OPTIMAL SENSORY ENCODING

Published in *The Handbook of Brain Theory and Neural Networks: The Second Edition* page 815-819.
Editor: Michael A. Arbib. MIT Press, 2002.

L. Zhaoping
Department of Psychology
University College London, U.K.
z.li@ucl.ac.uk

INTRODUCTION

What is optimal depends on computational tasks. Many recent works base optimality on information theoretical terms such as information transmission rates. This can be particularly revelant in the early stages of vision which are mainly concerned with transmitting information indiscriminately. We focus on the better known visual system to discuss optimal sensory coding, although coding in other sensory systems are expected to address similar concerns.

Consider a simplified visual input model, with, say, 1000x1000 pixels arranged in a regular grid at one byte per pixel and 20 images per second. It provides many megabytes/second of raw data. Given the information bottleneck in the long optic nerve from retina to thalamus and the limited firing rates (thus limited data capacity) of cortical neurons (see article SENSORY CODING AND INFORMATION TRANSMISSION), early vision can greatly benefit from a data encoding that reduces the data rate without significant information loss. Since nearby image pixels tend to convey similar signals (e.g., luminance values) and thus carry redundant information, significant savings can be made by avoiding transmitting the information redundantly. If, within a particular time window, each original pixel codes one byte of information, 80% of which is redundant information shared with neighboring pixels, then one million pixels code only 200 Kbytes of non-redundant information. One way to avoid redundancy is to transform the original signal $\mathbf{S} = \{S_1, S_2, ..., S_N\}$ in the N neurons (e.g., photoreceptors) to signals $\mathbf{O} = \{O_1, O_2, ..., O_M\}$ in another M (more/fewer) neurons (e.g., the retinal ganglion cells or cortical neurons), such that signals in O_i and O_j for all i, j are not significantly redundant. Consequently, 200 Kbytes of information in $\mathbf{S}$ could be coded by only 0.2 byte in each neuron O_i if $M = N$, which needs a much reduced firing rate. Loss-less encoding means that, if needed, $\mathbf{S}$ can be reconstructed from $\mathbf{O}$. Such observations have led to the "infomax" proposal that early vision constructs an "optimal coding" of input to allow maximum information transmission from retina to cortex under limited channel capacity of the optic nerve or neural activities (Attneave 1954, Barlow 1961, Linsker 1990, and Atick 1992). This principle has provided many insights in the properties of the receptive fields (RFs) in early vision.

OPTIMAL CODING ILLUSTRATED BY STEREO VISION

Consider the redundancy and encoding of stereo signals (Li and Atick 1994a). Let S_L and S_R be the signals to the left and right eyes (Fig. (1)). They may be the average luminance in the images, or the Fourier components (of a particular frequency) of the images. Assume that they have zero mean

Figure 1: A stereo pair input to the two eyes.

(for simplicity) and equal variance (or signal power) $< S_L^2 >=< S_R^2 >$ ($< .. >$ denotes average over the input ensemble). The redundancy is seen in the correlation matrix:

$$R^S \equiv \begin{pmatrix} < S_L^2 > & < S_L S_R > \\ < S_R S_L > & < S_R^2 > \end{pmatrix} =< S_L^2 > \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$$

where $0 \leq r \leq 1$ is the correlation coefficient between S_L and S_R. The value of r is high, $r \rightarrow 1$, for mean luminance signals $S_{L,R}$ but low, $r \rightarrow 0$, if $S_{L,R}$ are a high spatial frequency Fourier component of the respective images. A simplifying assumption is that $\mathbf{S}$ are Gaussian signals, which are defined as to have a probability distribution $P(\mathbf{S}) \propto exp(-\sum_{ij} S_i S_j (R^S)^{-1}_{ij}/2))$. An encoding

$$O_+ = S_+ \equiv (S_L + S_R)/\sqrt{2}, \qquad O_- = S_- \equiv (S_L - S_R)/\sqrt{2}$$

gives zero correlation $< O_+ O_- >$ in $\mathbf{O}$, leaving output probability $P(\mathbf{O}) = \Pi_i P(O_i)$ factorized, as easily verified. The transform $\mathbf{S} \rightarrow \mathbf{O}$ is linear, which approximates the cell response properties in the retina and, to a less degree, in primary visual cortex. The cell coding O_+ is a binocular cell due to the binocular summation of inputs, while the cell coding O_- is monocular or ocularly opponent. Note that $S_\pm$ are the eigenvectors of the correlation matrix R^S or the principal components of the signals, and their signal power $< S_\pm^2 >= (1 \pm r) < S_L^2 >$ are the corresponding eigenvalues. In reality, input noise $\mathbf{N}$ is added on $\mathbf{S}$ and the coding transform introduces additional noise $\mathbf{N_o}$, hence, $O_\pm = [(S_L + N_L) \pm (S_R + N_R)]/\sqrt{2} + N_{o,\pm}$, giving effective output noise $N_\pm = (N_L \pm N_R)/\sqrt{2} + N_{o,\pm}$. For simplicity, the noise terms are assumed to be independent of each other and of the signals. Let $< N^2 >\equiv< N_L^2 >=< N_R^2 >$, and $< N_o^2 >\equiv< N_{o,+}^2 >=< N_{o,-}^2 >$. Input $S_{L,R} + N_{L,R}$ has

$$I_{L,R} = \frac{1}{2} \log_2 \frac{< S_{L,R}^2 > + < N^2 >}{< N^2 >}$$

bits of (mutual) information about $S_{L,R}$, since, for Gaussian signals and noise, the information amount is $\frac{1}{2}\log_2$(signal-to-noise), whereas while $O_\pm$ has

$$I_\pm = \frac{1}{2} \log_2 \frac{< O_\pm^2 >}{< N_\pm^2 >} = \frac{1}{2} \log_2 \frac{< S_\pm^2 > + < N^2 > + < N_o^2 >}{< N^2 > + < N_o^2 >}$$

bits of information about $S_{L,R}$ or $S_\pm$. Note that the redundancy between S_L and S_R causes higher or lower signal powers $< O_+^2 >$ or $< O_-^2 >$ in O_+ or O_- respectively, leading to higher or lower information rate I_+ or I_-. As an initial choice, define cost as the total signal power, although there

can be many other cost considerations (see later). Since $I_\pm = \frac{1}{2}\log_2(< O_\pm^2 >)$ + constant = $\frac{1}{2}\log_2$ (cost) + constant, we note that the gain in information per unit cost ($\Delta I/\Delta$ cost) is smaller in the O_+ than that in the O_- channel. This motivates reduction and increment of costs in the O_+ and O_- channels respectively, by introducing the gains $V_\pm$, such that $O_\pm = V_\pm[(S_L + N_L) \pm (S_R + N_R)]/\sqrt{2} + N_{o,\pm}$, at the expense or benefit of the information transmitted

$$I_\pm = \frac{1}{2}\log_2 \frac{V_\pm^2(< S_\pm^2 > + < N^2 >) + < N_o^2 >}{V_\pm^2 < N^2 > + < N_o^2 >} \tag{1}$$

Hence, the optimal encoding, balancing the cost and information extraction, is to find the gains $V_\pm$ to minimize

$$E(V_\pm) \equiv \sum_a (< O_a^2 >) - \lambda \sum_a (I_a) = \text{cost} - \lambda \cdot \text{Information} \tag{2}$$

where λ is the Lagrange multiplier whose value determines the balance. The optimal gains can be obtained by $\partial E/\partial V_\pm = 0$ to give

$$V_\pm^2 \propto \text{Max}\left\{[\frac{1}{2}\frac{< S_\pm^2 >}{< S_\pm^2 > + < N^2 >}(1 + \sqrt{1 + \frac{4\lambda}{\log 2 < N_o^2 >}\frac{< N^2 >}{< S_\pm^2 >}}) - 1], 0\right\}. \tag{3}$$

In the zero noise limit when $\frac{<S_\pm^2>}{<N^2>} \gg 1$, $V_\pm^2 \propto < S_\pm^2 >^{-1}$. As expected, this suppresses the stronger ocular summation signal S_+ and amplifies the weaker ocular contrast signal S_-, in order to save the cost, since the cost increases linearly with $V_\pm^2$, but the extracted information increases only logarithmically with $V_\pm^2$. Hence, for instance, when the coding noise $\mathbf{N_o}$ is negligible (i.e., $\frac{<N_o^2>}{V_\pm^2 <N^2>} \ll 1$), output $\mathbf{O}$ and the original input $\mathbf{S} + \mathbf{N}$ contain about the same amount of information about the true signal $\mathbf{S}$, but $\mathbf{O}$ consumes much less power with $V_+ \ll V_- < 1$, when $r \sim 1$. This gain $V_\pm \propto < S_\pm^2 >^{-1/2}$ also equalizes output power $< O_+^2 > \approx < O_-^2 >$, since $< O_\pm^2 > = V_\pm^2 < S_\pm^2 >$ + noise power, making the output correlation matrix R^o (with elements $R_{ab}^o = < O_a O_b >$) proportional to an identity matrix (since $< O_+ O_- > = 0$). Such a transform $\mathbf{S} \to \mathbf{O}$, which leaves output channels decorrelated and equally powered, is called whitening. Any rotation $\mathbf{O} \to \mathbf{UO}$ via a rotation or unitary transform $\mathbf{U}$ ($\mathbf{UU^T} = 1$), by angle θ in the two dimensional space $\mathbf{O}$, multiplexes the channels O_+ and O_- to give two alternative channels

$$\begin{pmatrix} O_1 \\ O_2 \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} O_+ \\ O_- \end{pmatrix} = \begin{pmatrix} \cos(\theta)O_+ + \sin(\theta)O_- \\ -\sin(\theta)O_+ + \cos(\theta)O_- \end{pmatrix}.$$

which are also decorrelated ($< O_1 O_2 > = 0$). Furthermore, note from equations (2) and (1) that cost = Tr (R^o) and Information = $\frac{1}{2}\log\frac{\det R^o}{\det R^N}$, where R^N is the correlation matrix of the noises in the output channel, Tr(.) and det(.) denote the trace and determinant of a matrix. Since both the trace and determinant are invariant to unitary transforms (rotations), the optimized objective function E = (cost -λ Information) is invariant to this rotation $O_\pm \to O_{1,2}$. Hence, both encoding schemes $S_{L,R} \to O_\pm$ and $S_{L,R} \to O_{1,2}$, with former a special case of the latter, are equally optimal in making the output decorrelated (non-redundant), in extracing information about $S_{L,R}$, and in saving the coding cost $\sum_a < (O_a)^2 >$. Since

$$\begin{pmatrix} O_1 \\ O_2 \end{pmatrix} = \begin{pmatrix} S_L(\cos(\theta)V_+ + \sin(\theta)V_-) + S_R(\cos(\theta)V_+ - \sin(\theta)V_-) \\ S_L(-\sin(\theta)V_+ + \cos(\theta)V_-) + S_R(-\sin(\theta)V_+ - \cos(\theta)V_-) \end{pmatrix},$$

in general O_1 and O_2 prefer different eyes. In particular, $\theta = -45^o$ gives $O_{1,2} \propto S_L(V_+ \mp V_-) + S_R(V_+ \pm V_-)$. The visual cortex indeed has neurons of a whole spectrum of ocularities.

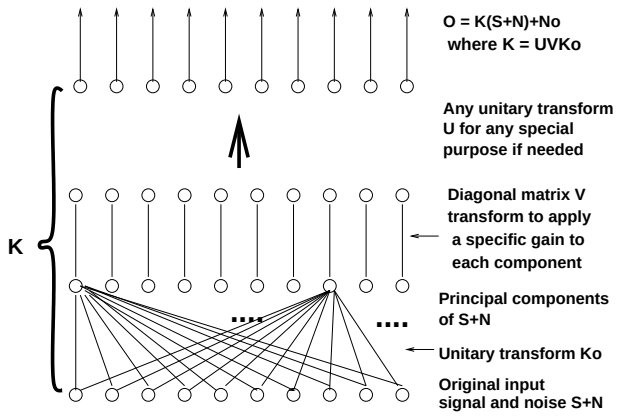

Figure 2: A schematic of the steps to obtain infomax (linear) code for Gaussian signals.

VARIATIONS OF OPTIMAL CODINGS

It is now apparent that infomax coding as defined in the Equation (2) is related to whitening, decorrelation, principal component analysis, and *factorial codes*, defined as when probabilities of signals factorizes $P(\mathbf{O}) = \Pi_a P(O_a)$. Many other relatives of optimal codings are: *minimum entropy* or *minimum description length*, since minimizing $< O_1^2 > + < O_2^2 >$ reduces the total output entropy $H(O_1) + H(O_2)$ ($H(.)$ stands for entropy) for Gaussian signals O_a, *independent components analysis* since principal components are independent components for Gaussian signals, *redundancy reduction* since the well known inequality $\sum_a H(O_a) > H(\mathbf{O})$ means that minimizing $\sum_a H(O_a)$ reduces the redundancy, intuitively defined as $\sum_a H(O_a)/H(\mathbf{O}) - 1 \geq 0$ (equal to zero when there is no redundancy), between output channels, *sparse coding* since it is defined as lowering the coding bits $H(O_a)$ for all channels a, *maximum entropy code* since $H(\mathbf{O})$ is maximized given $\sum_a H(O_a)$ when redundancy is removed, *predictive codes* since the code effectively predicts or explains away S_R from S_L to achieve minimum $\sum_a < O_a^2 >$ for given $I(\mathbf{O}; \mathbf{S})$ (information in $\mathbf{O}$ about $\mathbf{S}$), and *minimum predictability codes* or *least mutual information* between output channels since $\sum_a H(O_a) = H(\mathbf{O})$ means zero mutual information between output channels O_a and O_b. All these variations of "optimal coding" often mean approximately or exactly the same (Nadal and Parga 1997) depending on their precise definitions and the statistics of the signals concerned, and should not be thought of as independent coding principles.

OPTIMAL VISUAL CODING IN SPACE, TIME, COLOR, AND SCALE

In general, for simple linear encoding of approximately Gaussian signals $\mathbf{S}$, a recipe for optimal coding is visualized in Fig. (2). Given input signal $\mathbf{S}$ with noise $\mathbf{N}$, the encoding transform $\mathbf{K}$ and additional coding noise $\mathbf{N_o}$ gives output signal $\mathbf{O} = \mathbf{K}(\mathbf{S} + \mathbf{N}) + \mathbf{N_o}$. The optimal transform $\mathbf{K}$ is dictated by the input statistics characterized by the correlation matrix R^S. The first step is principal component analysis, transforming $\{S_a\}$, via a matrix $\mathbf{K_o}$ to the principal components $\{\mathcal{S}_k\}$, i.e., $\mathcal{S} = K_o\mathbf{S}$. The powers of the components $\mathcal{S}$ are the eigenvalues of R^S. Next, the optimal gain V_k to $\mathcal{S}_k$ is determined by $\mathcal{S}_k$'s signal-to-noise ratio via equation (3). A particular optimal coding transform

is $\mathbf{K} = \mathbf{VK_o}$, where $\mathbf{V}$ is a diagonal matrix with diagonal elements equal to the optimal gains V_k or $V(k)$. The resulting $\mathbf{O}$ have decorrelated components and retains the maximum information about $\mathbf{S}$ given output cost $\sum_a < O_a^2 >$. Furthermore, any transform in the class $\mathbf{K} = \mathbf{UVK_o}$, where $\mathbf{U}$ is any unitary transformation (rotation, $\mathbf{UU}^T = 1$), is equally optimal since it leaves the outputs $\mathbf{O}$ with the same information extraction and cost, and, in the zero noise limit, the same decorrelation. The conceptual steps above correspond mathematically to finding the (degenerate) solution $\mathbf{K}$ of $\partial E/\partial \mathbf{K} = 0$ where $E(\mathbf{K}) =$ cost - λ Information.

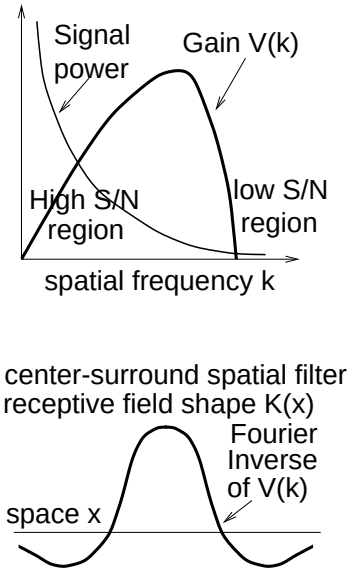

Figure 3: The contrast gain $V(k)$ as a function of spatial frequency k, determined from the signal-to-noise (S/N) of the inputs (S+N) at that frequency. The corresponding spatial filter $K(x)$ is the Fourier inverse of $V(k)$, adopted by the retinal ganglion cells on the photoreceptor inputs.

In spatial coding (Atick 1992), signal at visual location x is S_x. Since the signal correlation is translation invariant, i.e., $< S_x S_{x'} >$ is a function of only $x - x'$, the principal components are Fourier modes, and K_o is the Fourier transform $K_o^{kx} \sim e^{-ikx}$ such that $S_x \rightarrow \mathcal{S}_k \sim \sum_x K_o^{kx} S_x \sim \sum_x e^{-ikx} S_x$. Field (1987) measured the power spectrum as $< \mathcal{S}_k^2 > \sim 1/k^2$ with Fourier frequency k. Assuming white noise power $< \mathcal{N}^2 >$, the high signal-to-noise $\mathcal{S}^2/\mathcal{N}^2$ in the low k region leads to the gain V_k or $V(k) \propto k$ that increases with k. However, for high k where $\mathcal{S}^2/\mathcal{N}^2$ is low, $V(k)$ quickly decays with increasing k to zero according to equation (3) in order not to amplify noise. This gives a band-pass $V(k)$ as a function of k (Fig. (3)). If $\mathbf{U}$ is the inverse Fourier transform $U^{x'k} \sim e^{ikx'}$, then the whole transform $\mathbf{K} = \mathbf{UVK_o}$ transforms signal S_x to activities $O_{x'}$ of a neuron with a receptive field (RF) at location x' as a band-pass filter, i.e., $O_{x'} \sim \sum_k V(k) \sum_x e^{ik(x'-x)} S_x$ + noise. This is roughly what retinal output (ganglion) cells do, achieving a center-surround transform on the input image and emphasizing the intermediate frequency band where signal-to-noise is of order 1. Function $V(k)$ is the well known contrast sensitivity function. When the visual environment dims down, reducing the overall signal-to-noise $\frac{<\mathcal{S}_k^2>}{<\mathcal{N}^2>}$ in all frequencies, say from $\frac{<\mathcal{S}_k^2>}{<\mathcal{N}^2>} \sim 100/k^2$ to $\frac{<\mathcal{S}_k^2>}{<\mathcal{N}^2>} \sim 1/k^2$, the band-pass region should shift towards lower frequencies, effectively making $V(k)$ a low pass. This explains the dark adaptation of the retinal ganglion cells'

RFs, from center-surround contrast enhancing (band-pass) filter to Gaussian-like smoothing (low-pass) filter, to integrate signals and smooth out noise.

Coding in time is analogous to coding in space. Image statistics in time (Dong and Atick 1995) determine the temporal frequency sensitivities $V(\omega)$ (of frequency ω) of the optimal temporal filter. Given a sustained input $S(t)$ over time t, the output $O(t)$ may be more sustained or transient depending on whether the filter is more low pass or band pass. By an appropriate choice of the rotation transform $\mathbf{U}$ (Dong and Atick 1995 and Li 1996), the temporal filter can be made causal, i.e., the output $\mathbf{O}$ depends only on input $\mathbf{S}$ of the past but not the future.

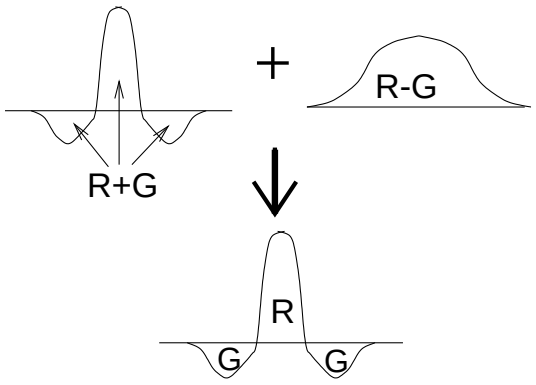

Figure 4: Multiplexing the center-surround achromatic (R+G) filter with the chromatic (R-G) gaussian-like filter gives a red-center-green-surround double (in space and in color) opponency RF observed in retina.

Visual color coding (Atick 1992) is analogous to the stereo coding. The inputs are three dimensional (3D), S_r, S_g, and S_b for red, green, and blue signals. The principal components include a strong luminance channel, a weighted summation of the cone inputs, and two weaker chrominance channels, one roughly red-green opponency and another yellow-blue opponency. Optimal coding then involves appropriate gains to these channals and additional multiplexing of them as needed. Physiologically, color and space codings are coupled, resulting for instance in the red-center-green-surround receptive fields (Fig. (4)) of the retinal ganglion cells. This can be understood in a simplified two cone system, red and green. The high signal-to-noise luminance channel ($S_r + S_g$) needs a center-surround or band pass spatial filter, while the low signal-to-noise chromatic channel ($S_r - S_g$) needs a smoothing or low pass filter. The multiplexing of these two channels, a rotational operation $\mathbf{U}$ in the 2-dimensional color space, leads to addition or subtraction of these two filters. The results are the red-center-green-surround or green-center-red-surround RFs. In the retina and/or primary visual cortex, codings in space, time, color, and stereo are all coupled together (Atick 1992, Li and Atick 1994ab, Li 1996).

MULTISCALE CODING IN THE PRIMARY VISUAL CORTEX

Primary visual cortex receives the retinal outputs via the lateral geniculate nucleus. Its RFs are orientation selective in the shape of small bars or edges. Different receptive fields have different

orientations and different sizes (or tuned to different spatial frequency bands), in a multiscale fashion such that RFs of different sizes are roughly scaled versions of each other, also called wavelet coding. These RFs can be seen as components of another optimal code by a particular choice of the rotation (unitary) matrix $\mathbf{U}$ in the coding transform $\mathbf{K} = \mathbf{U}\mathbf{V}\mathbf{K_o}$. Retinal RFs are given when $\mathbf{U} = \mathbf{K_o}^{-1}$, and are theoretically the same for all retinal ganglion cells except for a spatial translation. Another optimal code, apparently not adopted anywhere in our visual system, is when $\mathbf{U} = \mathbb{I}$, an identity matrix. The RFs would be infinitely large, each would be unique and a particular principal component (Fourier component) with a particular gain. The $\mathbf{U}$ transform for the multiscale coding is when $\mathbf{U}$ is somewhere in-between the two extremes $\mathbf{U} = \mathbf{K_o}^{-1}$ and $\mathbf{U} = \mathbb{I}$. To construct a cortical RF, $\mathbf{U}$ multiplexes the principal components (Fourier waves) within a finite frequency range $\mathbf{k} \in (\mathbf{k}_1, \mathbf{k}_2)$ such that the resulting RF is responsive only to a restricted range of orientations and spatial frequencies $\mathbf{k}$. The code can be viewed as an intermediate between the Fourier wave code, when each RF is infinitely large and responds to only one frequency and orientation, and the retinal code, where each RF is small and responsive to all frequencies k and all orientations. Different cortical units cover different ranges of frequencies to give a complete sampling (Li and Atick 1994b).

It has been argued (reviewed by Simoncelli and Olshausen 2001) that the multiscale code, which should be as good as the retinal code if the visual inputs assume gaussian statistics, is actually better in the light of the actual non-gaussian nature of the signals. Oriented RFs have been argued to capture the non-trivial 3rd order statistics, in particular, the third order correlation $< S_a S_b S_c >$ between signals from three image pixels a, b, c, which are not accounted for by the Gaussian statistics. Previous works (Simoncelli and Olshausen 2001) argued that the cortical orientation selective RFs match the orientation features in inputs, and that the neurons are inactive unless those matchs happen. The code is thus argued as a sparser code, since the activities of different cells are supposedly less correlated (see article SPARSE CODING IN THE PPRIMATE CORTEX). Why doesn't retina adopt this code? One reason could be that the cortical representation is in addition overcomplete, i.e., the number M of cortical units (output units O_a) is orders of magnitude larger than the number N of the retinal units (input units S_a). The overcompleteness has been argued to improve sparseness, though at the expense of the neural proliferation, by noting that cells tuned to different image features can not be active together. However, it should be noted that if cortical activities $\mathbf{O}$ depend linearly on visual input $\mathbf{S}$, the $\mathbf{O}$ units are necessarily (mathematically) dependent on, or correlated with, each other in an overcomplete representation where $M > N$ (Li 1996). Cortical response $\mathbf{O}$ depend on visual input $\mathbf{S}$ nonlinearly, by rectification, thresholding, saturation, and normalization etc (Simoncelli and Olshausen 2001). The observed nonlinearity is unlikely to be sufficient to achieve decorrelation. However, the nonlinearity and the overcomplete representation are more likely to serve non-trivial cognitive computations (Li, 2002) beyond the traditional coding considerations.

DISCUSSION

It is clear that maximizing information transmission alone is not enough to specify optimal codes. One may prefer one code or another when considering other costs and benefits, e.g. Levy and Baxter (1996). The retinal code has the advantage of small and identical RF shapes, involving shorter neural wiring and easier specifications. It also has stronger correlation between output signals than the Fourier wave codes outside the zero noise limit (both codes should have zero 2nd

order correlation in zero noise limit), making it easier for error correction purposes. Its translation invariance also allows an object translated laterally to induce the same pattern of neural activities except for a change in the responding neurons. When this invariance is extended to objects moving in depth (when images of objects change sizes), the cortical multiscale code is preferred. In this case many different RFs are scaled and/or translated versions of each other, leading to translation invariance within a scale and scale invariance between scales (Simoncelli et al 1992).

More significantly are optimality measures not based on information measures. For example, to give a best estimation $\hat{S}$ of input S from $O = K(S + N) + N_o$, the optimal coding transform K to minimize the estimation error $< (S - \hat{S})^2 >$ given output power $< O^2 >$ certainly does not satisfy infomax. Another example is the two classes of the retina ganglion cells. Whereas Infomax principle applies well to explain the RFs of the more numerous class of retinal ganglion cells, the P cells in monkeys or X cells in cats, another class of ganglion cells, M cells in monkeys or Y cells in cats, have RFs that are relatively larger, color unselective, and are tuned to higher temporal frequencies. These M cells do not extract the maximum information possible (infomax) about input S, but can serve to extract the information as fast as possible (Li 1992), i.e., the temporal outputs $(O(t = -\infty), ...O(t-1), O(t))$ should contain some information about $S(t' \leq t)$ with a shortest possible delay $t - t'$. This observation should have significant implications on how P and M pathways should interact at later processing stages.

Information theory provides excellent means to *quantify the amount* of information, to design optimal coding for *information transmission*. Cognitive functions often requires selections over the *quality or modality* of information, which is beyond Information Theory. Information theory is more likely to find its application in the early stages of the sensory processing, before information is selected or discriminated for any specific cognitive task, when general purpose information transmission is the main concern. This explains the successes of information theory in the retina and partly in the primary visual cortex, to the extent of quantitative agreement with experiments and predictive power for new data (Dong and Atick 1995, Chen and Li 1998). Optimal sensory coding in later stages of sensory pathways is expected to depend on cognitive tasks beyond simple information transmission, and should require applications of alternative theories in future research.

REFERENCE

Atick JJ 1992 "Could information theory provide an ecological theory of sensory processing?" Network: computation in neural systems 3, 213-251.

Attneave F. 1954, Informational aspects of visual perception, Psych. Rev. 61: 183-193.

Barlow HB, 1961, "Possible principles underlying the transformations of sensory messages." In: Sensory Communication W.A. Rosenblith, ed., MIT Press, pp. 217-234.

Chen Danmei and Li Zhaoping 1998 A psychophysical experiment to test the efficient stereo coding theory in Theoretical aspects of neural computation K.M. Wong, I. King, and D.Y. Yeung (eds) Springer-verlag

Dong DW, Atick JJ 1995 "Temporal decorrelation: a theory of lagged and non-lagged responses in the lateral geniculate nucleus," Network: Computation in Neural Systems, 6:159-178.

Dong DW, Atick JJ (1995) "Statistics of natural time-varying images"
Network: Computation in Neural Systems, 6:345-358.

Field DJ 1987 Relations between the statistics of natural images and the response properties of cortical cells. Journal of Optical Society of America, A 4(12):2379-94. 1987

Levy WB and Baxter RA 1996 Energy efficient neural codes. Neural Computation, 8(3) 531-43.

Li Zhaoping and Atick J. J. 1994a, "Efficient stereo coding in the multiscale representation" Network: computation in neural systems Vol.5 1-18.

Li Zhaoping and Atick J. J. 1994b, "Towards a theory of striate cortex" Neural Computation **6, 127-146**

Li Zhaoping 1996 "A theory of the visual motion coding in the primary visual cortex" Neural Computation vol. 8, no.4, p705-30.

Li Zhaoping 1992 "Different retinal ganglion cells have different functional goals" International J. of Neural Systems Vol. 3, No.3 237-248.

Li Zhaoping 2002, "A saliency map in primary visual cortex " Trends in Cognitive Sciences Vol 6. No.1. Jan. 2002, page 9-16

Linsker R. 1990 Perceptual neural organization: some approaches based on network models and information theory. Annu Rev Neurosci. 13:257-81.

Nadal J-P, Parga N. 1997 "Redundancy reduction and independent component analysis: conditions on cumulants and adaptive approaches" Neural Computation 9(7) p.1421-1456

Simoncelli E. P., Freeman W. T., Adelson E. H., and Heeger D. J. 1992 "Shiftable multiscale transforms", IEEE Trans. Informat. Theory Vol.38, p 587-607.

Simoncelli E and Olshausen B. 2001 "Natural image statistics and neural representation" Annual Review of Neuroscience, 24, 1193-216.