich, A. N. 1990. Towards a theory of early visual processing. *Neural Comp.*, **2**, 308–320.

[4] Atick, J. J. and Redlich, A. N. 1992. What does the retina know about natural scenes? *Neural Comp.*, **4**, 196–210.

[5] Atick, J. J., Li, Z. and Redlich, A. N. 1992. Understanding retinal color coding from first principles. *Neural Comp.*, **4**, 559–572.

[6] Auscher, P. 1992. Wavelet bases for $L^2(R)$ with rational dilation factor. In *Wavelets and their applications*, pp. 439–451, ed. Ruskai, M B. Jones and Bartlett, Boston.

[7] Daubechies, I. 1988. Orthonormal bases of compactly supported waves. *Commun. Pure Appl. Math.*, vol. **41**, 909–996.

[8] Daugman, J. G. 1985. Uncertainty relations for resolution in space, spatial frequency and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am.* **A 2**, 1160–1169.

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

[10] Field, D. J. 1989. What the statistics of natural images tell us about visual coding. SPIE Vol. **1077** Human Vision, Visual Processing, and Digital Display, 269–276.

[11] Grossmann, A. and Morlet, J. 1984. Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM J. Math.* vol. **15**, pp. 17–34.

[12] Hopfield, J. J. 1991. Olfactory computation and object perception. *Proc. Natl. Acad. Sci.* USA, **88**, 6462–6466.

[13] Kulikowski, J. J., Marcelja, S. and Bishop, P. 1982. Theory of spatial position and spatial frequency relations in the receptive fields of simple cells in the visual cortex. *Biol. Cybern.*, **43**, 187–198.

[14] Kulikowski, J. J. and Bishop, P. 1981. Linear analysis of the responses of simple cells in the cat visual cortex. *Exp. Brain Res.* **44**, 386–400.

[15] Linsker, R. 1992. Private communication. See also, talk at NIPS 92.

[16] Livingstone M. S. and Hubel D. H. 1984 Anatomy and physiology of a color system in the primate visual cortex. *J. of Neurosci.* Vol. 4, No.1. 309-356.

[17] Mallat, S. 1989. A theory of multiresolution signal decomposition: The wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, vol. **11**, 674–693.

[18] Meyer, Y. 1985. Principe d'incertitude, bases hilbertiennes et algebres d'operateurs. *Seminaire Bourbaki*, nr. **662**, 209–223.

[19] Olshausen, B., Anderson, C. H. and Van Essen D. C. 1992. A Neural model of visual attention and invariant pattern recognition. Caltech Report no. CNS MEMO 18, August.

[20] Pollen, D. A. and Ronner, S. F. 1981. Phase relationships between adjacent simple cells in the cat. *Science* **212**, 1409–1411.

[21] Ruderman, D. L. 1992. private communication and to appear.

[22] Schreiber, W. F. 1956. The measurement of third order probability distributions of television signals. IRE *Trans. Inform. Theory* **IT-2** pp. 94–105.

[23] Shannon, C. E. and Weaver, W. 1949. *The mathematical theory of communication*. (Urbana, Il; University of Illinois Press).

[24] Ts's D. Y. and Gilbert C. D. 1988 The organization of chromatic and spatial interactions in the primate striate cortex. *J. of Neurosci.* 8(5):1712-1727

[25] Van Essen, D. C., Olshausen B., Anderson C. H and Gallant, J. L. 1991. Pattern recognition, attention, and information bottlenecks in the primate visual system *Conf. on Visual Information Processing: From Neurons to Chips (SPIE Proc. 1473)*

Figures

A

B

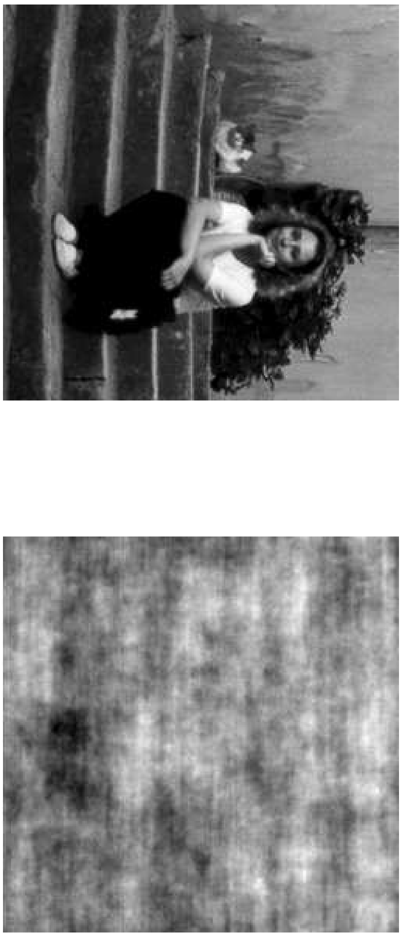

Figure 1: A, B: Demonstration of the uselessness of second-order statistics for form definition and discrimination. Following (Field 1989), image B is constructed by first fourier transforming A, randomizing the phases of the coefficients and then taking the inverse fourier transform. The two images thus have the same second-order statistics but B has no higher-order ones. All relevant object features disappeared from B.

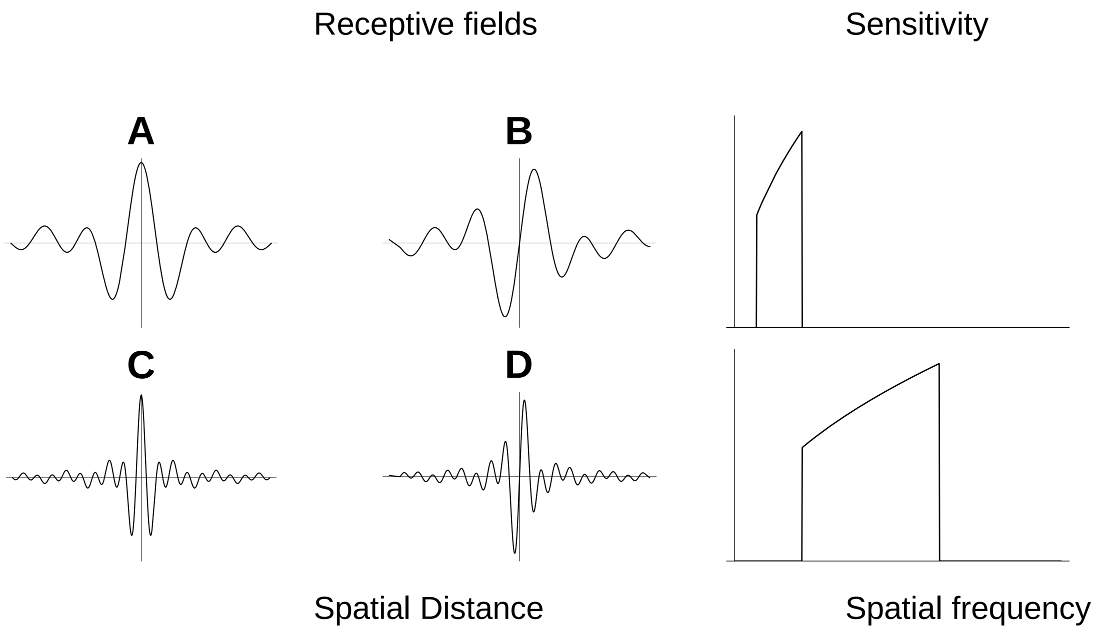

Figure 2: Even-symmetric”, A, C, and “odd-symmetric”, B, D, kernels predicted for the scale factor 3 (equivalently for $B_{oct} = 1.6$ octaves) for two neighboring scales (top and bottom rows, respectively), together with their spectra (frequency sensitivities or selectivities).

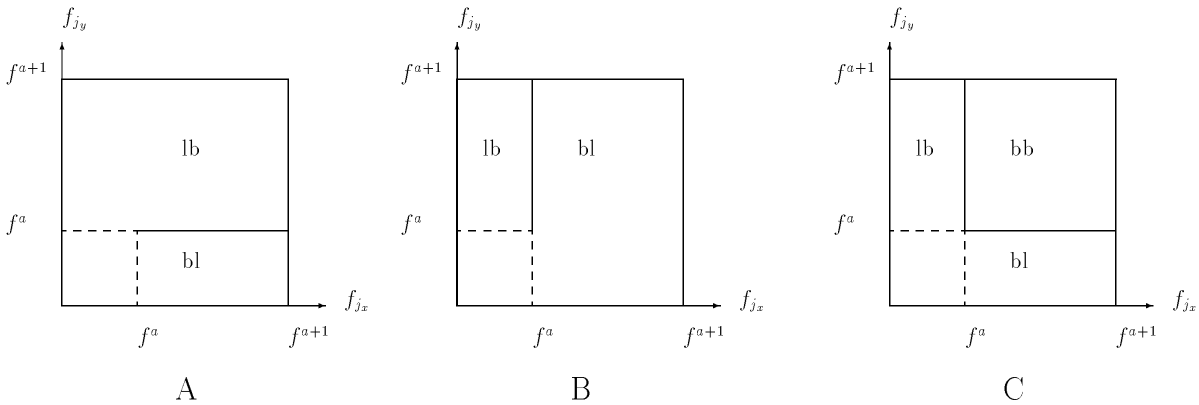

Figure 3: $A - C$: Proliferation of more cell types by the break-down of the frequency sampling region in $2D$ within a given scale a. Ignoring the negative frequencies, the frequencies f within the scale are inside the large solid box but outside the small dashed box. The solid lines within the large solid box further partition the sampling into subregions denoted by 'bl', 'lb', and 'bb', which indicate bandpass-lowpass, lowpass-bandpass, and bandpass-bandpass, respectively, in x-y directions. A and B give asymmetric break down between x and y directions, the 'lb' cells are not equivalent to a 90^o rotation of the 'bl' cells. C gives symmetric break-down between x and y directions. The 'bb' cells are significantly different from the others, see Fig. 4.

1

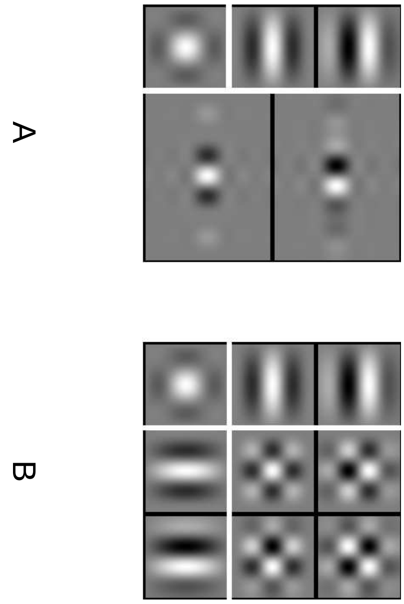

Figure 4: Fig. 4 A, B: The predicted variety of cell receptive fields in $2D$. The five cell types in A and the nine cell types in B arise from the frequency partitioning schemes in Fig. $3B$ and Fig. $3C$, respectively. The kernels in the lower-left corner of both images demonstrate the lowpass-lowpass filter K^0 in $2D$ and they are non-oriented. All others are bandpass in at least one direction. Those are actually significantly smaller but are expanded in size in this figure for demonstration. The 'bb' cells in the upper-right part of B come in four varieties (even-even, odd-odd, even-odd and odd-even when $\theta = 0$ is taken for both x and y directions) and should exist in the cortex if the scheme in Fig. $3C$ is favored. All kernels are constructed taking into account the optical MTF of the eye.

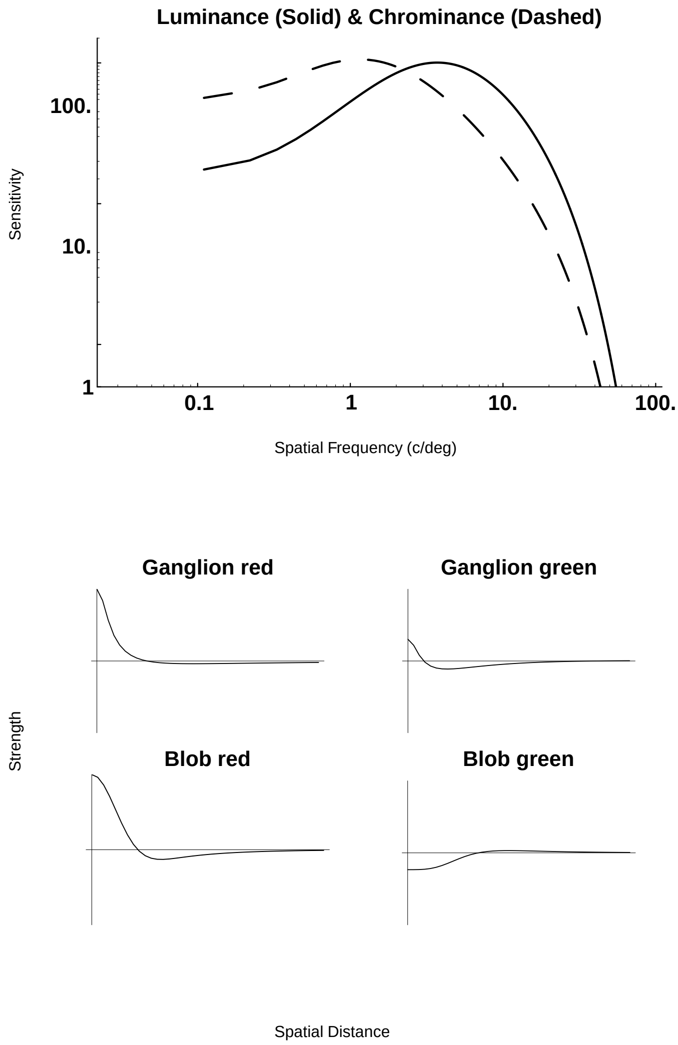

Figure 5: Change of color coding from retina to cortex. The top plot shows the visual contrast sensitivities to the luminance and chrominance signals. The bottom plot demonstrates the receptive field profiles (sensitivity to red or green cone inputs) of the color selective cells in the retina (or ganglion) and the cortex. The parameters used for the ganglion cells are the same as those in Atick et al 1992. The blob cells are constructed by lowpass filtering the ganglion cell outputs with a filter frequency sensitivity of $e^{-f^2/(2f_{low}^2)}$ where $f_{low} = 1.5$ c/deg. The strengths of the cell profiles are individually normalized for both the ganglion and the blob cells. The range of the spatial distance axes, or the size, of the blob cells is 3.7 times larger than that of ganglion cells. This means that each blob cell sums the outputs from (on the order of) at least about $(3.7)^2 \sim 16$ local ganglion cells.