Vision Res. Vol. 33, No. 1, pp. 123–129, 1993
Printed in Great Britain.

0042-6989/93 $5.00 + 0.00

What Does Post-Adaptation Color Appearance Reveal About Cortical Color Representation?

JOSEPH J. ATICK,*†‡ ZHAOPING LI,*† A. NORMAN REDLICH*

Received 8 April 1992; in revised form 4 August 1992

We examine the implications of the hypothesis that color information in the cortex is adaptively coded into a factorial (statistically independent) and gain-controlled representation. We show that this hypothesis explains the results of the recent experiments by Webster and Mollon [(1991) *Nature, 349,* 235–238] on changes in color appearance following post-receptoral adaptation. We also give a neural network with a deterministically convergent, unsupervised learning algorithm that reproduces the adaptation observed.

Color coding Adaptation Decorrelation

INTRODUCTION

The ultimate goals of any sensory processing is to locate, identify and discover associations among objects in an animal's environment. In vision this task entails segregation and identification of objects from input data consisting of the light levels in the scene as signaled by the array of photoreceptors. Recently, it has been argued that achieving such cognitive goals can be facilitated if the nervous system first preprocesses input data by recording it into a special type of representation known as a *factorial* code (Barlow, 1961, 1989; Atick & Redlich, 1990, 1992, 1993; for related ideas, see Linsker, 1988). This code has the special property that elements (e.g. pixels) of the representation are statistically independent. One immediate advantage of factorial codes is that the probability of any complex stimulus—the joint probability of the elements—can be computed simply from the individual probabilities of the elements that it activates. This is because statistical independence of the elements means that every joint probability factorizes into a product of individual probabilities (and hence the name factorial). Since the number of individual probabilities is far smaller than the number of possible joint probabilities, factorial coding allows the brain to learn, store, and access far more statistical knowledge than would otherwise be possible. Knowledge of these statistical properties then provides a set of constraints which are expected to be useful in solving the object recognition problem.

Actually, factorial coding does more than just create an efficient way to represent statistical properties; it also preprocesses sensory signals in a way which prepares them for some further useful types of processing: one of these is data compression which can be achieved following the factorial coding by applying a simple (quantizing) gain control to each of the statistically independent outputs. This eliminates redundant bits of data, and is likely to be an early step needed to fit sensory signals into what appears to be a very tight computational (attention) bottleneck later in the processing stream (Van Essen, Olshausen, Anderson & Gallant, 1991). Another use of factorial coding is as a first stage in segmenting an image into statistically independent parts which should aid in the object segregation problem. Finally, factorial coding helps in separating objects from background since it can be shown to give most weight to those parts of an image (such as boundaries) which are less predictable, and it also allows discovery of true associations as deviations from independence, as emphasized by Barlow (1989).

In vision the sampled representation of natural scenes is known to possess a high degree of correlations among pixels (photoreceptor responses) and hence is far from factorial. The first step in producing a factorial representation appears to take place in the retina where at high luminance (high signal-to-noise) pairwise correlations in the input are eliminated at the ganglion cell outputs. Of course this is just a first step in producing a factorial code since images also have important higher order structure (higher order correlations). It is therefore important to test for evidence for factorial coding beyond the retina, in the cortex. However, the problem there becomes very complicated since, for example, groups of neurons have complex lateral connections in the cortex so identifying

*School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540, U.S.A.
†Present address: The Rockfeller University, 1230 York Avenue, New York, NY 10021, U.S.A.
‡To whom all correspondence should be addressed.

the statistically independent elements becomes far more difficult than in the retina.

One way to test the general idea of factorial coding, while avoiding some of the complications of the detailed cortical processing at the neural level, is through quantitative psychophysical studies. The idea is to look for changes in perception following adaptation to an environment with changed statistical properties. The reason this is interesting is that the mapping from the input signals to the factorial representation depends on the statistical properties of the stimulus environment. Modifying those statistical properties will modify the mapping to a factorial representation. Therefore, if the brain is sufficiently plastic and the idea of coding for statistical independence is correct then one can alter the cortical transformations through clever environmental adaptation. This alteration will exhibit itself in experiments as a change in perception following adaptation.

In this paper we examine one such recent psychophysical adaptation experiment in the color domain and analyze at the quantitative level what it implies about the underlying representation of color information in the brain [this builds on earlier works of Buchsbaum and Gottschalk (1983) and of Atick, Li and Redlich (1992)]. We show that the idea that color in the cortex is adaptively decomposed into two* statistically independent channels can easily account for observed changes in color appearance following adaptation to an environment with one particular axis in color space: the theory makes quantitative predictions for most (14 out of 16 points) of the experimental data, given two initial data points. (We also propose a modified experiment where the theory would make predictions for all of the results.) In addition we give a local biologically plausible learning algorithm that can achieve the adaptation observed.

COLOR CODING: AN EFFICIENT REPRESENTATION?

We adopt the hypothesis that one goal of the visual pathway is to build a statistically independent representation of the image. In the particular problem of color coding at hand, this means that the activity of the L and M cones (ignoring the S cones) which is highly correlated needs to be recoded to achieve the desired decorrelated representation (factorial representation). We next examine the issue of decorrelation of two channels mathematically.

Let us denote the activity of the two types of photoreceptors by S_a, where $a = 1, 2$ stands for L, M respectively. Then the autocorrelator which captures the degree of correlation between these signals is given by $\langle S_a S_b \rangle \equiv R_{ab}$, where the brackets denote an average over the ensemble of signals. To decorrelate one can use a linear transformation K_{ab} on the input signals S_b to produce the output

$$O_a = \sum_{b=1}^{2} K_{ab} S_b \tag{1}$$

such that

$$\langle O_a O_b \rangle = (\mathbf{K} \cdot \mathbf{R} \cdot \mathbf{K}^T)_{ab} = 0 \quad \text{if } a \neq b \tag{2}$$

where bold-faced quantities are matrices (below boldface will be used to denote vectors as well). For this 2×2 problem the simplest transformation $\mathbf{K}$ needed is just a rotation $\mathbf{U}$ that diagonalizes $\mathbf{R}$:

$$\mathbf{U} \cdot \mathbf{R} \cdot \mathbf{U}^T = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \tag{3}$$

where λ_1 and λ_2 are the two eigenvalues of $\mathbf{R}$.

Actually, we go one step beyond decorrelation and normalize the output signals such that

$$\langle O_a O_b \rangle = (\mathbf{K} \cdot \mathbf{R} \cdot \mathbf{K}^T)_{ab} = \delta_{ab}. \tag{4}$$

This is a stronger condition than decorrelation alone since it involves an additional step of gain control. This additional gain control is the one which as mentioned in the Introduction is needed to use the factorial code to achieve data compression. In this way the two incoming signals S_1, S_2 can be fit into a couple of channels with the smallest possible dynamic range. One particular transformation $\mathbf{K}_0$ that satisfies equation (4) can be parameterized in the product form

$$\mathbf{K}_0 = \mathbf{V} \cdot \mathbf{U} \tag{5}$$

where $\mathbf{U}$ is the rotation in equation (3) and $\mathbf{V}$ is the gain control, which in this linear model is

$$\mathbf{V} = \begin{pmatrix} 1/\sqrt{\lambda_1} & 0 \\ 0 & 1/\sqrt{\lambda_2} \end{pmatrix}. \tag{6}$$

The specific transformation $\mathbf{K}_0$ in equation (5) is not the only transformation, however, that satisfies equation (4). In general we can always construct another transformation $\mathbf{K} = \mathbf{M} \cdot \mathbf{K}_0$ where $\mathbf{M}$ is any orthogonal matrix, $\mathbf{M} \cdot \mathbf{M}^T = \mathbb{1}$, such that $\mathbf{K} \cdot \mathbf{R} \cdot \mathbf{K}^T = \mathbb{1}$. This means that the problem of finding a factorial representation with gain control, as proposed by Barlow (1989, 1992) and Barlow and Foldiak (1989), has no unique solution. Rather there is a whole class of representations (in this case a one parameter class) which are all equivalent in the sense that they all have decorrelated outputs with unit variance. This is the same nonuniqueness that was faced in previous work on spatial decorrelation in the retina (Atick & Redlich, 1992, 1993). There it was shown that although all the transformations $\mathbf{K} = \mathbf{M} \cdot \mathbf{K}_0$ are equivalent in their decorrelation properties they are not all biologically plausible. In fact, it was shown that the condition of locality of the transformation selects out a unique transformation—leads to a unique choice of $\mathbf{M}$—which is the one observed in the retina.

In the present context, we fix this $\mathbf{M}$ symmetry by using a formal generalization of the principle of locality. Taking the output $\mathbf{O} = \mathbf{K} \cdot \mathbf{S}$ with $\mathbf{K} = \mathbf{M} \cdot \mathbf{K}_0 = \mathbf{M} \cdot \mathbf{V} \cdot \mathbf{U}$, we insist that $\mathbf{M}$ be chosen so as to minimize

*For simplicity we ignore the S (or blue) cone system contribution to color coding and focus on the two dimensional subspace spanned by the L (red) and M (green) cones.

the quantity $\mathrm{Tr}\langle(\mathbf{S}-\mathbf{O})^2\rangle$. In the purely spatial domain this can be shown to be the same as the condition of locality. In general, what this condition means is that we pick from among all the equivalently decorrelated and compressed representations **O** the one that remains as close as possible to the original representation S. This is in a sense a general statement of locality: the recordings that are favored biologically are those which require the least perturbation of the original signal but at the same time achieve the goal intended.

We can find the optimal local map by finding the solution to the variational equations $\delta E\{\mathbf{M}\}/\delta\mathbf{M}=0$ where

$$E\{\mathbf{M}\}=\mathrm{Tr}\langle(\mathbf{S}-\mathbf{M}\cdot\mathbf{V}\cdot\mathbf{U}\cdot\mathbf{S})^2\rangle - \mathrm{Tr}[\rho(\mathbf{M}\cdot\mathbf{M}^{\mathrm{T}}-\mathbb{1})] \quad (7)$$

and the matrix $\rho=\rho^T$ is a Lagrange multiplier enforcing the orthogonality constraint $\mathbf{M}\cdot\mathbf{M}^T=\mathbb{1}$. It is not difficult to show that the optimal solution is

$$\mathbf{M}=\mathbf{U}^{-1}=\mathbf{U}^T. \quad (8)$$

Thus in general we propose that the "most local" transformation is

$$\mathbf{K}=\mathbf{U}^T\cdot\mathbf{V}\cdot\mathbf{U}. \quad (9)$$

In the next section we demonstrate that the transformation (9) predicts the correct color adaptation. On the other hand, the minimal rotation plus gain control in equation (5) does not make correct predictions. This does not mean, however, that we disagree with Barlow's principle since *both* equations (5) and (9) produce statistically independent outputs. Rather the principle of statistical independence alone is incomplete, because of the arbitrary M rotation. To fix this symmetry requires an additional principle, which is why we introduced "locality" to determine M. It is the combination of Barlow's idea together with "locality" which gives an unambiguous and correct prediction of the experimental results. Also, in our previous work on color coding in the retina it was precisely this same combination of principles, i.e. the transformation of the form in equation (9) (rotation-scaling-rotation) which we showed agrees with retinal color opponent cells (Atick *et al.*, 1992). As mentioned above it is this type of transformation which is the color analog of the spatio-temporal transformation derived by Atick and Redlich (1992). Later in this paper we show that it is also the type of transformation that is learned by a biologically plausible neural network algorithm, one that is guaranteed to converge. In what follows, we show how the transformation **K** in (9) can explain the results of the adaptation experiment of Webster and Mollon (1991).

IMPLICATIONS OF THE THEORY TO COLOR ADAPTATION

We start by describing the experiment of Webster and Mollon (1991). In this experiment subjects view a spatially uniform chromatic stimulus that is temporally modulated and is presented in a restricted area of their visual field. The stimulus is modulated along a fixed axis in color space as indicated by the angle θ with the luminance axis in Fig. 1, while its stimulation strength or saturation is given by the radial distance. The temporal modulation does not affect the angle, but varies the saturation about a fixed mean defined as the origin in Fig. 1. After a certain period of presenting this adapting stimulus, subjects are then presented in the same visual area with a test stimulus. The task is to match the perceived color of the test stimulus by adjusting the color of a matching stimulus simultaneously presented in another visual area placed symmetrically on the other side of fixation (see Fig. 1). The lower part of Fig. 1 shows the results of a typical experiment from Webster and Mollon (1991) in the luminance, $L+M$, and chrominance, $L-M$ subspace. The points on the circle represent test colors (the axes have been normalized so that a unit distance along each axis is one detection threshold unit) while the experimental points (triangles) represent the perceived color as determined through the matching

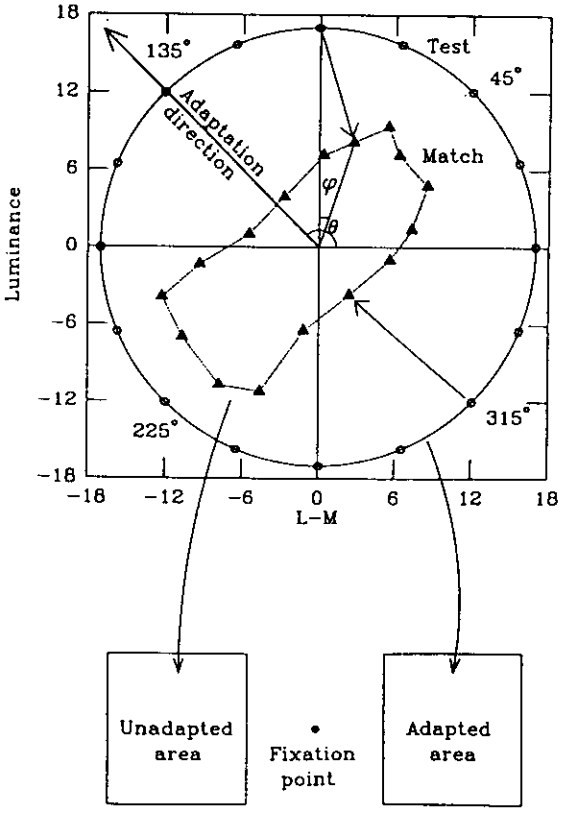

FIGURE 1. Color adaptation experiment by Webster and Mollon (1991). A stimulus in chromatic space is marked by a vector in the coordinate system with luminance $(L+M)$ and chrominance $(L-M)$ as axes. During adaptation, the subject views a restricted visual field in which the stimulus strength is modulated about a fixed mean, the origin in this coordinate, along a particular hue or adaptation direction. After adaptation, one of the test stimuli, indicated by small open circles on the large circle of radius 17 threshold units, is presented in the adapted visual area. This test stimulus is matched in appearance to a stimulus (solid triangle) presented in the upadapted visual area, as indicated for two examples by arrows from small open circles to triangles. For example, the test stimulus (0, 17) on the luminance axis is matched to a stimulus of a smaller strength and hue angle ϕ from the vertical.

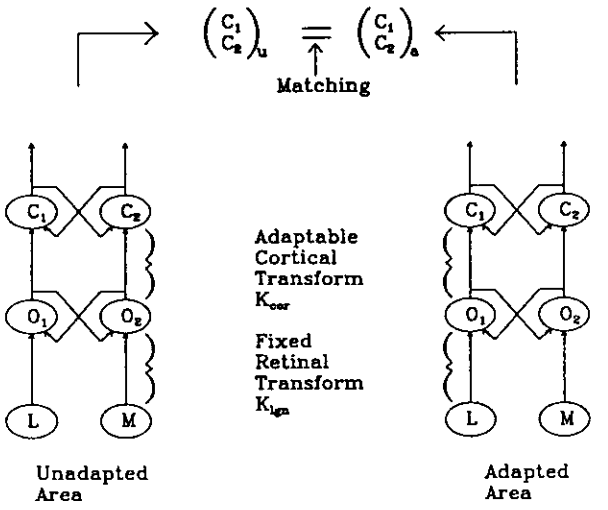

FIGURE 2. Schematic of the color adaptation and matching processes proposed for the brain. Cone signals L and M are transformed to retinal outputs O_1, O_2 by a fixed retinal (or lateral geniculate) transform K_{lgn} which decorrelates O_1 and O_2 for the ensemble of natural signals. O_1, O_2 are further transformed to cortical outputs C_1 and C_2 by the transform $\mathbf{K}_{cor}$ which adapts to keep C_1 and C_2 decorrelated even when new correlations between O_1 and O_2 are introduced in the adapting environment. In the unadapted area, $\mathbf{K}_{cor}$ is an identity transform $\mathbf{1}$.

procedure just described. It is clear from this data that adaptation causes major changes in both the perceived angle (perceived axis in color space) and saturation of the stimulus (radial distance). Since it is known that the type of stimulus used in the Webster and Mollon experiment does not affect the sensitivities of the photoreceptors or even retinal or geniculate neurons (Derrington, Krauskopf & Lennie, 1984) the observed adaptation must be of cortical origin.

We next show that the coding $\mathbf{K}$ [equation (9)] does quantitatively explain the Webster and Mollon data. Figure 2 gives a schematic of the color processing stages assumed for the visual pathway. We have broken up the color transformation into two stages: the first, given by the transformation $\mathbf{K}_{lgn}$ from L, M to O_1, O_2, is supposed to be fixed and nonadaptable (on the time scale and for the specific stimulus of the Webster and Mollon experiment) as just mentioned in the previous paragraph. As discussed in an earlier publication (Atick *et al.*, 1992) there is evidence that the fixed transformation $\mathbf{K}_{lgn}$ decorrelates color signals from the *natural environment*. It is therefore given by the transformation (9) with $\mathbf{U}$ and $\mathbf{V}$ determined by the autocorrelator of the photoreceptor signals L, M in response to natural stimulation. Therefore, we identify the axes O_1, O_2 (which are rescaled by V in $\mathbf{K}_{lgn}$, in the threshold units) with the experimental axes used by Webster and Mollon.*

The second transformation $\mathbf{K}_{cor}$ in Fig. 2 from O_1, O_2 to some cortical modules C_1, C_2 is the one that is assumed to be plastic or adaptable. Before adaptation, or in the unadapted patch—since the input to the cortex O_1, O_2 is already decorrelated—this second transformation $\mathbf{K}_{cor}$ is trivial $\mathbf{K}_{cor} = \mathbb{1}$. However, by exposing a subject to an adapting stimulus, the cortical inputs O_1, O_2 which were statistically independent in the natural environment become correlated. To restore decorrelation at C_1, C_2 the cortex has to apply a nontrivial map $\mathbf{K}_{cor}$ to O_1, O_2. Thus we can write

$$\begin{pmatrix} C_1 \\ C_2 \end{pmatrix}_u = \mathbb{1} \cdot \begin{pmatrix} O_1 \\ O_2 \end{pmatrix}_u$$
$$\begin{pmatrix} C_1 \\ C_2 \end{pmatrix}_a = \mathbf{K}_{cor} \cdot \begin{pmatrix} O_1 \\ O_2 \end{pmatrix}_a \tag{10}$$

where u and a stand for the unadapted and adapted patch respectively. In matching one attempts to find the stimulus $(O_1, O_2)_u$ that gives rise to the same level of cortical activity in the unadapted patch as does the test stimulus $(O_1, O_2)_a$ in the adapted patch. Mathematically, the matching condition is

$$\begin{pmatrix} C_1 \\ C_2 \end{pmatrix}_u = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}_a \tag{11}$$

which leads to the following equation

$$\begin{pmatrix} O_1 \\ O^2 \end{pmatrix}_u = \mathbf{K}_{cor} \cdot \begin{pmatrix} O_1 \\ O_2 \end{pmatrix}_a$$
$$= \mathbf{U}^T \cdot \mathbf{V} \cdot \mathbf{U} \cdot \begin{pmatrix} O_1 \\ O_2 \end{pmatrix}_a \tag{12}$$

where $\mathbf{U}$, $\mathbf{V}$ for $\mathbf{K}_{cor}$ are determined by equations (3) and (6) for the autocorrelator $\mathbf{R}$ of the cortical inputs (O_1, O_2) for the adapting stimulus (environment).

Unfortunately, the adaptation ensemble in the Webster–Mollon experiment does not produce a well defined autocorrelator: strictly speaking a single stimulus, modulated in time along a fixed axis, gives a correlation matrix with one vanishing eigenvalue. Actually, this eigenvalue must be nonvanishing due to the existence of at least some noise (at least some signal quantization) at all processing levels. However, the noise contribution to the autocorrelator is not known. This turns out to mean that the eigenvalues λ_1 and λ_2 of the autocorrelator $\mathbf{R}$ are not determined by the experiment. On the other hand, the basis where the autocorrelator $\mathbf{R}$ is diagonal *is* determined by experiment; it is obtained by rotating the unadapted axis by the adaptation angle θ (see Appendix). Thus we can parameterize the autocorrelator as

$$\mathbf{R} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix} \times \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{13}$$

where λ_1 and λ_2 are the unknown eigenvalues of the autocorrelator which we treat as parameters in what follows. A minor change in the experiment, using an adapting ensemble with at least two colors, would give control over the ratio λ_1/λ_2 which as we show below

*Actually the axes used in the experiments of Webster and Mollon are a fixed rotation of the model LGN O_1, O_2 axes. Ignoring this distinction does not alter the theoretical predictions below but simplifies the discussion. Our analysis below can be repeated keeping explicit this fixed rotation without any difficulty.

would result in a parameter-free prediction of color axis shifts.

We should mention that the observed ratio λ_1/λ_2 (see below) does turn out to imply a rather low signal-to-noise ratio (see Appendix). Therefore, although noise must be at least partly responsible for nonvanishing eigenvalues, it may not explain why the ratio λ_1/λ_2 is as low as it is. An alternative explanation is that the cortex is unable to completely decorrelate the input signals. One way to test this idea would be to use the learning algorithm (see below) with a high signal-to-noise ratio, but to stop the algorithm before it converges. The problem with this is that the algorithm requires a relatively low signal-to-noise ratio for stability. We have not found any other quantitative way to explore incomplete decorrelation.

With the parameterization (13), it is clear from equations (3) and (6) that $\mathbf{U}$ is a rotation by θ, while $\mathbf{V}$ is a scaling matrix $\mathrm{diag}(1/\sqrt{\lambda_1}, 1/\sqrt{\lambda_2})$. Thus, according to equation (9) the cortical transformation $\mathbf{K}_{cor}$ is

$$\mathbf{K}_{cor} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \cdot \begin{pmatrix} 1/\sqrt{\lambda_1} & 0 \\ 0 & 1/\sqrt{\lambda_2} \end{pmatrix} \times \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}. \quad (14)$$

If the points in $(O_1, O_2)_a$ lie on a circle then from equations (14) and (12) it is clear that the matching points $(O_1, O_2)_u$ will lie on an ellipse whose minor and major axes are in the θ direction and the direction orthogonal to it, respectively. The lengths of the minor and major axes of the ellipse are $1/\sqrt{\lambda_1}$ and $1/\sqrt{\lambda_2}$ respectively. Thus the theory predicts that the matching data should lie on an ellipse whose minor axis is oriented along the adapting axis, but whose major and minor axes lengths—for the Webster and Mollon experiment—are not otherwise determined.

We can now make some quantitative comparisons with the Webster and Mollon data. Because we do not know λ_1 and λ_2 we cannot predict all of their data, but this does not mean we can make no quantitative predictions. In fact, we need only two data points to determine λ_1 and λ_2, after which we can predict all of the other 14 data points per subject. One interesting type of prediction to make for these 14 data points is of the angle φ between the test stimulus axis and the matching stimulus axis. Recall, in this experiment that a subject is adapted to a stimulus at angle θ and *then* is exposed to one of the 16 different test stimuli which lie on the circle, as shown in Fig. 1. For each test point on that circle the subject selects a matching stimulus, giving the set of points which lie approximately on an ellipse (Fig. 1). Now for each test stimulus (one point on the circle), one can ask how much *its* matching stimulus (one triangle on the ellipse) is rotated from it. For example, if the test stimulus is at 45° and the match is at 60° then the shift angle $\varphi = 15°$. The magnitude of the shift will depend on the adapting angle θ. We can therefore measure the set of shift angles for one particular test stimulus, say at 90 or 270°, to obtain $\varphi(\theta)$, as was done by Webster and Mollon (see Fig. 3).

To compare these experimental measurements to the theoretical predictions we now need to compute $\varphi(\theta)$. For the 90° test stimulus $(O_1, O_2)_a = (0, 1)$ so from equations (12) and (14) the shift angle $\tan\varphi(\theta) = (-O_1/O_2)_u$ is

$$\tan\varphi(\theta) = \frac{(-K_{cor})_{12}}{(K_{cor})_{22}} = \frac{\cos\theta\sin\theta(\sqrt{\lambda_1/\lambda_2} - 1)}{\sin^2\theta + \sqrt{\lambda_1/\lambda_2}\cos^2\theta} \quad (15)$$

which only depends on the ratio λ_1/λ_2. From two points of the experimental response data [Fig. 2(C) in Webster

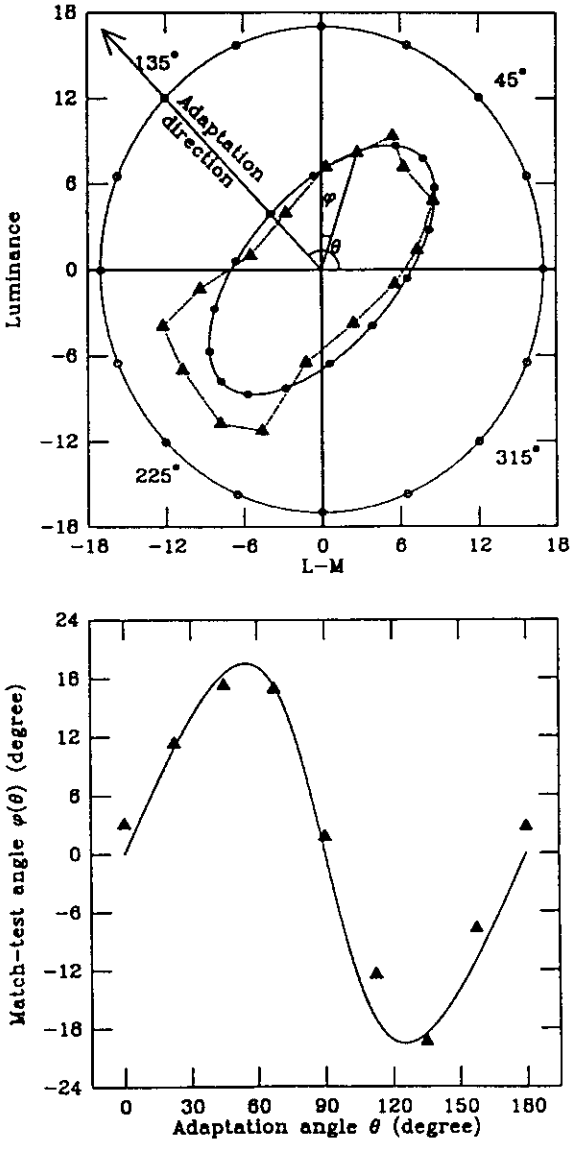

FIGURE 3. Comparison of theoretical prediction and experimental measurement of hue shifts φ. The lower graph shows the shifts in angle of a test stimulus on the luminance axis as a function of adaptation direction. The triangles are the experimental data obtained by averaging the measured shifts for the positive and negative luminance axis at each adaptation angle θ (Webster & Mollon, 1991). In the theory, the amount of shift predicted depends only on λ_1/λ_2, which is roughly constant for all adaptation directions, with average $\lambda_1/\lambda_2 = 4$ from the experimental response data. With this λ_1/λ_2, the solid curve gives the predicted $\varphi(\theta)$. The upper graph gives the predicted (solid circles) and measured (triangles) match stimuli at one adaptation direction. The prediction is derived with the same λ_1/λ_2 as in the lower graph, and $\lambda_1 = 9.6$.

and Mollon (1991)], we estimate that this ratio is 4.0, so in Fig. 3 we plot $\varphi(\theta)$ for $\lambda_1/\lambda_2 = 4.0$. In that figure the continuous curve is the predicted shift from equation (15) while the solid triangles are the experimental measurements. Also, from equation (15) we can determine the maximum shift possible φ_{max} and the angle θ_{max} at which it occurs. Solving $d\phi(\theta)/d\theta = 0$, we get $\sin^2\theta_{max} = \sqrt{\lambda_1/\lambda_2}/(1+\sqrt{\lambda_1/\lambda_2})$ and $\tan\varphi_{max} = (\sqrt{\lambda_1/\lambda_2}-1)/[2(\lambda_1/\lambda_2)^{1/4}]$, which for $\lambda_1/\lambda_2 = 4$ gives $\theta_{max} \sim 54.7°$ and $\phi_{max} = 19.5°$, consistent with what is typically found by Webster and Mollon.

NETWORK IMPLEMENTATION

In what follows we present a neural network with a learning algorithm that allows the network to adaptively diagonalize the current autocorrelation matrix: the network learns and implements the transform $\mathbf{K}_{cor} = \mathbf{U}^T \cdot \mathbf{V} \cdot \mathbf{U}$ of equation (9). The network architecture is shown in Fig. 2, where the feedforward connections from O_1, O_2 to C_1, C_2 are assumed to be fixed and set to unity (we use a convention where neurons are labeled by their outputs). The adaptable links are assumed to be the lateral feedback links which connect the outputs $C_1(C_2)$ back to the input of neuron $C_2(C_1)$ with link strength $W_{12}(W_{21})$. The dynamics is thus

$$T\frac{dC_i}{dt} = O_i - \sum_{j=1}^{2} W_{ij}C_j.$$

At equilibrium, $dO_i/dt = 0$, one gets $O_i = \Sigma_{j=1}^2 W_{ij}C_j$ which shows that $\mathbf{W}$ is the inverse of the transformation of $\mathbf{K}_{cor}$ from O_1, O_2 to C_1, C_2 given in equation (10). The proposed update algorithm is

$$\tau\frac{dW_{ij}}{dt} = O_iC_j - W_{ij}. \tag{16}$$

This algorithm was previously applied to pairwise decorrelate in space the photoreceptor signals for a stimulus ensemble with the same power spectrum as natural scenes, and it was shown to converge to ganglion cell receptive fields (Atick & Redlich, 1993). The algorithm was originally proposed by Goodall (1960) in a different context.

The algorithm in equation (16) when averaged over presentations of signals drives W_{ij} to a configuration where $\langle C_iC_j\rangle = [\mathbf{W}^{-1}\cdot\mathbf{R}\cdot(\mathbf{W}^{-1})^T]_{ij} = \delta_{ij}$. For a proof of this the reader can consult Atick and Redlich (1993). In that paper it was also noted that the algorithm possesses a symmetry of multiplication by any orthogonal matrix $\mathbf{M}$. For example, if $\mathbf{W}$ achieves decorrelation then so does $\mathbf{W}\cdot\mathbf{M}$. This is the same symmetry that we discussed in the paragraph following equation (6), where we argued that although the principle of decorrelation does not select a unique $\mathbf{M}$, there may be biological reasons for favoring a particular $\mathbf{M}$. That led to the choice $\mathbf{M} = \mathbf{U}^T$ based on the principle of locality. What we wish to do now is to give an additional argument for the choice $\mathbf{M} = \mathbf{U}^T$. Namely, that $\mathbf{M} = \mathbf{U}^T$ is exactly the $\mathbf{M}$ that results by applying the developmental algorithm (16) to learn color decorrelation starting with the initial condition—before adaptation—$\mathbf{K}_{cor} = \mathbf{1}$ (equivalently $\mathbf{W} = \mathbf{1}$) as in equation (10).

In Atick and Redlich (1993) it was shown that the final $\mathbf{M}$ found by the algorithm after convergence depends on the initial condition for $\mathbf{W}$. It was further proven that with the initial condition $\mathbf{W} = \mathbf{1}$ the algorithm is guaranteed to converge to a configuration with $\mathbf{W} = \mathbf{W}^T$ ($\mathbf{K}_{cor} = \mathbf{K}_{cor}^T$). This condition is exactly equivalent to $\mathbf{M} = \mathbf{U}^T$ since

$$\mathbf{K}_{cor} = \mathbf{M}\cdot\mathbf{V}\cdot\mathbf{U} = \mathbf{U}^T\cdot\mathbf{V}\cdot\mathbf{M}^T = \mathbf{K}_{cor}^T \tag{17}$$

has only one solution $\mathbf{M} = \mathbf{U}^T$, assuming orthogonality for $\mathbf{M}$.

The algorithm (16) was simulated using Gaussian signals O_1, O_2 with a fixed autocorrelator. We find that the algorithm, as expected, learns a $\mathbf{K}_{cor}$ (or equivalently $\mathbf{W}$) which can be factorized into the form in equation (9) with $\mathbf{U}$ and $\mathbf{V}$ given by the correct rotation and scaling matrices for the particular signals used. For additional discussion on detailed simulations of algorithms of the type (16) see Atick and Redlich (1993).

DISCUSSION

At this point we should point out that Webster and Mollon (1991) suggested, without going into extensive detail, that their data might be explained in two possible ways. One way is close to our explanation here and refers to some earlier ideas by Barlow and Foldiak (1989) on decorrelated representations. As discussed below equation (9) the principle of decorrelation alone is insufficient to completely determine a transformation, since it leaves an arbitrary rotation which we fix based on the additional principle of "locality". This choice is different from Barlow and Foldiak's (1989) which does not correctly predict the Webster and Mollon (1991) result. However, since the basic principle in both cases is decorrelation plus gain control we consider our results a confirmation of the same fundamental ideas discussed by Barlow and Foldiak (1989). That is, we have shown concretely that coding color in the cortex in a factorial (decorrelated) and gain controlled representation, where the elements (chromatic channels) are coupled and actively adapt to the current environment, provides a very simple quantitative explanation of the Webster and Mollon data.

The alternate proposal mentioned by Webster and Mollon (1991) is very different from the type of explanation that we have presented here and is based on the hypothesis of fatigue. This hypothesis relies on the assumption that repeated activation of a neuron during the adaptation period diminishes its sensitivity. So if a perceptual quantity (e.g. hue) is coded by a collection of neurons (analyzers) and the adaptation stimulus activates these neurons differently, then the fatigue hypothesis predicts that each neuron would be desensitized by different amounts. This in principle can lead to a variety of perceptual shifts.

The fact that fatigue can produce hue shifts is not an issue of debate, what is to be seen is whether it can

produce the correct quantitative shifts. It is not hard to show, as do Webster and Mollon (1991), that the simplest model of fatigue with only two channels cannot explain the data of Webster and Mollon. It does indeed lead to color axis shifts since, in general, adapting the two color neurons with a stimulus defined by the angle θ does produce different activation levels for the two neurons and hence different levels of fatigue. However, it predicts the wrong shifts. For example this model predicts that the matching ellipse has its major and minor axes always along the luminance and $L - M$ axes. Also, when adapting at angle $\theta = 45°$ the matching locus is not an ellipse but a circle in contradiction to what is observed. A more sophisticated model of fatigue with multiple chromatic channels might explain the data, but at the cost of several additional parameters and assumptions. We also find the idea of fatigue uncompelling; it is hard to accept that a system such as the brain—which is known to exhibit all sorts of intricate adaptations--does so merely because of a breakdown of its neuronal response abilities and not in order to serve some function. This is hard to believe especially in view of the fact that active adaptation is a strategy that can enhance the brain's computational power.

Unfortunately, at this stage we do not have any experiment that can unequivocally rule out fatigue in favor of functional adaptation and hence it would be very interesting to try to design some. Actually, a hint that adaptation is not the result of fatigue can be seen in the experimental data of Movshon and Lennie (1979) on pattern selective adaptation as has been pointed out by Barlow (1992). Also, more quantitative control over the statistical properties of the stimulus in the Webster and Mollon experiment might lead to stronger theoretical predictions and hence might lend more credence to the active functional adaptation philosophy. One experiment that we would like to see done, is one where the adaptation stimulus consists of at least two colors modulated in a way which produces a nonsingular correlation matrix. In that experiment one could control the ratio λ_1/λ_2 and hence have parameter free predictions from the theory.

REFERENCES

Atick, J. J. & Redlich, A. N. (1990). Towards a theory of early visual processing. *Neural Computations, 2,* 308–320.

Atick, J. J. & Redlich, A. N. (1992). What does the retina know about natural scenes? *Neural Computations, 4,* 196–210.

Atick, J. J. & Redlich, A. N. (1993). Convergent algorithm for sensory receptive field development. *Neural Computations.* In press.

Atick, J. J., Li, Z. & Redlich, A. N. (1992). Understanding retinal color coding from first principles. *Neural Computations, 4,* 559–572.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In Rosenblith, W. A. (Ed.), *Sensory communication.* Cambridge, Mass.: MIT Press.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computations, 1,* 295–311.

Barlow, H. B. (1992). The biological role of neocortex. *Proceedings of the brain theory meeting held at Ringberg, April 1990.* Berlin: Springer.

Barlow, H. B. & Foldiak, P. (1989). *The computing neuron.* New York: Addison-Wesley.

Buchsbaum, G. & Gottschalk, A. (1983). Trichromacy, opponent colors coding and optimum color information transmission in the retina. *Proceedings of the Royal Society of London B, 220,* 89–113.

Derrington, A. M., Krauskopf, J. & Lennie, P. (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. *Journal of Physiology, 357,* 241–265.

Goodall, M. C. (1960). Performance of stochastic net. *Nature, 185,* 557–558.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer, 21,* 105–117.

Movshon, J. A. & Lennie, P. (1979). Pattern-selective adaptation in visual cortical neurones. *Nature, 278,* 850–852.

Van Essen, D. C., Olshausen, B., Anderson, C. H. & Gallant, J. L. (1991). Pattern recognition, attention, and information bottlenecks in the primate visual system. In *Proceedings of the SPIE conference on visual information processing: From neurons to chips* (Vol. 1473).

Webster, M. A. & Mollon, J. D. (1991). Change in colour appearance following post-receptoral adaptation. *Nature, 349,* 235–238.

Acknowledgement—Supported in part by a grant from the Seaver Institute.

APPENDIX

To demonstrate that the autocorrelator **R** takes the form (13) we need to explicitly add noise N_i to the signal O_i. In the experiment the signals $O_i(t)$ have a time-independent ratio $O_2(t)/O_1(t) = \tan\theta$ determined by the adapting angle θ and have a time-dependent magnitude $P^2(t) = O_1^2(t) + O_2^2(t)$. Therefore the signal can be parameterized as

$$O_i(t) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \cdot \begin{pmatrix} P(t) \\ 0 \end{pmatrix} \equiv \mathbf{U}^T \cdot \begin{pmatrix} P(t) \\ 0 \end{pmatrix}. \tag{A1}$$

Thus without noise the correlation matrix $R_{ij} = \langle O_i O_j \rangle$, where $\langle \cdots \rangle$ here denotes temporal averaging, is

$$\mathbf{R}|_{\text{noise}=0} = \mathbf{U}^T \cdot \begin{pmatrix} \langle P^2(t) \rangle & 0 \\ 0 & 0 \end{pmatrix} \cdot \mathbf{U}, \tag{A2}$$

which explicitly shows the zero eigenvalue of **R**. Now assuming that noise is totally decorrelated $\langle N_i N_j \rangle = N^2 \delta_{ij}$ and has no correlations with the signal $\langle N_i O_j \rangle = 0$, the autocorrelator for $O_i + N_i$ is

$$\mathbf{R}|_{\text{noise}\neq 0} = \mathbf{U}^T \cdot \begin{pmatrix} \langle P^2(t) \rangle + N^2 & 0 \\ 0 & N^2 \end{pmatrix} \cdot \mathbf{U}. \tag{A3}$$

This is the form exhibited in equation (13) which has two nonvanishing eigenvalues as promised.