To appear in Neural Computation. Also, Technical Report 96.2 (revision of 95.4) National Resource Laboratory for the Study of Brain and Behavior, University of Rochester, November 1995.

Dynamic Model of Visual Recognition Predicts Neural Response Properties in the Visual Cortex*

Rajesh P.N. Rao and Dana H. Ballard

Department of Computer Science, University of Rochester,

Rochester, NY 14627-0226, USA

The responses of visual cortical neurons during fixation tasks can be significantly modulated by stimuli from beyond the classical receptive field. Modulatory effects in neural responses have also been recently reported in a task where a monkey freely views a natural scene. In this paper, we describe a hierarchical network model of visual recognition that explains these experimental observations by using a form of the extended Kalman filter as given by the Minimum Description Length (MDL) principle. The model dynamically combines input-driven bottom-up signals with expectation-driven top-down signals to predict current recognition state. Synaptic weights in the model are adapted in a Hebbian manner according to a learning rule also derived from the MDL principle. The resulting prediction/learning scheme can be viewed as implementing a form of the Expectation-Maximization (EM) algorithm. The architecture of the model posits an active computational role for the reciprocal connections between adjoining visual cortical areas in determining neural response properties. In particular, the model demonstrates the possible role of feedback from higher cortical areas in mediating neurophysiological effects due to stimuli from beyond the classical receptive field. Simulations of the model are provided that help explain the experimental observations regarding neural responses in both free viewing and fixating conditions.

1 Introduction

Much of the information regarding the response properties of cells in the primate visual cortex has been obtained by displaying visual stimuli within a cell's receptive field while a monkey fixates on a blank screen and observing, via microelectrode recordings, the neuronal responses elicited by the displayed stimuli [Hubel and Wiesel, 1968; Zeki, 1976; Baizer *et al.*, 1977; Andersen *et al.*, 1985; Maunsell and Newsome, 1987]. The initial assumption was that the responses thus recorded are not substantially different from those generated in a more natural setting, but later experiments showed that the responses of many visual cortical neurons can be significantly modulated by stimuli from beyond the classical receptive field [Nelson and Frost, 1978; Zeki, 1983; Allman *et al.*, 1985; Desimone and Schein, 1987; Gulyas *et al.*, 1987; Muller *et al.*, 1996]. Modulatory effects in neural responses have also been recently reported by [Gallant *et al.*, 1994; 1995; Gallant, 1996] in a task where a monkey freely

*A preliminary account of this work appeared as [Rao and Ballard, 1995b].

views a natural scene. In these experiments, responses from the visual areas V1, V2, and V4 were recorded while a monkey freely viewed natural images. The image subregions that fell within a cell's receptive field during free viewing were then extracted from the images and flashed within the cell's receptive field while the monkey fixated on a blank screen. Surprisingly, the responses obtained when the stimuli were flashed during the fixation task were found to be generally *much larger* than when the same stimuli entered the receptive field during free viewing.

In this paper, we describe a hierarchical network model of dynamics and synaptic learning in the visual cortex that provides explanations for the above experimental observations by exploiting two fundamental properties of the cortex: (a) the reciprocity of connections between any two connected areas [Rockland and Pandya, 1979; Van Essen, 1985; Felleman and Van Essen, 1991], and (b) the convergence of inputs from lower area neurons to a higher area neuron, resulting in larger receptive field sizes for neurons in the higher area [Van Essen and Maunsell, 1983; Desimone and Ungerleider, 1989]. The reciprocity of connections allows feedback pathways to convey top-down reconstructed signals that serve as predictions for lower level modules, while the feedforward pathways convey the *residuals* between current recognition state and the top-down predictions from the higher level. The larger receptive field size of higher level neurons allows them to integrate information from a larger spatial extent, which in turn permits them to influence, via feedback pathways, lower level units operating at a smaller spatial scale.

The model acknowledges the fact that vision is an active, dynamic process subserving the larger cognitive goals of the organism that the visual system is embedded in. In particular, the dynamics of a cortical module at a given hierarchical level is shown, using the Minimum Description Length (MDL) principle [Rissanen, 1989; Zemel, 1994], to assume the form of an extended *Kalman filter* [Kalman, 1960; Kalman and Bucy, 1961; Maybeck, 1979] which optimally estimates current recognition state by combining information from input-driven bottom-up signals and expectation-driven top-down signals. The dynamics also allows a nonlinear prediction step that facilitates the processing of time-varying stimuli, and helps to counteract the signal propagation delays introduced by the different hierarchical levels of the network. Prediction is mediated by a set of units whose weights can be adapted in order to minimize the prediction error. The transformation of signals from lower, less abstract areas to higher, more abstract areas is achieved via synaptic weights in the feedforward pathway while the transformation back is achieved via the feedback weights, both of which are learned during exposure to natural stimuli. This learning is mediated by a Hebbian learning rule also derived from the MDL principle. The weights are shown to approach the transposes of each other which in turn allows the fast dynamics of the Kalman filter to be efficiently implemented. The combined prediction/learning scheme can be viewed as implementing a form of the Expectation-Maximization (EM) algorithm [Baum *et al.*, 1970; Dempster *et al.*, 1977].

We first verify the hypothesis that reciprocal connections between different hierarchical levels can mediate

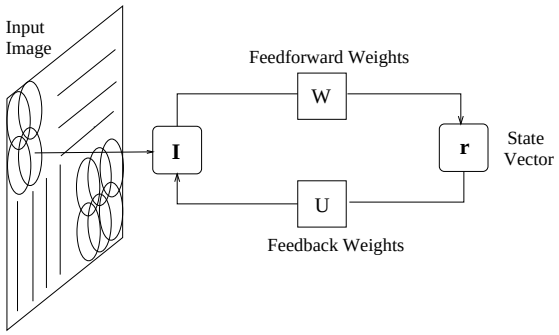

Figure 1: **A Simple Input Encoder Network**. The goal is to optimally encode inputs $\mathbf{I}$ by finding best estimates $\widehat{U}$, $\widehat{W}$, and $\widehat{\mathbf{r}}$ for the network parameters U, W, and $\mathbf{r}$. Optimality is defined in terms of least squared reconstruction error where the reconstruction of the input is given by $\mathbf{I}' = U\mathbf{r}$.

robust visual recognition by assessing the model's performance on a small training set of realistic objects. The model is shown to recognize objects from its training set even in the presence of partial occlusions and maintain a fine discrimination ability between highly similar visual stimuli. By exposing the network to natural images, we show that the receptive fields developed at the lowest two hierarchical levels are comparable to those of cells at the corresponding levels in the primate visual cortex. We then describe simulations of the model in both free viewing and fixating conditions and show that the simulated neural responses are qualitatively similar to those observed in the neurophysiological experiments of [Gallant *et al.*, 1994; 1995; Gallant, 1996]. We conclude by describing related work on cortical modeling, Kalman filters, and learning/parameter estimation techniques (including the relationship to the EM algorithm), and discuss the use of the model as a computational substrate for understanding a number of well-known psychophysical/physiological phenomena such as illusory contours, bistable percepts, amodal completions, visual imagery, end-stopping, and other effects from beyond the classical receptive field.

2 A Simple Input Encoding Example

We begin by examining a simple input encoding problem. In particular, consider the problem of encoding a collection of inputs $\mathbf{I}_1, \mathbf{I}_2, \ldots$ using a single-layer recurrent network (Figure 1) with linear activation units where the goal is to minimize the reconstruction error. In other words, for a given $n \times 1$ input vector $\mathbf{I}$, we would like to minimize the sum of component-wise squared errors as given by:

$$J(U, \mathbf{r}) = (\mathbf{I} - U\mathbf{r})^T (\mathbf{I} - U\mathbf{r}) \tag{1}$$

where U denotes the $n \times k$ feedback matrix, the rows of which correspond to the synaptic weights of units in the feedback pathway, $\mathbf{r}$ denotes the current $k \times 1$ activation (or internal state) vector and T denotes the transpose operator. The vector $U\mathbf{r}$ corresponds to the network's reconstruction $\mathbf{I}'$ of the input $\mathbf{I}$.

Note that the optimization function J is modulated by two sets of parameters U and $\mathbf{r}$. We can thus minimize J by performing gradient descent with respect to both of these parameters. For a given value of U, we can obtain the optimal estimate $\hat{\mathbf{r}}$ of the state vector $\mathbf{r}$ according to:

$$\dot{\hat{\mathbf{r}}} = -\frac{k_1}{2}\frac{\partial J}{\partial \mathbf{r}} = k_1 U^T(\mathbf{I} - U\hat{\mathbf{r}}) \tag{2}$$

where $k_1 > 0$ determines the adaptation rate for $\mathbf{r}$. The above equation can be equivalently expressed in its discrete form as:

$$\hat{\mathbf{r}}(t+1) = \hat{\mathbf{r}}(t) + k_1 U^T(\mathbf{I} - U\hat{\mathbf{r}}(t)) \tag{3}$$

Similarly, for a given value of $\mathbf{r}$, we can obtain the optimal estimate $\widehat{U}$ of the generative weight matrix U according to:

$$\dot{\widehat{U}} = -\frac{k_2}{2}\frac{\partial J}{\partial U} = k_2(\mathbf{I} - \widehat{U}\mathbf{r})\mathbf{r}^T \tag{4}$$

where $k_2 > 0$ determines the learning rate. Equivalently,

$$\widehat{U}(t+1) = \widehat{U}(t) + k_2(\mathbf{I} - \widehat{U}(t)\mathbf{r})\mathbf{r}^T \tag{5}$$

It is interesting to note that the above rule can be regarded as a form of Hebbian learning, with presynaptic activity $\mathbf{r}$ and postsynaptic activity $(\mathbf{I} - \widehat{U}\mathbf{r})$. The difference in the latter term can be computed, for instance, via inhibitory feedback. Typically, for a given input $\mathbf{I}$ during learning, $\mathbf{r} = \hat{\mathbf{r}}$ is allowed to converge to a stabilized value before the weights $\widehat{U}$ are updated. These new values for $\widehat{U}$ are subsequently used for computing the state $\hat{\mathbf{r}}$ for the next input. The resulting scheme can thus be regarded as implementing an on-line form of the well-known Expectation-Maximization (EM) algorithm [Dempster *et al.*, 1977] (see Section 9.5 for a discussion). Note that for appropriately small positive values of k_1, the quadratic form of J assures convergence of $\hat{\mathbf{r}}$ to the global minimum of J for a given input $\mathbf{I}$.

An important feature of the above estimation scheme is that the dynamics in Equation 3 can be readily implemented in the network of Figure 1 if we make the feedforward weight matrix $W = U^T$. By using a learning rule for U and W^T identical to Equation 5 but with a linear weight decay term, it can shown that $\widehat{W}$ converges approximately to $\widehat{U}^T$ even when initialized to arbitrary random values (see, for example, [Williams, 1985]).

A number of variants of the estimation scheme sketched above have previously appeared in the literature in various guises. For example, [Daugman, 1988] proposes a gradient descent rule for estimating the coefficients of a fixed basis set of non-orthogonal Gabor filters (see also [Pece, 1992]). Daugman's scheme can be seen to be essentially identical to Equation 2 where the rows of U^T comprise the Gabor filter set. On the other hand, rather than fixing the basis set of filters to be Gabor functions, one can attempt to *learn* the basis set for some *fixed* state vector $\mathbf{r}$. In particular, if we let $\mathbf{r} = U^T\mathbf{I}$ (the feedforward response assuming $W = U^T$), Equation 4 reduces

to Oja's subspace network learning algorithm [Oja, 1989] and Williams' symmetric error-correction learning rule [Williams, 1985], both of which were shown to generate orthogonal weight vectors that span the principal subspace (or dominant eigenvector subspace) of the input distribution. Thus, the purely feedforward form of the network performs an operation equivalent to *principal component analysis* (PCA) [Chatfield and Collins, 1980]. More recently, learning schemes similar to the one we presented above have been proposed independently by [Olshausen and Field, 1996] and [Harpur and Prager, 1996]. As above, these schemes combine the dynamics of the state vector $\mathbf{r}$ with the learning of the weights U. By additionally adding an activity penalty (derived from a "sparseness function") to the dynamics of the state vector $\mathbf{r}$, [Olshausen and Field, 1996] demonstrate the development of non-orthogonal and localized basis filters from natural images.

In the following sections, we show that the above framework can be regarded as a special case of a more general estimation and learning scheme that has its roots in stochastic modeling and Kalman filter theory. In particular, using a hierarchical stochastic model of the image generation process, we formulate an optimization function based on the minimum description length (MDL) principle. In the MDL formulation, the activity penalty of [Olshausen and Field, 1996] arises as a direct consequence of the assumed model prior. As we show in the following sections, the resulting learning/estimation scheme allows the modeling of the visual cortex as a hierarchical Kalman predictor.

3 The Architecture of the Model

The model is a hierarchically organized neural network wherein each intermediate level of the hierarchy receives bottom-up information from the preceding level as well as top-down information from a higher level (see Figure 2). At each level, the output signals of spatially adjacent modules, each processing its own local image patch, are combined and fed as input to the next higher level, whose outputs are in turn combined with those of its neighbors and fed into yet another higher level, until the entire visual field has been accounted for. As a result, the receptive fields of units become progressively larger as one ascends the hierarchical network in a manner similar to that observed in the occipitotemporal pathway (V1 $\rightarrow$ V2 $\rightarrow$ V4 $\rightarrow$ IT) of the primate visual system [Van Essen and Maunsell, 1983; Desimone and Ungerleider, 1989]. At the same time, feedback pathways allow top-down influences to modulate the output of noisy lower level modules. From a computational perspective, such an arrangement allows a hierarchy of abstract internal representations to be learned while simultaneously endowing the system with properties essential for dynamic recognition such as the ability to form pattern completions during occlusions. From a neuroanatomical perspective, such an architecture falls well within the known complexity of cortico-cortical and intracortical connections found in the primate visual cortex [Rockland and Pandya, 1979; Gilbert and Wiesel, 1981; Lund, 1981; Jones, 1981; Douglas *et al.*, 1989; Felleman and Van Essen, 1991].

At the lowest level, the input image is processed in parallel by a large number of identical network modules

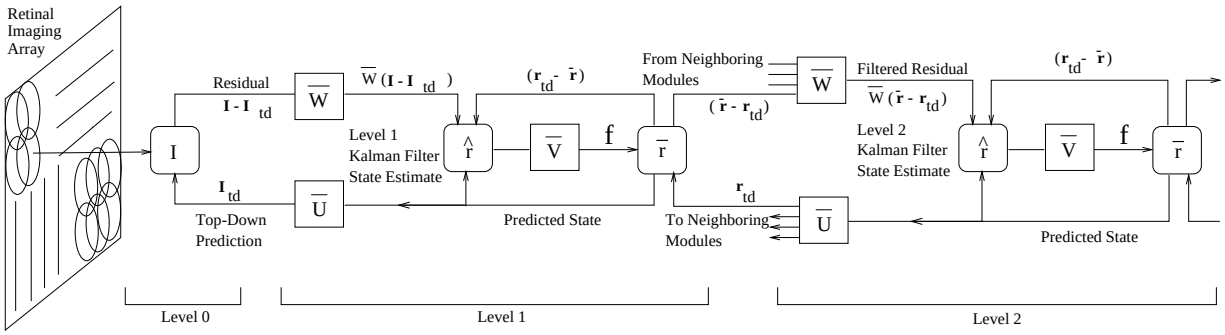

Figure 2: **The Architecture of the Model**. The feedforward pathways for the first two levels are shown in the top half of the figure while the bottom half represents the feedback pathways. Each feedforward weight matrix $\overline{W}$ is approximately the transpose of its corresponding feedback matrix $\overline{U}$. The feedback pathways carry top-down reconstructed signals $\mathbf{r}_{td}$ that serve as predictions for the lower level modules. The feedforward pathways carry residuals between the current state and the top-down predictions, and these residuals are filtered through the feedforward weight matrix $\overline{W}$ at the next level. The current state prediction $\overline{\mathbf{r}}(t)$ is computed from the previous state estimate $\widehat{\mathbf{r}}(t-1)$ using the prediction weights $\overline{V}$ and the activation function f. The current state estimate $\widehat{\mathbf{r}}(t)$ at each level is continually updated by combining the top-down and bottom-up residuals with the current state prediction according to the extended Kalman filter update equations derived in the text. The figure illustrates the architecture for the simple case of four level 1 modules feeding into a level 2 module. However, this arrangement can be easily generalized in a recursive manner to the case of an arbitrary number of lower level modules feeding into a higher level module, whose outputs are in turn combined with those of its neighbors and fed into yet another higher level and so on, until the entire visual field has been covered.

which tessellate the visual field. For a given module, the local image patch can be considered an n-dimensional vector $\mathbf{I}$, which is linearly filtered by a set of k modifiable filters as represented by the rows of a $k \times n$ feedforward matrix $\overline{W}$ of the synaptic weights of units. The synaptic weights comprising $\overline{W}$ are initialized to small random values and adapted in response to input stimuli according to the synaptic learning rules derived below. The image representation $\mathbf{I}$ in our model roughly corresponds to the representation at the LGN while the filters defined by the rows of weight matrix $\overline{W}$ perform an operation similar to that carried out by simple cells in layer IV of V1. The filtered image is defined by a response vector $\mathbf{y}$ given by:

$$\mathbf{y} = \overline{W}\mathbf{I} \tag{6}$$

The response vector $\mathbf{y}$ is fed into k (possibly nonlinear) "state" units which compute running estimates $\widehat{\mathbf{r}}$ of current internal state and predict the state $\overline{\mathbf{r}}$ for the next time instant using their synaptic weights $\overline{V}$ (see Figure 2). There also exist n backprojection or feedback units with an associated $n \times k$ matrix of "generative" weights $\overline{U}$ which allow reconstruction of the input at the lower, less abstract level from the current internal state at the higher, more

abstract level. Their output is given by:

$$\mathbf{I}' = \overline{U}\overline{\mathbf{r}} \tag{7}$$

In addition to bottom-up input $\mathbf{I}$, each level in the hierarchical network (except the highest level) also receives a top-down signal $\mathbf{r}_{td}$ from a higher level module. The signal $\mathbf{r}_{td}$ acts as a prediction of the lower level module's current state vector $\overline{\mathbf{r}}$, as estimated by a higher-level module after taking into account, for example, information from a larger spatial neighborhood in conjunction with the past history of inputs.

Note that while the above equations were defined for the first hierarchical level, where the bottom-up input consisted of the image $\mathbf{I}(t)$, the same analysis can be applied to all levels of the network where the bottom-up input is simply the vector obtained by combining the output vectors of several neighboring lower-level modules as illustrated in Figure 2 for the second level.

Given the above architectural setting, we are now left with the problem of prescribing the dynamics for the internal state vectors $\widehat{\mathbf{r}}$ and $\overline{\mathbf{r}}$, as well as appropriate learning rules for the synaptic weights $\overline{U}$, $\overline{V}$, and $\overline{W}$. In the previous section, we achieved our objective by simply minimizing the reconstruction error. As we shall see, a more general statistical approach is to model the inputs to the system as being stochastically generated by a hierarchical image generation process that can be characterized as follows. First, we assume that images are being generated by a linear imaging model of the form:

$$\mathbf{I}(t) = U(t)\mathbf{r}(t) + \mathbf{n}_{bu}(t) \tag{8}$$

where U is a generative matrix (corresponding to the generative feedback weights in the network) and $\mathbf{r}$ is a hidden state vector. We assume the mean of the bottom-up noise process $E(\mathbf{n}_{bu}(t)) = 0$. The corresponding noise covariance matrix $\mathbf{\Sigma}_{bu}$ at time t is given by

$$E[\mathbf{n}_{bu}(t)\mathbf{n}_{bu}(s)^T] = \mathbf{\Sigma}_{bu}(t)\delta(t,s) \tag{9}$$

where δ is the Kronecker delta function equaling 1 if $t = s$ and 0 otherwise.

We model the difference between the top-down prediction $\mathbf{r}_{td}$ from a higher level and the actual state vector $\mathbf{r}$ at the current level by another stochastic noise process $\mathbf{n}_{td}$:

$$\mathbf{r}_{td}(t) = \mathbf{r}(t) + \mathbf{n}_{td}(t) \tag{10}$$

with $E(\mathbf{n}_{td}(t)) = 0$. The top-down noise covariance matrix $\mathbf{\Sigma}_{td}$ at time t is given by:

$$E[\mathbf{n}_{td}(t)\mathbf{n}_{td}(s)^T] = \mathbf{\Sigma}_{td}(t)\delta(t,s) \tag{11}$$

Given the current state $\mathbf{r}(t)$, the transition to the state $\mathbf{r}(t+1)$ at the next time instant is modeled as:

$$\mathbf{r}(t+1) = f(V(t)\mathbf{r}(t)) + \mathbf{n}(t) \tag{12}$$

where f is a (possibly nonlinear) vector-valued activation function, V is the *state transition matrix* (corresponding to the "prediction" weights in the network), and $\mathbf{n}$ is a stochastic noise process with mean and covariance:

$$E[\mathbf{n}(t)] = \overline{\mathbf{n}}(t) \tag{13}$$

$$E[(\mathbf{n}(t) - \overline{\mathbf{n}}(t))(\mathbf{n}(s) - \overline{\mathbf{n}}(s))^T] = \mathbf{\Sigma}(t)\delta(t,s) \tag{14}$$

For the derivations below, we assume that the three stochastic noise processes introduced above are uncorrelated over time with each other and with the initial state vector.

It is convenient to view the $n \times k$ generative weight matrix U and the $k \times k$ prediction weight matrix V as an $nk \times 1$ vector $\mathbf{u}$ and a $k^2 \times 1$ vector $\mathbf{v}$ respectively, where the vectors are obtained by collapsing the rows of the matrix into a single vector. For example,

$$\mathbf{u} = \begin{bmatrix} U_1^T \\ U_2^T \\ \vdots \\ U_n^T \end{bmatrix} \tag{15}$$

where U_i denotes the ith row of U. We characterize the evolution of the different unknown weights we used in the hierarchical imaging model above by the following dynamic equations:

$$\mathbf{u}(t+1) = \mathbf{u}(t) + \mathbf{n}_u(t) \tag{16}$$

$$\mathbf{v}(t+1) = \mathbf{v}(t) + \mathbf{n}_v(t) \tag{17}$$

where $\mathbf{n}_u$ and $\mathbf{n}_v$ are stochastic noise processes with mean and covariances given by:

$$E[\mathbf{n}_u(t)] = \overline{\mathbf{n}}_u(t) \tag{18}$$

$$E[\mathbf{n}_v(t)] = \overline{\mathbf{n}}_v(t) \tag{19}$$

$$E[(\mathbf{n}_u(t) - \overline{\mathbf{n}}_u(t))(\mathbf{n}_u(s) - \overline{\mathbf{n}}_u(s))^T] = \mathbf{\Sigma}_u(t)\delta(t,s) \tag{20}$$

$$E[(\mathbf{n}_v(t) - \overline{\mathbf{n}}_v(t))(\mathbf{n}_v(s) - \overline{\mathbf{n}}_v(s))^T] = \mathbf{\Sigma}_v(t)\delta(t,s) \tag{21}$$

In summary, Equations 8 through 21 reflect our *assumed hierarchical model* of how visual inputs are being stochastically generated. The goal of a given cortical module is thus to estimate, as closely as possible, its counterpart in the input generation model as characterized by the system of equations above.[1] It does so by minimizing a cost function based on the Minimum Description Length principle.

4 The Minimum Description Length Based Optimization Function

In order to prescribe the dynamics of the network and derive the learning rules for modifying the synaptic weight matrices, we make use of the Minimum Description Length (MDL) principle [Rissanen, 1989; Zemel, 1994], a

[1] By convention, each estimate that the cortical module maintains will be differentiated from its counterpart in the input generation model by the occurrence of the hat symbol. For example, the mean of the current cortical estimate of the actual state vector $\mathbf{r}$ will be denoted by $\widehat{\mathbf{r}}$.

formal information-theoretic version of the well-known Occam's Razor principle commonly attributed to William of Ockham (13th-14th century A.D.) [Li and Vitanyi, 1993]. Briefly, the goal is to encode input data in a way that balances the cost of encoding the data given the use of a model with the cost of specifying the model itself. The underlying motivation behind such an approach is to accurately fit the model to the input data while at the same time avoiding overfitting and allowing generalization.

Suppose that we have already computed a prediction $\overline{\mathbf{r}}$ of the current state $\mathbf{r}$ based on prior data. In particular, let $\overline{\mathbf{r}}$ be the mean of the current state vector *before* measurement of the input data at the current time instant. The corresponding covariance matrix is given by:

$$E[(\mathbf{r} - \overline{\mathbf{r}})(\mathbf{r} - \overline{\mathbf{r}})^T] \quad = \quad M \tag{22}$$

Similarly, let $\overline{\mathbf{u}}$ and $\overline{\mathbf{v}}$ be estimates of the current weights $\mathbf{u}$ and $\mathbf{v}$ calculated from prior data with:

$$E[(\mathbf{u} - \overline{\mathbf{u}})(\mathbf{u} - \overline{\mathbf{u}})^T] \quad = \quad S \tag{23}$$

$$E[(\mathbf{v} - \overline{\mathbf{v}})(\mathbf{v} - \overline{\mathbf{v}})^T] \quad = \quad T \tag{24}$$

Given a description language $\mathcal{L}$, data $\mathcal{D}$ and model parameters $\mathcal{M}$, the MDL principle advocates minimizing the following cost function [Zemel, 1994]:

$$J(\mathcal{M}, \mathcal{D}) = |\mathcal{L}(\mathcal{M}, \mathcal{D})| = |\mathcal{L}(\mathcal{D}|\mathcal{M})| + |\mathcal{L}(\mathcal{M})| \tag{25}$$

$|\cdot|$ denotes length of the description. In our case, $\mathcal{D}$ consists of the current input image $\mathbf{I}(t)$ and top-down input $\mathbf{r}_{td}(t)$, and $\mathcal{M}$ consists of the parameters U, V, and $\mathbf{r}$ that can be modulated.[2] Thus,

$$|\mathcal{L}(\mathcal{D}|\mathcal{M})| = |\mathcal{L}(\mathbf{I} - U\mathbf{r})| + |\mathcal{L}(\mathbf{r}_{td} - \mathbf{r})| \tag{26}$$

The two terms on the right hand side of the preceding equation are simply the bottom-up and top-down reconstruction errors. The cost of the model is given by:

$$|\mathcal{L}(\mathcal{M})| = |\mathcal{L}(\mathbf{r})| + |\mathcal{L}(\mathbf{u})| + |\mathcal{L}(\mathbf{v})| \tag{27}$$

In a Bayesian framework, the above equation can be regarded as taking into account the prior model parameter distributions while Equation 26 represents the negative log-likelihood of the data given the model parameters. Minimizing the complete cost function J is thus equivalent to maximizing the posterior probability of the model $\mathcal{M}$ given data $\mathcal{D}$.

[2]The feedforward weight matrix W does not appear among the model terms since this matrix can be constrained to be the transpose of the feedback weight matrix U as a result of the synaptic learning rules described in Section 6.

Given the true probability distribution (over discrete events) of the various terms in the above equations, the expected length of the optimal code for each term is given by Shannon's *optimal coding theorem* [Shannon, 1948]:

$$|\mathcal{L}(x)| = -\log P(X = x) \tag{28}$$

where $P(X = x)$ denotes the probability of the discrete event x. Since the true distributions are unknown, we appeal to the Central Limit Theorem [Feller, 1968] and use multivariate Gaussians as the prior distributions for coding the various terms above. However, encoding from a continuous distribution requires the calculation of the probability *mass* of a particular small interval of values around a given value for use in Equation 28 above [Nowlan and Hinton, 1992; Zemel, 1994]. Using a trapezoidal approximation, we may estimate the mass under a continuous (in our case, Gaussian) density p in an interval of width w around a value x to be $P(X = x) \cong p(x)w$. For encoding the errors (Equation 26), we assume w to be a constant infinitesimal width which yields (using Equation 28 and ignoring the constant terms due to the coefficients of the multivariate Gaussians):

$$|\mathcal{L}(\mathcal{D}|\mathcal{M})| \quad = \quad (\mathbf{I} - U\mathbf{r})^T \Sigma_{bu}^{-1} (\mathbf{I} - U\mathbf{r}) + (\mathbf{r}_{td} - \mathbf{r})^T \Sigma_{td}^{-1} (\mathbf{r}_{td} - \mathbf{r}) \tag{29}$$

For encoding the model terms (Equation 27), a constant infinitesimal width w may be inappropriate since some values of the parameters may need to be encoded more accurately than others. For example, the network may be more sensitive to small changes in some parameter values than others [Nowlan and Hinton, 1992]. One solution, as suggested by [Nowlan and Hinton, 1992], is to make w inversely proportional to the curvature of the cost function surface, but this may be computationally very expensive. A simpler alternative, which favors small values for the parameters given the choice between small and large values, is to let w be a zero mean radially symmetric Gaussian for each of the model terms $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{r}$ with variances inversely proportional to α, β, and γ respectively. Using Equation 28, the model cost then reduces to (ignoring the constant terms):

$$\begin{aligned} |\mathcal{L}(\mathcal{M})| \quad = \quad & (\mathbf{r} - \overline{\mathbf{r}})^T M^{-1} (\mathbf{r} - \overline{\mathbf{r}}) + (\mathbf{u} - \overline{\mathbf{u}})^T S^{-1} (\mathbf{u} - \overline{\mathbf{u}}) + (\mathbf{v} - \overline{\mathbf{v}})^T T^{-1} (\mathbf{v} - \overline{\mathbf{v}}) + \\ & \gamma \mathbf{r}^T \mathbf{r} + \alpha \mathbf{u}^T \mathbf{u} + \beta \mathbf{v}^T \mathbf{v} \end{aligned} \tag{30}$$

The first three terms in the sum above arise from the prior Gaussian densities for $\mathbf{r}$, $\mathbf{u}$, and $\mathbf{v}$ as given by $G(\overline{\mathbf{r}}, M)$, $G(\overline{\mathbf{u}}, S)$, and $G(\overline{\mathbf{v}}, T)$ while the latter three terms are derived from the zero mean Gaussians associated with w. Note that the above equation implements an intuitively appealing trade-off between achieving a good fit with respect to prior predictions ($\overline{\mathbf{r}}$, $\overline{\mathbf{u}}$ and $\overline{\mathbf{v}}$) and simultaneously penalizing large values for the parameters (activities $\mathbf{r}$ and weights $\mathbf{u}$ and $\mathbf{v}$). One may also view the latter three terms as *regularizers* [Poggio *et al.*, 1985] that help prevent overfitting of data, thereby increasing the potential for generalization.[3]

[3]The variance-related regularization parameters α, β, and γ may be chosen according to a variety of techniques (see [Girosi *et al.*, 1995] for references) but for the experiments in this paper, we used appropriately small but arbitrary fixed values for these parameters.

5 Network Dynamics and Optimal Estimation of Recognition State

The derivation of the estimation algorithm for recognition state is analogous to that for the discrete Kalman filter as given in standard texts such as [Bryson and Ho, 1975] (the continuous case follows in a straightforward fashion by applying a few limits to the difference equations in the discrete case [Bryson and Ho, 1975]). Given new input $\mathbf{I}$ and top-down prediction $\mathbf{r}_{td}$, the optimal estimate of current state $\mathbf{r}$ after measurement of the new data is a stochastic code whose mean $\widehat{\mathbf{r}}$ and covariance P can be obtained by setting:

$$\frac{\partial J(\mathcal{M}, \mathcal{D})}{\partial \mathbf{r}} = 0 \tag{31}$$

which implies

$$M^{-1}(\widehat{\mathbf{r}} - \overline{\mathbf{r}}) - U^T \Sigma_{bu}^{-1}(\mathbf{I} - U\widehat{\mathbf{r}}) - \Sigma_{td}^{-1}(\mathbf{r}_{td} - \widehat{\mathbf{r}}) + \gamma\widehat{\mathbf{r}} = 0 \tag{32}$$

After some algebraic manipulation, the above equation yields the following update rule for the mean of the optimal stochastic code at time t:

$$\widehat{\mathbf{r}}(t) = \overline{\mathbf{r}}(t) + PU^T\Sigma_{bu}^{-1}(\mathbf{I} - U\overline{\mathbf{r}}(t)) + P\Sigma_{td}^{-1}(\mathbf{r}_{td} - \overline{\mathbf{r}}(t)) - \gamma P\overline{\mathbf{r}}(t) \tag{33}$$

where the prediction $\overline{\mathbf{r}}$ is given by:

$$\overline{\mathbf{r}}(t+1) = f(V(t)\widehat{\mathbf{r}}(t)) + \overline{\mathbf{n}}(t) \tag{34}$$

It follows that the covariance matrix P at time t is given by the update rule:

$$P(t) = (M^{-1}(t) + U^T\Sigma_{bu}^{-1}(t)U + \Sigma_{td}^{-1}(t) + \gamma I)^{-1} \tag{35}$$

where I is the $k \times k$ identity matrix. The prediction error covariance matrix M is updated according to:

$$M(t+1) = \frac{\partial f}{\partial \mathbf{r}} P(t) \left(\frac{\partial f}{\partial \mathbf{r}}\right)^T + \Sigma(t) \tag{36}$$

with the partial derivative being evaluated at $\mathbf{r} = \widehat{\mathbf{r}}(t)$. The partial derivatives arise from a first-order Taylor series approximation to the activation function f around $\widehat{\mathbf{r}}(t)$. The above set of equations can be regarded as implementing a form of the *extended Kalman filter* [Maybeck, 1979].

During implementation of the algorithm, the covariance matrices Σ, Σ_{bu}, and Σ_{td} can be approximated by matrices $\widehat{\Sigma}$, $\widehat{\Sigma}_{bu}$, and $\widehat{\Sigma}_{td}$ estimated from sample data. Further, by appealing to a version of the EM algorithm (see Section 9.5), we may use $U = \overline{U}$ and $V = \overline{V}$ in the equations above, where $\overline{U}$ and $\overline{V}$ are learned estimates whose update rules are described in Section 6 and the appendix respectively. For neural implementation, the update equations above can be simplified considerably by noting that the basis filters that form the rows of the feedforward matrix $\overline{W}$ ($\cong \overline{U}^T$; see learning rule below) also approximately decorrelate their input, thereby effectively *diagonalizing* the noise covariance matrices (see also [Pentland, 1992] for related ideas). This allows

the recursive update rules to be implemented locally in an efficient manner through the use of recurrent axon collaterals.

Equation 33 forms the heart of the dynamic state estimation algorithm. Note that Equation 3 which was derived via gradient descent in Section 2 is actually a special case of Equation 33. More importantly, Equation 33 implements an efficient trade-off between information from three different sources: the system prediction $\overline{\mathbf{r}}(t)$, the top-down prediction $\mathbf{r}_{td}$, and the bottom-up data $\mathbf{I}$. Intuitively, the bottom-up and top-down gain matrices $P\overline{U}^T\widehat{\Sigma}_{bu}^{-1}$ and $P\widehat{\Sigma}_{td}^{-1}$ can be interpreted as *signal-to-noise ratios*. Thus, when the bottom-up noise variance $\widehat{\Sigma}_{bu}$ is high (for instance, due to occlusions), the bottom-up term $(\mathbf{I} - \overline{U}\overline{\mathbf{r}}(t))$ is given less weight in the state estimation step (due to a lower gain matrix) and the estimate $\widehat{\mathbf{r}}(t)$ relies more on the top-down term $(\mathbf{r}_{td} - \overline{\mathbf{r}}(t))$ and the system prediction $\overline{\mathbf{r}}(t)$. On the other hand, when the top-down noise variance $\widehat{\Sigma}_{td}$ is high (for instance, due to ambiguity in interpretation by the higher-level modules), the estimate $\widehat{\mathbf{r}}(t)$ relies more heavily on the bottom-up term $(\mathbf{I} - \overline{U}\overline{\mathbf{r}}(t))$ and the system prediction $\overline{\mathbf{r}}(t)$. Finally, if the system prediction $\overline{\mathbf{r}}(t)$ has a large noise variance, the matrix P assumes larger values which implies that the system relies more heavily on the top-down and bottom-up input data rather than on its noisy prediction $\overline{\mathbf{r}}(t)$. The dynamics of the network thus strives to achieve a delicate balance between the current prediction and the inputs from various cortical sources by exploiting the signal-to-noise characteristics of the corresponding input channels. Note that by keeping track of the degree of correlations between units at any given level, the covariance matrices also dictate the degree of *lateral interactions* between the units as determined by Equation 33.

In summary, the optimal estimate of the current state *before measurement* of current data is a stochastic code with mean $\overline{\mathbf{r}}$ and covariance M. After measurement and update as given by Equations 33 and 35, the mean and covariance become $\widehat{\mathbf{r}}$ and P respectively. The mean $\widehat{\mathbf{r}}$ is updated by filtering the current *residuals* or "innovations" [Maybeck, 1979] $(\mathbf{I} - \overline{U}\overline{\mathbf{r}})$ and $(\mathbf{r}_{td} - \overline{\mathbf{r}})$ using $P\overline{U}^T\widehat{\Sigma}_{bu}^{-1}$ and $P\widehat{\Sigma}_{td}^{-1}$ respectively. Note that this requires the existence of the matrix $\overline{U}^T$ for filtering the bottom-up residual $(\mathbf{I} - \overline{U}\overline{\mathbf{r}})$ that is being communicated from the lower level. This computation can be readily implemented if the feedforward weights $\overline{W}$ happen to be symmetric with the feedback weights $\overline{U}$ i.e. $\overline{W} = \overline{U}^T$. Therefore, we need to address the question of how such feedforward and feedback weights can be developed.

6 Learning the Feedforward and Feedback Synaptic Weights

The previous section described one method of minimizing the optimization function $J(\mathcal{M}, \mathcal{D})$: updating the estimate of the current state $\mathbf{r}$ as given by $\widehat{\mathbf{r}}$ and P in response to the input data. However, J can be further minimized by additionally adapting the weights $\mathbf{u}$, $\mathbf{v}$, and the feedforward weights $\mathbf{w}$ as well. Here, we derive the learning rules for $\mathbf{u}$ and $\mathbf{w}$. A possible learning procedure for the prediction weights $\mathbf{v}$ is described in the appendix. Note that in order to achieve stability in the estimation of state $\mathbf{r}$, the weights need to be adapted at a

much slower rate than the dynamic process that is estimating $\mathbf{r}$.

We first derive the weight update rule for estimating the backprojection (generative) weights $\mathbf{u}$. Notice that:

$$(\mathbf{I} - U\mathbf{r}) = (\mathbf{I} - R\mathbf{u}) \tag{37}$$

where R is the $n \times nk$ matrix given by:

$$R = \begin{bmatrix} \mathbf{r}^T & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{r}^T & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{r}^T \end{bmatrix} \tag{38}$$

Differentiating J with respect to the backprojection weights $\mathbf{u}$ and setting the result to 0:

$$\frac{\partial J(\mathcal{M}, \mathcal{D})}{\partial \mathbf{u}} = 0 \tag{39}$$

we obtain, after some algebraic manipulation, the following update rule for the mean of the optimal stochastic weight vector at time t (note that we use $\mathbf{r} = \overline{\mathbf{r}}(t)$ as prescribed by a version of the EM algorithm - see Section 9.5):

$$\widehat{\mathbf{u}}(t) = \overline{\mathbf{u}}(t) + P_u \overline{R}(t)^T \Sigma_{bu}^{-1}(\mathbf{I} - \overline{R}(t)\overline{\mathbf{u}}(t)) - \alpha P_u \overline{\mathbf{u}}(t) \tag{40}$$

where $\overline{\mathbf{u}}(t+1) = \widehat{\mathbf{u}}(t) + \overline{\mathbf{n}}_u(t)$ and $\overline{R}(t)$ is the matrix obtained by replacing $\mathbf{r}$ with $\overline{\mathbf{r}}(t)$ in the definition of R above. The covariance matrices are updated according to:

$$P_u(t) = (S^{-1}(t) + \overline{R}(t)^T \Sigma_{bu}^{-1}(t)\overline{R}(t) + \alpha I)^{-1} \tag{41}$$

$$S(t+1) = P_u(t) + \Sigma_u(t) \tag{42}$$

where I is the $nk \times nk$ identity matrix. Note that Equation 40 is simply a *Hebbian learning rule with decay*, the presynaptic activity being the current responses $\overline{R}(t)$ and the postsynaptic activity being the residual $(\mathbf{I} - \overline{R}(t)\overline{\mathbf{u}}(t))$ (see feedback pathway in Figure 2).

It is relatively easy to see that Equation 5 which we derived via gradient descent in Section 2 is in fact a special case of Equation 40 above. Thus, as we mentioned in Section 2, for the case of feedforward processing, where $\mathbf{r} = \overline{W}\mathbf{I}$ and $\overline{U} = \overline{W}^T$, this learning rule (without the covariance matrices) becomes identical to Williams' symmetric error-correction learning rule [Williams, 1985] and Oja's subspace network learning algorithm [Oja, 1989], both of which perform an operation equivalent to *principal component analysis* (PCA) [Chatfield and Collins, 1980]. The more general form above which additionally incorporates dynamics for the state vector allows the development of non-orthogonal and local representations as well. Another interesting feature of the above learning rule is the transition from $\widehat{\mathbf{u}}(t)$ to $\overline{\mathbf{u}}(t+1)$. This step allows the modeling of intrinsic neuronal noise via the term $\overline{\mathbf{n}}_u(t)$. By assuming a distribution for this noise term (for example, a zero-mean Gaussian distribution)

and sampling from this distribution, the transition from $\hat{\mathbf{u}}(t)$ to $\overline{\mathbf{u}}(t+1)$ can be made stochastic. This helps the learning algorithm avoid local minima and facilitates the search for a global minimum. The weight decay term $-\alpha P_u \overline{\mathbf{u}}(t)$ (due to the model term $|\mathcal{L}(\mathbf{u})|$ in the MDL optimization function) penalizes overfitting of the data and helps increase the potential for generalization.

Having specified a learning rule for the feedback weights, we need to formulate one for the feedforward weights such that the dynamics expressed by Equation 33 is satisfied. In other words, the feedforward weight matrix $\overline{W}$ should be updated so as to approximate the *transpose* of the feedback matrix $\overline{U}$. One solution is to assume that the feedforward weights are initialized to the same random values as the transpose of the feedback weights, and then to apply the rule given in Equation 40. However, it is highly unlikely that biological neurons can have their synapses initialized in such a convenient manner. A more reasonable assumption is that the feedforward weights are initialized *independently* of the feedback weights.

Let $\mathbf{w}$ represent the vectorized form of the matrix W^T analogous to Equation 15. Consider the following weight modification rule for the feedforward weights at time t:

$$\hat{\mathbf{w}}(t) = \overline{\mathbf{w}}(t) + P_u \overline{R}(t)^T \mathbf{\Sigma}_{bu}^{-1} (\mathbf{I} - \overline{R}(t)\overline{\mathbf{u}}(t)) - \alpha P_u \overline{\mathbf{w}}(t) \tag{43}$$

where we use $\overline{\mathbf{w}}(t+1) = \hat{\mathbf{w}}(t) + \overline{\mathbf{n}}_w(t)$ with $\overline{\mathbf{n}}_w(t)$ assumed to be a zero-mean Gaussian noise process. This is again a form of Hebbian learning (with decay), the presynaptic activity this time being the residual image and the postsynaptic activity being the current $\overline{\mathbf{r}}$ (see feedforward pathway in Figure 2). It is relatively easy to show that the above rule causes the feedforward matrix to approach the transpose of the feedback matrix, i.e. $\overline{W} \cong \overline{U}^T$, even when they start from different initial conditions. Define a difference vector $\mathbf{d}$ as:

$$\mathbf{d} = \hat{\mathbf{u}}(t) - \hat{\mathbf{w}}(t) \tag{44}$$

Then, from Equations 40 and 43,

$$\mathbf{d} = (\overline{\mathbf{u}}(t) - \overline{\mathbf{w}}(t)) - \alpha P_u (\overline{\mathbf{u}}(t) - \overline{\mathbf{w}}(t)) \tag{45}$$

Asymptotically, as $\overline{\mathbf{u}}(t) \rightarrow \hat{\mathbf{u}}(t)$ and $\overline{\mathbf{w}}(t) \rightarrow \hat{\mathbf{w}}(t)$, $(\overline{\mathbf{u}}(t) - \overline{\mathbf{w}}(t)) \rightarrow \mathbf{d}$ and Equation 45 reduces to:

$$-\alpha P_u \mathbf{d} \cong 0 \tag{46}$$

Thus, given that $\alpha > 0$ and P_u is positive definite, the expected value of $\mathbf{d}$ approaches zero asymptotically, which implies $\overline{\mathbf{u}} \cong \overline{\mathbf{w}}$. In other words (since $\overline{\mathbf{w}}$ is $\overline{W}^T$ by definition), $\overline{W} \cong \overline{U}^T$ which satisfies the condition needed for the dynamics of the network (Equation 33) to converge to the optimal response vector for a given input.

7 Experimental Results

Before presenting simulation results for the free viewing and fixation experiments, we verify the viability of the model by describing results from two related experiments.

7.1 Experiment 1: Learning and Recognizing Simple Objects

The first experiment was designed to test the model's performance in a simple recognition task. A hierarchical network as shown in Figure 2 was used with four level 1 modules, each processing its local image patch, feeding into a single level 2 module. The four level 1 feedforward modules each contained 10 units (rows of $\overline{W}$) while the second level module contained 25. The constants α and γ that determine the variances of the MDL model terms were set at 0.02 and 0.1 respectively. The feedforward and feedback matrices were trained by exposing the network to grey scale images of five objects as shown in Figure 3 (A). Each input image was of size 128×128 and was split into four equal subimages of size 64×64 that were fed to the four level 1 modules. The average grey level values were subtracted from each image and the resulting image normalized to unit length before being input to the network. For each input training image, the entire network was allowed to stabilize before the weights were updated. Since only static images were used, we used $\overline{\mathbf{r}}(t+1) = \widehat{\mathbf{r}}(t)$. Note that even without a non-linear prediction step, for an arbitrary set of possibly non-orthogonal weight vectors, the optimal response of the network at a given hierarchical level remains a highly non-linear function of the input image that cannot be computed by a single feedforward pass but rather must be recursively estimated by using Equation 33 until stabilization is achieved (cf. [Daugman, 1988]). Figure 3 (B) through (E) illustrate the response of the trained network to various input images.

7.2 Experiment 2: Learning Receptive Fields from Natural Images

The previous section demonstrated the model's ability to learn internal representations for various man-made objects and subsequently use these representations for robust recognition. In this section, we ask the question: what internal representations does the network develop if it is exposed to a sequence of arbitrary natural images? Note that this question becomes especially relevant if the network is to be used as a model of the visual cortical processing during, for example, free viewing of natural images and therefore the internal representations (such as the weights $\overline{W}$) developed by the model will need to be at least qualitatively similar to corresponding representations in the cortex that have been reported in the literature.

We once again employed a hierarchical network as shown in Figure 2 with four level 1 modules feeding into a single level 2 module. The four level 1 feedforward modules each contained 20 units (rows of $\overline{W}$) while the level 2 module contained 25. Eight grey scale images of natural outdoor scenes were used for training the network. However, unlike the previous experiment, the receptive fields of the four level 1 modules were allowed to overlap as shown at the right of Figure 4 in order to qualitatively model the overlapping organization of receptive fields of neighboring cells in the cortex [Hubel, 1988]. Thus, for each natural image, a 20×20 image patch was chosen and four overlapping 16×16 subimages, offset by 4 pixels horizontally and/or vertically from each other, were fed as input to the lower-level modules. Each of the subimages was preprocessed by subtracting the average grey level

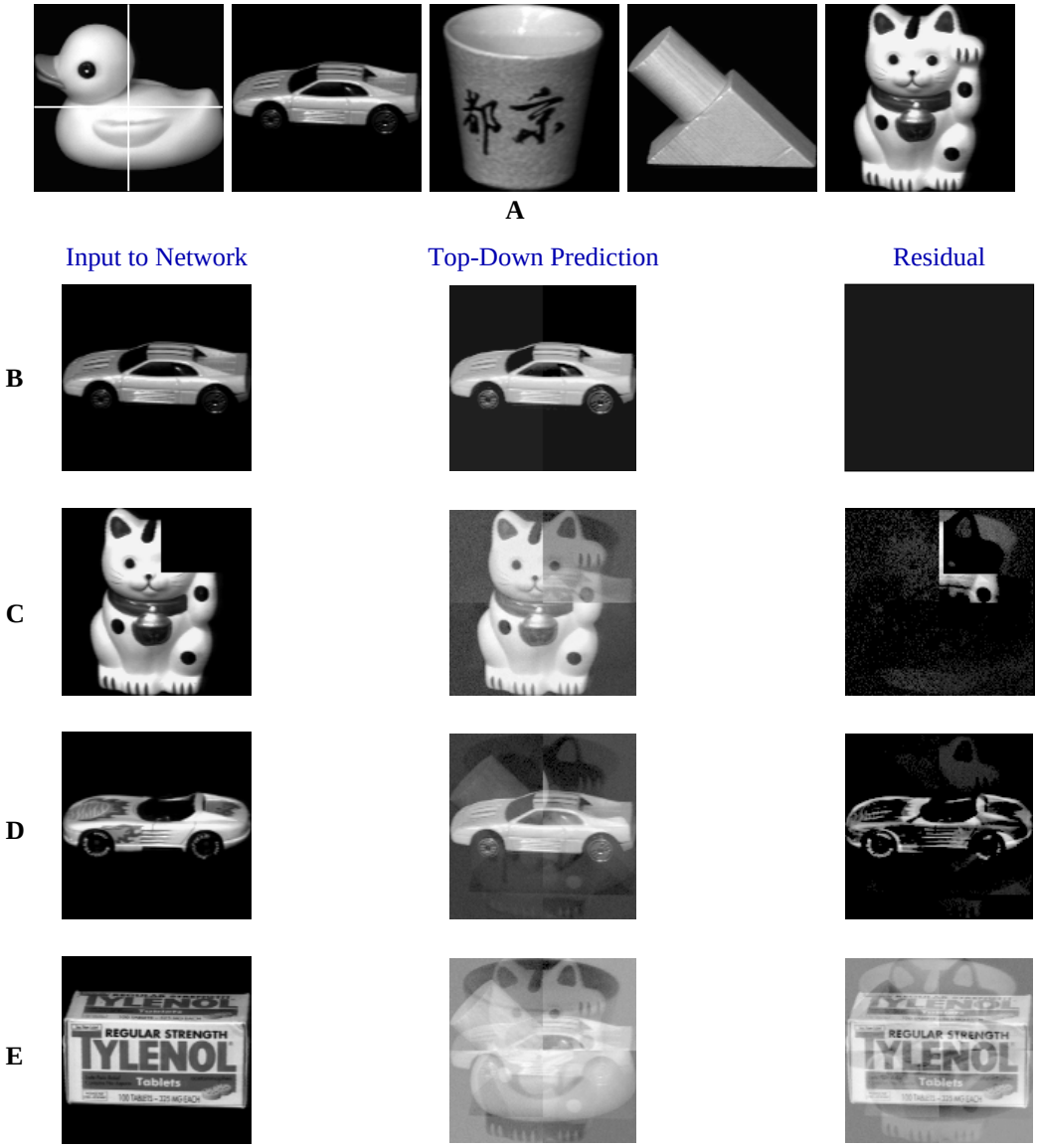

Figure 3: **Results from Experiment 1: Recognition of Simple Objects**. **(A)** The five objects used for training the hierarchical network. The first image additionally shows how a given image was partitioned into four local subimages that were fed to the four corresponding level 1 modules. **(B)** through **(E)** illustrate the response of the trained network to various input images. **(B)** When a training image is input, the network predicts an almost perfect reconstructed image resulting in a residual that is almost everywhere zero, which indicates correct recognition. **(C)** If a partially occluded object from the training set is input, the unoccluded portions of the image together contribute, via the level 2 module, to predict and fill in the missing portions of the input. **(D)** When the network is presented with an object that is highly similar to a trained object, the prediction is that of the closest resembling object (the car in the training set). However, the large residual allows the network to judge the input as a *new object* rather than classifying it as the training object and incurring a false positive error, an occurrence that is common in most purely feedforward systems. **(E)** shows that a completely novel object results in a prediction image that is an arbitrary mixture of the training images and as such, generates large residuals. The system can then choose to either learn the new object using Equations 40 and 43, or ignore the object if it happens to be behaviorally irrelevant.

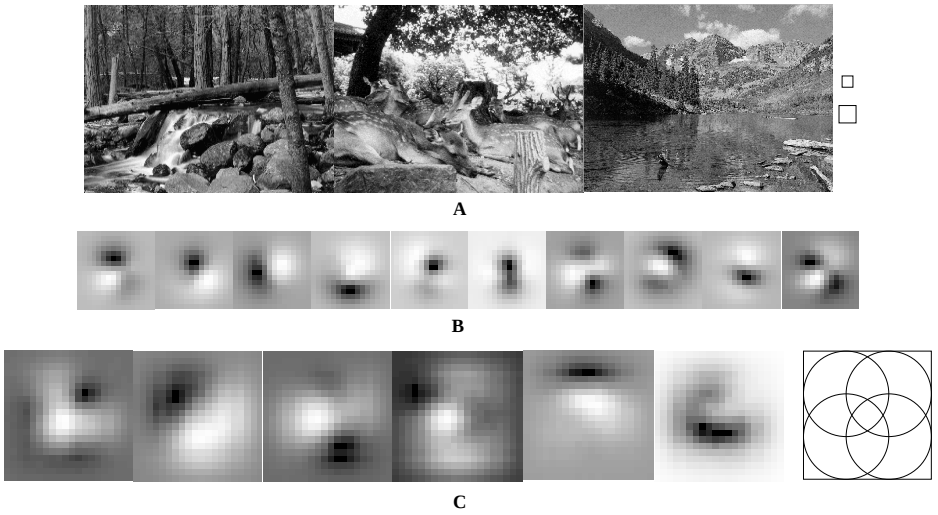

Figure 4: **Results from Experiment 2: Learning Receptive Fields from Natural Images**. **(A)** Three of the eight images used in training the weights. The boxes on the right show the size of the first and second-level receptive fields respectively on the same scale as the images. Four overlapping Gaussian windowed image patches extracted from the natural image were fed to the level 1 modules as shown at the extreme left in Figure 2. The responses from the four level 1 modules were in turn combined in a single vector and fed to the level 2 module. The effective level 2 receptive field thus encompasses the image region spanned by the four overlapping level 1 receptive fields as shown on the right in (C). **(B)** Ten of the twenty receptive fields (or spatial filters) as given by the rows of $\overline{W}$ in a level 1 module. These resemble classical oriented edge/bar detectors, which have been previously modeled as difference of offset Gaussians or Gabor functions. **(C)** Six of twenty receptive fields as given by the rows of $\overline{W}$ at level 2. These show non-linearities that result from higher-order correlations captured by the level 2 units.

value from each pixel to approximate the removal of DC bias at the level of the LGN. The resulting subimage was windowed by a 16×16 Gaussian in order to prevent the spatial frequency biases introduced by a rectangular image sampling lattice. A total of 12114 image patches were used for training the hierarchical network. For each image patch, the network was allowed to converge to the different optimal response vectors $\widehat{\mathbf{r}}$ for the different modules at each level before the corresponding weights ($\overline{W} = \overline{U}^T$) were updated. The variance-dependent constants α and γ were set to 0.02 and 0.1 respectively and the temporal prediction step was simply $\overline{\mathbf{r}}(t+1) = \widehat{\mathbf{r}}(t)$ as in the previous section.

Figure 4 shows feedforward synaptic weights $\overline{W}$ for the level 1 (equivalent to V1) and level 2 (V2) units that were learned using Equation 43. The level 1 receptive fields resemble non-orthogonal wavelet-like edge/bar detectors at different orientations similar to the receptive fields of simple cells in V1 [Hubel and Wiesel, 1962; 1968; Palmer *et al.*, 1991] (see also [Olshausen and Field, 1996; Harpur and Prager, 1996; Bell and Sejnowski, 1996]). These have previously been modeled as 2D Gabor functions [Daugman, 1980; Marcelja, 1980] or difference of

Gaussian operators [Young, 1985]. The level 2 receptive fields were constructed by propagating a unit impulse response for each row of the weight matrix $\overline{W}$ at the second level to level 0. These receptive fields appear to be tuned towards more complex shapes such as corners, curves, and localized edges. The existence of cells coding for features more complex than simple edges and bars have also been reported in several electrophysiological studies of V2 [Hubel and Wiesel, 1965; Baizer *et al.*, 1977; Zeki, 1978]. Thus, both the first and second level units in the model develop internal representations that are comparable to those of their counterparts in the primate visual cortex.

8 Simulations of Free Viewing and Fixation Experiments

The experimental results from the previous section suggest that the model may provide a useful platform for studying cortical function by (a) demonstrating the model's efficacy in a simple visual recognition task and (b) showing that the learning rules employed by the model are capable of developing receptive fields that resemble cortical cell receptive field weighting profiles. In this section, we return to the question posed in the introduction: why do the responses of some visual cortical neurons differ so drastically when the same image regions enter the neuron's receptive field in free viewing and fixation conditions [Gallant *et al.*, 1994; 1995; Gallant, 1996]?

A possible answer is suggested by the recognition results in Figure 3 (C). In particular, the figure shows that the responses of neurons depend not only on the image in the cell's receptive field but also on the images in the receptive fields of neighboring cells. This suggests that a cell's response may be modulated to a considerable extent by the surrounding context in which the cell's stimulus appears, due to the presence of top-down feedback in conjunction with lateral interactions mediated by the covariance matrices P, $\widehat{\Sigma}_{bu}^{-1}$, and $\widehat{\Sigma}_{td}^{-1}$ (see Equation 33).

For comparison with Gallant *et al.*'s electrophysiological recording results, we assumed that the cell responses recorded correspond to the residual $(\overline{\mathbf{r}} - \mathbf{r}_{td})$ in the model (see Figure 2).[4] Qualitatively similar results are obtained by measuring the filtered bottom-up residual. For the simulation experiments, we modeled cells in V1 and V2 as shown in Figure 2 (with four level 1 modules feeding into a single level 2 module) but the model can be readily extended to more abstract higher visual areas with a higher degree of convergence at each level. The parameters were set to the same values as in Section 7.2. The feedforward and feedback weights were set equal to the values

[4] The primate visual cortex is a laminar structure that can be divided into six layers [Felleman and Van Essen, 1991]. The supragranular pyramidal cells (e.g. layer III) typically project to layer IV of the next higher visual area while infragranular pyramidals (e.g. layer V/VI) send axons back to layers I, V and/or VI of the preceding lower area [Van Essen, 1985; Felleman and Van Essen, 1991]. The architecture of the model (Figure 2) suggests that the supragranular cells may carry the residual $(\overline{\mathbf{r}} - \mathbf{r}_{td})$ to the next level while the infragranular cell axons could convey the current top-down prediction $\overline{U}\overline{\mathbf{r}} (= \mathbf{I}_{td}$ or $\mathbf{r}_{td})$ to the lower level. The bottom-up residual $(\mathbf{I} - \overline{U}\overline{\mathbf{r}})$ would then be filtered by layer IV cells (whose synapses encode $\overline{W}$) with the help of interneurons which implement the lateral interactions due to the covariances. The estimate $\widehat{\mathbf{r}}(t)$ would presumably be computed by infragranular cells (whose synapses could encode $\overline{V}$) during the course of predicting $\overline{\mathbf{r}}(t+1)$; the presence of recurrent axon collaterals [Lund, 1981] among these cells is especially suggestive of such a recursive computation. These functional interpretations of cortical circuitry are however speculative and require rigorous experimental confirmation.

learned during prior exposure to natural images as described in Section 7.2. The free viewing experiments were simulated by making random "eye-movements" to image patches within a given natural image and allowing the entire network to converge till all cell responses stabilized; the response values (as given by the residual) were then recorded. For determining the response of a cell during the fixation task, only the image region that fell in the cell's receptive field during free viewing was displayed and the cell's response noted. The cell responses recorded were numerical quantities that could be used to generate spikes in a more detailed neuron model. Since the eye-movements were assumed to be random, the temporal prediction step involving V was not used in these simulations but we nevertheless expect it to play an important role in future simulations that employ temporal as well as spatial context in predicting stimuli.

Figure 5 shows the simulation results. As noted above, the response of a given cell at any level is determined not just by the image falling within its receptive field but also by the images in the receptive fields of neighboring cells since the neighboring modules feed into the higher level which conveys the appropriate top-down prediction back to the lower level: the better the top-down prediction $\mathbf{r}_{td}$, the smaller the residual $(\overline{\mathbf{r}} - \mathbf{r}_{td})$ and hence, the smaller the response. In the case of free viewing an image, the neighboring modules play an active role in contributing to the top-down prediction for any cell and the existence of this larger spatial context allows for smaller responses in general. This is shown in Figure 5 (A) where the given cell's response initially increases until appropriate top-down signals cause the response to diminish to a stabilized level. On the other hand, in the fixation task, only the image subregion that fell in a cell's receptive field during free viewing is displayed on a blank screen. The absence of any contextual information results in a much less accurate top-down prediction due to the lack of contributions from neighboring modules and hence, a relatively larger response is elicited as illustrated in Figure 5 (B). The histograms of responses in the simulations of fixation and free viewing conditions are shown in Figure 5 (C). The results show that in most but not all cases the responses in free viewing are substantially reduced. This is consistent with the free viewing data [Gallant, 1996] and the data from accompanying control experiments (see below). Perhaps the most interesting feature of the free viewing versus fixation histograms is that they overlap, suggesting that the reduction in responses is not a binary all or none phenomenon but rather one that is dependent on the specific ability of the higher level units to predict the current state at the lower level.

9 Discussion

In related experiments [Gallant *et al.*, 1995], it was observed that responses were sometimes also suppressed by enlarging the stimulus beyond the cell's receptive field in the fixation task. More recently, Gallant has shown that in many cases, the response attenuation observed during free viewing can be duplicated in a static fixation task in which a review "movie" recreating the spatiotemporal sequence of image patches that entered the receptive field during free viewing is presented to the monkey, the review stimuli being three times the size of

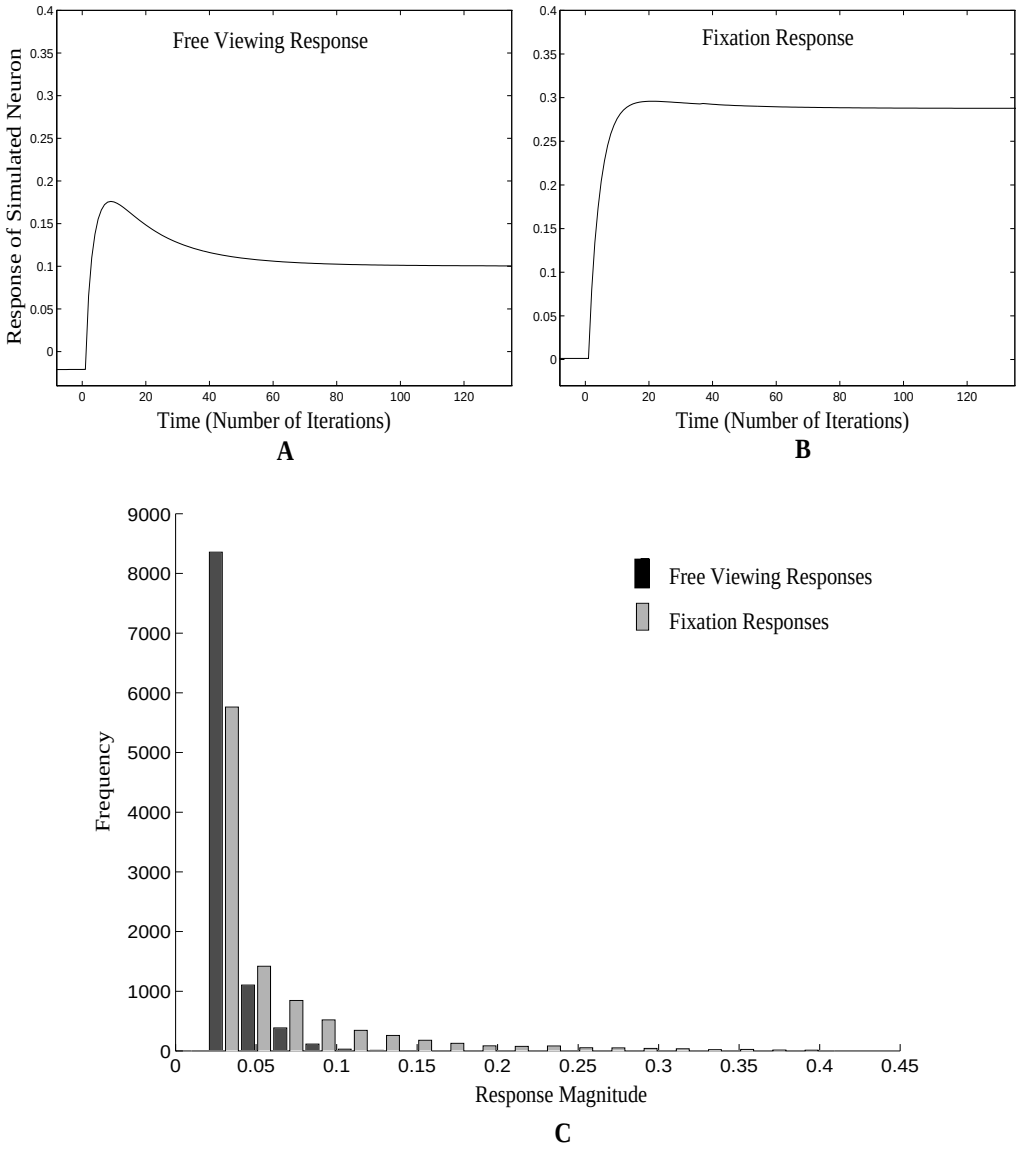

Figure 5: **Free Viewing and Fixation Simulation Results**. **(A)** An example of the dynamic response of a cell to an image patch in the simulated case of free viewing. **(B)** The response of the same cell to the same image patch but under conditions of fixation. **(C)** The histograms of responses accumulated over 20 cells in a first level module for 500 image presentations, showing the fixation responses (light grey) compared to the free viewing responses (black). Responses obtained during free viewing were found to be in general much diminished than those obtained during the fixation task. As the graph depicts, more than 80% of the free viewing responses lie within the first bin in the histogram; the fixation distribution, on the other hand, has a much longer tail than the free viewing distribution.

the classical receptive field [Gallant, 1996]. The results from both of these experiments, which do not involve any eye movements, lend support to our explanation of the suppression of responses as occurring largely due to contributions from neighboring modules in response to the global input stimuli, rather than being the result of oculomotor predictive mechanisms (such as the "shifter circuit" of [Andersen and Van Essen, 1987]). However, this does not preclude the possibility of suppression due to oculomotor predictions occurring in other visual cortical areas, especially those in the parietal cortex, which are connected to oculomotor structures and where such predictive remapping effects have been observed [Duhamel *et al.*, 1992].

A number of neurophysiological effects due to stimuli from beyond the classical receptive field can be explained within the context of the present model as occurring due to the influence of top-down predictions. Take, for example, the classical phenomenon of *end-stopping*, or suppression of responses when the stimulus (e.g. a line) extends beyond the cell's receptive field [Hubel and Wiesel, 1965; 1968]. This is clearly a highly "nonlinear" property of the "hypercomplex" cell [Hubel and Wiesel, 1965] if one considers the cell in isolation; however, the model suggests that such "nonlinearities" may in fact arise due to cortico-cortical/geniculo-cortical and intracortical interactions between linear cells as well. An example of this phenomenon is Figure 3 (C) where a neural unit participates in the completion of an occluded image: the unit's response appears to be a highly non-linear function of the input image patch even though its activation function is linear. In the case of end-stopping, the present model predicts that the ends are "stopped" via inhibitory feedback whenever the surrounding and encompassing receptive fields can predict the cell's local input. The behavior of the cell appears non-linear only if the cell is viewed in isolation without regard to possible interactions between its neighbors within a level and across levels (as suggested by the covariance terms and top-down feedback terms respectively in Equation 33).[5] Indeed, experiments by [Murphy and Sillito, 1987] confirm the contribution of cortical feedback in mediating end-inhibition in cells in the dorsal LGN. In addition, Bolz and Gilbert have shown that supragranular cells in the cat striate cortex (V1) lose the property of end-stopping when infragranular cells (in particular, cells in layer VI) are inactivated by using the inhibitory transmitter GABA [Bolz and Gilbert, 1986]. Since infragranular cells in V1 receive feedback from higher extrastriate visual cortical areas [Van Essen, 1985; Felleman and Van Essen, 1991], it is reasonable to assume that inactivation of these cells prevents top-down feedback from influencing the supragranular cells, thereby causing them to lose their property of end-stopping. Note that other properties of the supragranular cells such as orientation selectivity were found to remain intact, which is in agreement with our model. Further evidence for the fundamental role of top-down feedback in

[5] Note that this does not preclude the existence of non-linear neurons in the cortex. Non-linearities (within the dynamic range of a cortical neuron) are accounted for in two ways by the present model: (a) single units in the model code both positive and negative quantities whereas the cortex may employ multiple neurons for this purpose, thereby causing the appearance of a non-linearity; (b) the model allows for neurons with non-linear activation functions in the prediction step (Equation 34) at each hierarchical level.

influencing neural activity at lower cortical levels is provided by the work of [Mignard and Malpeli, 1991] who show that cells in the upper layers of V1 can be driven in the absence of direct bottom-up input from the LGN. However, additionally destroying V2 profoundly reduced the upper layer neural activity which led [Mignard and Malpeli, 1991] to conclude that the cause of the upper layer activity was in fact top-down feedback from V2. Given the above observations, it may be possible to attribute other types of "non-specific suppression" in V1 cells (such as *cross-orientation inhibition* [DeAngelis *et al.*, 1992]) to bottom-up/top-down cortico-cortical and lateral interactions such as those occurring in Equation 33 rather than to purely normalization-based suppression mechanisms [Heeger *et al.*, 1996]. Also, in the light of the above discussion, the model predicts that much less response suppression should be observed in Gallant *et al.*'s free-viewing experiments if top-down feedback is precluded, for instance, by intracortical pharmacological means as in the experiments of [Bolz and Gilbert, 1986] or via destruction of higher cortical areas as in the experiments of [Mignard and Malpeli, 1991].

Suppression of neural responses has also been observed in higher visual cortical areas such as MT [Allman *et al.*, 1985], V4 [Desimone and Schein, 1987; Muller *et al.*, 1996] and IT [Miller *et al.*, 1991]. Many of these experiments employ time-varying stimuli such as drifting gratings and we would therefore expect the responses to be modulated with respect to not just spatial context (as in this paper) but also recent temporal context. This would necessitate learning the prediction matrix $\overline{V}$ (a possible method is suggested in the appendix) and using it in conjunction with an appropriate activation function f for predicting spatiotemporal inputs as given by Equation 34. Simulation of such spatiotemporal cortical models constitutes an active direction of future research.

9.1 Free Viewing of Natural Scenes by Humans

An interesting question is whether the model conforms to behavioral data obtained from human subjects during free viewing of natural images. Experiments by McConkie and others [McConkie, 1991; Grimes and McConkie, 1995] appear to be especially relevant in answering this equation. In these experiments, subjects were given the primary task of remembering an image but they were also informed that changes might be introduced in the image as they examine it. They were asked to press a button when they detected a change. During a given trial, parts of the image was changed *during* certain saccades. Surprisingly, the subjects were remarkably unaware of such prominent changes as the addition or deletion of objects, and changes in color.

The remarkable insensitivity of human subjects to changes in the image during "free viewing" might appear to contradict the present model since, after all, a change in the image from a previous fixation should introduce a large residual[6] and hence direct the subject's attention to it. Recent experiments in our laboratory [Ballard *et al.*, 1996] and in McConkie's laboratory [Irwin *et al.*, 1994] in fact do corroborate this expectation: subject's

[6]This observation applies to models employing temporal context to predict post-saccadic stimuli whereas the simulations described in this paper used spatial context only.

do indeed become aware of the changes *if* the object that was subject to a change was relevant to the task at hand. For example, [Ballard *et al.*, 1996] report an increase in fixation time of subjects whenever task-relevant changes are introduced during the course of a block copying task. In addition, neurophysiological experiments by [Miller *et al.*, 1991] indicate that responses in IT neurons in the monkey are attenuated according to whether or not the presented stimulus matches a previously memorized object, yielding large responses (residuals) during mismatches and response suppression (small residuals) during matches. Indeed, the theoretical framework of Crick and Koch (see, for example, [Koch, 1996]) proposes that visual awareness may be correlated with responses in higher visual cortical areas such as V4 and IT, but probably not lower level areas such as V1. This supports our argument that high residuals in lower level areas, especially when in image regions not related to the task at hand, need not always imply awareness of the corresponding changes as in McConkie's experiments.

9.2 Illusory Contours, Amodal Completions, Visual Imagery, and Bi-Stable Percepts

We have shown that the ability of adjacent modules to contribute to the output of a neighbor (via feedback from the next level) endows the model with important properties such as the ability to form completions in the presence of occlusions (for example, see Figure 3). The same property also allows alternate interpretations of phenomena such as *subjective/illusory contours* and *amodal completions* [Kanizsa, 1990]. For example, [Grosof *et al.*, 1992] and [von der Heydt *et al.*, 1984] report the existence of neurons responding to illusory contours in V1 and V2 respectively. The present model suggests that these responses may be the result of expectation-based top-down feedback as illustrated by Figure 3 (C). In this regard, the existence of overlap between adjacent receptive fields considerably helps in providing more accurate completions as compared to non-overlapping receptive fields. The feedback pathways in the model also provide a convenient substrate for realizing *visual imagery* as illustrated by the top-down reconstructed images in Figure 3. Indeed, recent PET studies have shown that imagery shares some of the same neural substrates as perception and can activate even the lowest visual areas [Kosslyn *et al.*, 1995].

Equation 34 allows the effects of intrinsic neuronal noise to be modeled via the additive noise term $\overline{\mathbf{n}}(t)$. By assuming a probability distribution for this term and sampling from this distribution in Equation 34, the computation of the recognition state estimates can be made stochastic. An alternate but equally plausible method to make the model stochastic is to directly sample the current recognition state from a Gaussian distribution with mean $\overline{\mathbf{r}}$ and covariance M. Either of these settings could conceivably allow modeling of bistable percepts and Gestalt inversions such as *Rubin's vase* or *Necker's cube*. Such phenomena could be attributed to a lack of convergence at the highest level to a stable object hypothesis due to large noise variances in the top-down component of the architecture. The stochastic nature of the model would in turn allow periodic transitions between two alternate interpretations of the same visual image (local minima in recognition state space). Evaluating the efficacy of these ideas remains the subject of ongoing simulations.

9.3 Comparison with Related Work on Cortical Modeling

There has been considerable work in recent years on determining the computational nature of the cortex and the brain [Churchland and Sejnowski, 1992]. This work includes models ranging from feedforward networks such as the hierarchical "perceptual network" of [Linsker, 1988] and the "HBF" networks of Poggio and collaborators [Poggio, 1990; Poggio *et al.*, 1992] to bi-directional "flow"/control-based models, such as Ullman's Counterstreams model [Ullman, 1994], Deacon's "counter-current diffusion" model [Deacon, 1989] and Van Essen *et al.*'s Dynamic Routing Circuit model [Van Essen *et al.*, 1994]. While an exhaustive survey is beyond the scope of this paper, we briefly outline below comparisons of the present model with some closely related models that employ both feedforward as well as feedback connections for perception and learning.

Perhaps the earliest use of feedback mechanisms as integral components of perception can be found in the classic paper by [Pitts and McCulloch, 1947] in which a negative feedback model of oculomotor control is proposed. Another early model in which feedback plays a central role is that of [MacKay, 1956] who describes an automaton that generates "imitative" internal responses to adaptively match incoming signals. The present model can be regarded as a concrete stochastic solution to the "epistemological problem" that is the subject of MacKay's work. Feedback also plays an important role in several more recent theories such as Carpenter and Grossberg's adaptive resonance theory [Carpenter and Grossberg, 1987], Edelman's "re-entrant" signaling theory [Edelman, 1978], Fukushima's extended Neocognitron model [Fukushima, 1988], Harth *et al.*'s "Alopex" optimization model [Harth *et al.*, 1987], Mumford's model of the cortex [Mumford, 1994] based on Grenander's Pattern Theory [Grenander, 1976-81], and Rolls' theories on the hippocampus and cortical backprojections [Rolls, 1989], all of which utilize feedback connections to instantiate *forward models* [Jordan and Rumelhart, 1992] of the input data but differ in the way top-down feedback is compared and integrated with bottom-up input. Of these, Mumford's model is closest in spirit to the ideas presented in this paper. However, iterative relaxation schemes such as those proposed by Mumford and others have been previously dismissed as models of perception due to the large number of iterations that may be required to assure convergence, a fact that contradicts the rapidity of recognition of static stimuli in humans and primates [Thorpe and Imbert, 1989; Oram and Perrett, 1992; Tovee *et al.*, 1993]. The model proposed in this paper avoids the pitfalls of steepest descent relaxation algorithms by employing prediction weights $\overline{V}$ and approximate *inverse models* (as given by $\overline{W}$) in the feedforward pathway that allow rapid (though perhaps crude and not always correct) one-shot estimates of static stimuli that are further refined by the state estimation algorithm if necessary.

Inverse models also play a crucial role in the forward/inverse optics model of [Kawato *et al.*, 1993] and in the Helmholtz machine [Dayan *et al.*, 1995; Hinton *et al.*, 1995], both of which share considerable similarities with the model proposed in this paper and both of which inspired many of the choices made in the present model.

The Kawato *et al.* model minimizes the sum of the reconstruction error and a regularizer term incorporating a priori knowledge about the visual world such as smoothness of representations. This is similar to the MDL optimization function that we use, but without the stochastic perspective. The Helmholtz machine [Dayan *et al.*, 1995] is a hierarchical network that builds a stochastic generative model by using an inverse ("recognition") model to approximate the true posterior distribution of the input data. For modifying the feedback (generative) and feedforward (recognition) weights of a given stochastic Helmholtz machine, [Hinton *et al.*, 1995] propose an elegant learning method known as the "Wake-Sleep" algorithm, which is derived using the MDL principle. While the derivation of the algorithm necessitates some approximations and simplifications, which may sometimes prevent it from converging to the correct posterior distribution, experimental results using the algorithm have been positive [Hinton *et al.*, 1995]. In this regard, the non-linear form of the model presented in this paper also uses an approximation, namely, a first-order Taylor series approximation to the non-linear activation function f. A possible weakness of the Helmholtz machine, as stated by [Dayan *et al.*, 1995], is that recognition in the hierarchical network is a purely bottom-up feedforward process; the generative weights are superfluous and play no role in mediating perception. In addition, there are no lateral interactions between units at a given level.[7] The model presented in this paper allows for relatively complex dynamic interactions to occur between top-down and bottom-up signals in addition to some lateral interactions as given by Equation 33 among the units at each level. Such top-down/bottom-up interactions have been shown to occur in a wide variety of visual tasks in humans (see [Churchland *et al.*, 1994] for a review). Another issue not directly addressed by many of the models above is that of modeling the dynamic nature of vision and the role of prediction in visual processing (see [Softky, 1996] for some arguments regarding the necessity of prediction during perception). The model presented herein on the other hand is essentially a hierarchical predictor/estimator that continually refines its recognition estimates at each level, allowing the visual system to anticipate incoming stimuli.

9.4 Comparison with Related Work on Kalman Filters

During the past decade, Kalman filters have been used in computer vision and robotics for tackling a wide variety of problems ranging from motion estimation to contour tracking [Hallam, 1983; Broida and Chellappa, 1986; Ayache and Faugeras, 1986; Matthies *et al.*, 1989; Blake and Yuille, 1992; Dickmanns and Mysliwetz, 1992; Pentland, 1992]. Much of this work crucially hinges on the ability to formulate accurate dynamic/physical models of the object properties being estimated. The formulation of such hand-coded models however becomes increasingly difficult in more complex dynamic environments. A crucial difference between the present approach and the approach predominant in much of the Kalman filter literature is that rather than being concerned solely with the estimation of state for a *predefined* model, we suggest that Kalman filter based estimation algorithms

[7] However, see [Dayan and Hinton, 1996] for several interesting variants of the Helmholtz machine that address some of these issues.

may also be used for simultaneously *learning* the feedforward, feedback, and prediction models ("measurement" and "system" models in Kalman filter terminology) as given by their respective weight matrices. It remains to be seen whether these learned models perform favorably in environments where hand-coded models have proved to be unsatisfactory. Another important difference is the use of a hierarchical form of the Kalman filter. Our formulation in this regard is in many ways similar to the recent proposals of [Chou *et al.*, 1994a; 1994b; Luettgen and Willsky, 1995] who describe an optimal estimation algorithm for a class of multiscale dynamic models. These dynamic models are similar in spirit to the one we described in Section 3 albeit without the temporal prediction step. In particular, Chou *et al.* exploit the analogy between time and scale, and define *scale-recursive* linear dynamic models evolving on dyadic trees:

$$\mathbf{r}(s) = A(s)\mathbf{r}(s-1) + B(s)\mathbf{w}(s) \tag{47}$$

$$\mathbf{I}(s) = C(s)\mathbf{r}(s) + \mathbf{v}(s) \tag{48}$$

where s denotes the current scale on the dyadic tree (highest values denote the leaves), and **w** and **v** are zero-mean white noise processes. This formulation leads to an elegant estimation algorithm recursive in *scale*, consisting of an upward sweep in which information in a subtree is fused in a fine-to-coarse fashion, followed by a downward sweep in which information is spread back through the tree. For the simple linear case without the temporal prediction step, the iterative algorithm presented in this paper can be seen to converge to approximately the same solution as that found by the upward/downward sweep algorithm proposed by Chou *et al.*.[8]

The equations used by Chou *et al.* above can be seen to be closely related to the following two equations (when defined on a dyadic tree) that formed part of the more general stochastic model described in Section 3:

$$\mathbf{r}_{td}(s,t) = \mathbf{r}(s,t) + \mathbf{n}_{td}(s,t) \tag{49}$$

$$\mathbf{I}(s,t) = U(s,t)\mathbf{r}(s,t) + \mathbf{n}_{bu}(s,t) \tag{50}$$

where we may use $\mathbf{r}_{td}(s,t) = U^i(s-1,t)\mathbf{r}(s-1,t)$ if $\mathbf{r}_{td}$ happens to be information from the next higher level and U^i is the part of the higher level weight matrix U that generates inputs for the given lower level module (in Figure 2, for example, the matrix U^i for the first level module corresponds to the top one-fourth rows of the second level matrix U as depicted in the figure). The correspondence between the various terms in the two sets of equations (47 and 49, 48 and 50) is quite striking. However, a few important differences are also evident. While the linear dynamic system model in Chou *et al.*'s framework is used only to relate transitions from one scale to the next, the model used in this paper can be seen to be recursive in both scale and time. The recursion in scale is however implicit since information from another scale is handled in a modular fashion simply as input from a

[8] We thank Peter Dayan for pointing out this equivalence.

separate source (as in Equations 10 and 49). The corresponding estimation algorithm is thus recursive in *time* as in the standard Kalman filter and allows for prediction in the temporal domain at each hierarchical level. As stressed previously, the ability to predict is crucial for perception and action. Being strictly recursive in scale, the Chou *et al.* algorithm requires that covariance matrix information be propagated across hierarchical levels. In contrast, the model presented here maintains covariance information locally at each level. In a biological setting, the computation of covariance information may be a limiting step and is perhaps best kept local. A further advantage of our formulation is that it lends itself naturally to arbitrary varieties of *modularization* by allowing estimates of a given state variable from disparate sources to be integrated according to their signal-to-noise ratios (this paper illustrates the example of top-down and bottom-up sources). A direct consequence of such a formulation is that one can model arbitrary connections among different (but related) cortical areas *which do not necessarily have to be related in a strictly multiscale, hierarchical fashion* as required by the framework of Chou *et al.* A final but important difference, as mentioned in the previous paragraph, is the prominent role accorded to learning in our approach, a subject not addressed by Chou *et al.*

9.5 Comparison with Related Work on Learning and Parameter Estimation

The subject of learning synaptic weights that resemble receptive field weighting profiles in the visual cortex has received considerable attention within the neural modeling community. Early work showed that receptive fields resembling those of simple cells in the striate cortex could be learned from natural images using algorithms ranging from competitive learning [Barrow, 1987] to principal component analysis [Hancock *et al.*, 1992]. However, these fell short of providing a complete characterization due to limitations such as proclivity towards producing global and/or orthogonal receptive fields. More recently, algorithms that produce better characterizations of cortical receptive fields have been proposed independently by [Olshausen and Field, 1996] and [Harpur and Prager, 1996] (see also [Bell and Sejnowski, 1996]), building on previous work by [Daugman, 1988] and [Pece, 1992]. As alluded to in Section 2, these algorithms can be seen to be special cases of the learning rule expressed in Equation 40. In particular, the "sparseness" function of [Olshausen and Field, 1996] corresponds to the MDL model prior term $|\mathcal{L}(\mathbf{r})|$ in Equation 27. Using different encoding distributions gives rise to different degrees of sparseness and localization in the internal representations. For example, [Olshausen and Field, 1996] show how a variety of non-linear sparseness functions can be used to generate localized receptive fields.[9] In our case, the Gaussian distribution assumed when evaluating $|\mathcal{L}(\mathbf{r})|$ leads to a linear activity penalty that favors distributed codes; localization in our model is provided by the already existing cortical topology of hierarchical, overlapping receptive fields, whose size at each level is predetermined by the local dendritic field. Since the learning algorithms of [Harpur and Prager, 1996] and [Olshausen and Field, 1996] are based on gradient descent

[9] However, [Baddeley, 1996] presents some arguments against sparseness maximization.

(as described in Section 2), they require the specification of a schedule for adapting the learning rate parameter whereas this adaptation occurs naturally in Equation 40 due to the gradual reduction in the noise covariance as more inputs become available. More importantly, the learning rule expressed by Equation 40 as well as those proposed for the other weights in the model take into account not only bottom-up information but also top-down expectations during weight modification. This allows for the possibility of developing top-down guided *task-relevant* internal representations.[10]

The learning method presented in this paper is closely related to mean field methods for Bayesian belief networks [Saul *et al.*, 1996; Jaakkola *et al.*, 1996] which use mean field equations for computing lower bounds on the log-likelihood based cost function. The network parameters (the weights and biases) are subsequently learned by performing gradient ascent on the mean field derived bound for the log-likelihood. In comparison, the learning scheme presented in this paper can be regarded essentially as employing an on-line form of the Expectation-Maximization (EM) algorithm [Dempster *et al.*, 1977] in which the M-step performs maximum a posteriori (MAP) estimation rather than maximum likelihood (ML) estimation. Note that ML estimation is simply a special case of MAP estimation in which there is no prior information (this corresponds to the case where no predictions are computed from prior data as, for example, in Section 2). In particular, the on-line MAP form of the EM algorithm is invoked in three different ways in the present framework:

- *Estimation of the state* **r**: In this case, the "hidden" variables (or "missing" data) consist of the weights U and V while we treat the state vector $\mathbf{r}$ as the parameter to be estimated. Thus, the E-step involves computing the Gaussian distributions given by $(\overline{\mathbf{u}}, S)$ and $(\overline{\mathbf{v}}, T)$. The M-step then uses these distributions from the E-step to compute $\widehat{\mathbf{r}}$ and P, thereby maximizing the posterior probability of the parameter vector $\mathbf{r}$ (by minimizing the MDL-based cost function of Section 4). Note however that in the case where the activation function f is non-linear, the maximization in the M-step is not exact due to the Taylor series approximation to f used in the various update equations (the same applies to the other M-steps below).

- *Estimation of the generative weights* **u**: Here, the hidden variables consist of the state vector $\mathbf{r}$ and the weights V, while the parameter to be estimated is U. The E-step thus involves computing the distributions given by $(\overline{\mathbf{r}}, M)$ (or when feasible, $(\widehat{\mathbf{r}}(t|N), P(t|N))$ - see below) and $(\overline{\mathbf{v}}, T)$ which allow the calculation of $\widehat{\mathbf{u}}$ and P_u in the M-step as discussed in Section 6.

- *Estimation of prediction weights* **v**: In this case, the hidden variables comprise the state vector $\mathbf{r}$ and the weights U, and the parameter to be estimated is the state transition matrix V. Since V implements a Markov

[10] A hierarchical learning method for vector quantization in the presence of noise is presented in [Luttrell, 1992]. This method allows single-cause models whereas the present model can encode *multiple-cause models* [Dayan and Zemel, 1995; Saund, 1995] (see also [Luttrell, 1995]).

chain relating the states $\mathbf{r}$ in time (see Equation 12), the E-step in this case becomes more complicated than in the two cases above since the optimal estimate of the state $\mathbf{r}$ is no longer causal but depends temporally on future as well as past and current data values. In the case of hidden Markov models (HMMs) [Rabiner and Juang, 1986] (which are analogous to Kalman filters with discrete hidden states), this impasse is solved by using the familiar forward-backward procedure within the context of the Baum-Welch algorithm [Baum *et al.*, 1970] (the Baum-Welch algorithm is incidentally a special case of the EM algorithm). In our case, we may use a form of *optimal smoothing* [Bryson and Ho, 1975] which, when given the batch input data for time $t = 1, \ldots, N$, optimally assimilates past, current, and future data into temporally-smoothed state estimates $\widehat{\mathbf{r}}(t|N)$ (with covariance $P(t|N)$) for each of the time instants $t = 1, \ldots, N$. This, along with the computation of $(\overline{\mathbf{u}}, S)$, comprises the E-step. The M-step then involves computing $\widehat{\mathbf{v}}$ and P_v as described in the appendix.

It is worth noting that it may be desirable in many cases to implement the various E- and M-steps above for each set of parameters in an asynchronous manner rather than in strict alternation. For example, the cortex may choose to update its estimates for the state $\mathbf{r}$ much more frequently than those for U and V in order to act in real-time. The updates for U and V could then be carried out either on-line but at a slower rate than $\mathbf{r}$, or alternately when the organism is in a rest phase. The latter suggestion is especially attractive in the case of learning V given that the batch processing of data required by the optimal smoothing step suggested above may not be appropriate in many circumstances that demand real-time responses from the organism. In such cases, it might be beneficial for the organism to act based on its current on-line state estimates while simultaneously storing behaviorally relevant input sequences in memory as batch data. Such stored data may then be replayed out at a later time when the organism is in a more quiescent phase in order to learn or fine-tune the prediction weights V. This suggests a possible computational role for dream sleep in facilitating learning and memory.

9.6 Summary and Conclusion

Vision is fundamentally a dynamic process. Any putative model of the visual cortex must therefore acknowledge this fact and allow the modeling of dynamic processes. This paper suggests that the hierarchical structure of the visual cortex is well-suited to implement a multiscale estimation algorithm that can be regarded as a hierarchical form of the *extended Kalman filter* [Maybeck, 1979]. At each hierarchical level, the algorithm recursively estimates the current recognition state by combining "measurement" information from top-down and bottom-up cortical modules with the prediction that was made by using a "system" model. The "system" matrix (the prediction weights) as well as the "measurement" matrices (feedforward and feedback weights) can be *learned* from visual experiences in a dynamic environment. Thus, the adaptive processes in the cortex are modeled at two different time scales: a fast dynamic state estimation process that allows the visual system to anticipate incoming stimuli and a

slower Hebbian synaptic weight change mechanism. Both these processes are characterized by their respective mean and covariance estimates which are continually updated.

Some of the salient features of the model as a recognition system were illustrated by training it on a small number of man-made objects and testing recognition performance. The model was shown to be robust to partial occlusions and confusing stimuli. An ongoing effort involves a more quantitative evaluation of the model's recognition performance using realistic objects in natural scenes and the validation of the model in the domain of human visual search. Preliminary results along these two directions have been promising [Rao and Ballard, 1995a; Rao *et al.*, 1996]. When trained on natural images, the model developed feedforward receptive fields qualitatively resembling those found at the early hierarchical levels of the primate visual cortex. Current simulations involve applying the learning algorithms to successively higher levels of the network and comparing the receptive field properties that develop to those found at higher cortical levels such as IT. The hierarchical model also provides a computational platform for understanding a number of neurophysiological/psychophysical phenomena such as illusory contours, amodal completions, visual imagery, bistable percepts, end-stopping, and other effects that occur due to stimuli from beyond the classical receptive field. In particular, we provided simulations of the model suggesting that some of the recent results of [Gallant *et al.*, 1994; 1995; Gallant, 1996] regarding neural response suppression during free viewing of natural scenes may be interpreted as occurring due to predictions from higher levels of the cortex. While the present framework illustrates how information from two different sources, one top-down and the other bottom-up, may be integrated in a given cortical area, it is not hard to envision extensions of the model where more than two areas feed into a single area as is evident in the visual cortex [Felleman and Van Essen, 1991]. This would entail the relatively straightforward exercise of adding additional error/model terms to the optimization equations 26 and 27 to account for inputs from the additional areas before deriving the estimation algorithms.

A reassuring feature of the model is that it does not explicitly depend on the input signals being visual. Thus, given that the neocortex is organized along similar laminar input-output principles regardless of cortical area or input modality [Creutzfeldt, 1977], it is reasonable to assume that the general framework proposed herein may be uniformly applicable to other cortical areas such as the parietal, auditory, or motor cortex as well.[11] We expect the results obtained using the current model to play a crucial role in guiding our future cortical modeling efforts.

[11] We have recently shown in [Rao and Ballard, 1996] how the model presented in this paper can be extended to handle transformations in the image plane. The resulting estimation scheme parallels the functional dichotomy between the dorsal (occipitoparietal) and ventral (occipitotemporal) pathways in the primate visual cortex [Felleman and Van Essen, 1991]. Support for the possibility of modeling the motor cortex using Kalman filter-like mechanisms comes from the work of [Wolpert *et al.*, 1995] who provide strong psychophysical evidence for the existence of such mechanisms in the context of sensorimotor integration.

Acknowledgments

We are grateful to Peter Dayan and the two anonymous reviewers for their constructive comments and suggestions that immensely helped in improving the quality of the paper. We are particularly indebted to Peter Dayan for extensive criticism of an earlier version of this paper, which inspired the more general formulation of the approach as described herein. Thanks are also due to Jack Gallant for many interesting conversations, for reading drafts of the paper, and for the large number of suggestions and clarifications he provided during the course of this work. We would additionally like to thank Helder Araujo, Jessica Bayliss, Chris Brown, Avis Cohen, Mary Hayhoe, Kiriakos Kutulakos, Peter Lennie, Bill Merigan, Robbie Jacobs, Gerhard Sagerer, David Williams, and Gregory Zelinsky for pointers, comments, and discussions regarding the paper. Some of the images used in the simulations were taken from the Columbia object database, courtesy of Shree Nayar. This research was supported by NIH/PHS research grants 1-P41-RR09283 and 1-R24-RR06853-02, and by NSF research grants IRI-9406481 and IRI-8903582.

Appendix

Learning the Prediction Weights

While the simulations in the paper used static stimuli and therefore did not make use of the prediction step, it is almost certain that such predictions play an important role in the processing of *time-varying* stimuli. Therefore, we suggest in this appendix a possible learning procedure for modifying the synaptic weights $\mathbf{v}$ of the (possibly nonlinear) prediction units during exposure to time-varying stimuli. The derivation is similar to that for the feedback weights $\mathbf{u}$. Given the current vector of responses $\widehat{\mathbf{r}}$ at time t, define the $k \times k^2$ matrix $\widehat{R}$ to be:

$$\widehat{R} = \begin{bmatrix} \widehat{\mathbf{r}}^T & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{r}}^T & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \widehat{\mathbf{r}}^T \end{bmatrix} \tag{51}$$

Notice that the prediction step can be stated as:

$$\begin{aligned} \overline{\mathbf{r}}(t+1) &= f(V(t)\widehat{\mathbf{r}}(t)) + \overline{\mathbf{n}}(t) \\ &= f(\widehat{R}(t)\mathbf{v}(t)) + \overline{\mathbf{n}}(t) \end{aligned} \tag{52}$$

Differentiating the cost function J with respect to the vector of prediction weights $\mathbf{v}$ and setting the result to zero, we obtain (using Equation 52 for $\overline{\mathbf{r}}(t+1)$):

$$-\left(\frac{\partial f}{\partial \mathbf{v}}\right)^T M^{-1}(\mathbf{r}(t+1) - f(\widehat{R}(t)\widehat{\mathbf{v}}(t)) - \overline{\mathbf{n}}(t)) + T^{-1}(\widehat{\mathbf{v}}(t) - \overline{\mathbf{v}}(t)) + \beta\widehat{\mathbf{v}}(t) = 0 \tag{53}$$

which yields the update rule (for time t):

$$\widehat{\mathbf{v}}(t) = \overline{\mathbf{v}}(t) + P_v \left(\frac{\partial f}{\partial \mathbf{v}}\right)^T M^{-1}(\mathbf{r}(t+1) - f(\widehat{R}(t)\overline{\mathbf{v}}(t)) - \overline{\mathbf{n}}(t)) - \beta P_v \overline{\mathbf{v}}(t) \tag{54}$$

or more simply:

$$\widehat{\mathbf{v}}(t) = \overline{\mathbf{v}}(t) + P_v \left(\frac{\partial f}{\partial \mathbf{v}}\right)^T M^{-1}(\mathbf{r}(t+1) - \overline{\mathbf{r}}(t+1)) - \beta P_v \overline{\mathbf{v}}(t) \tag{55}$$

where $\overline{\mathbf{v}}(t+1) = \widehat{\mathbf{v}}(t) + \overline{\mathbf{n}}_v(t)$ and $\overline{\mathbf{r}}(t+1) = f(\widehat{R}(t)\overline{\mathbf{v}}(t)) + \overline{\mathbf{n}}(t)$. The corresponding covariance matrix is updated at time t as follows:

$$P_v(t) = (T(t)^{-1} + \left(\frac{\partial f}{\partial \mathbf{v}}\right)^T M(t)^{-1}\frac{\partial f}{\partial \mathbf{v}} + \beta I)^{-1} \tag{56}$$

where I is the $k^2 \times k^2$ identity matrix. The partial derivatives are evaluated at $\mathbf{v}(t) = \overline{\mathbf{v}}(t)$. Note that in order to obtain a closed-form expression for updating $\widehat{\mathbf{v}}$ and P_v, we once again applied a first-order Taylor series approximation to f around the current estimate $\overline{\mathbf{v}}(t)$. The covariance matrix T gets updated according to:

$$T(t+1) = P_v(t) + \mathbf{\Sigma}_v(t) \tag{57}$$

By appealing to a version of the EM algorithm (see Section 9.5), we may use $\mathbf{r}(t+1) = \widehat{\mathbf{r}}(t+1|N)$ in Equation 55 above, where $\widehat{\mathbf{r}}(t+1|N)$ is the optimal temporally *smoothed* state estimate [Bryson and Ho, 1975] for time $t+1$ ($\leq N$), given input data for each of the time instants $1, \ldots, N$ (see discussion in Section 9.5). In certain circumstances, it might be possible to use the on-line estimate $\widehat{\mathbf{r}}(t+1)$ in place of its computationally more expensive smoothed counterpart $\widehat{\mathbf{r}}(t+1|N)$, though such an approximation has been known to yield negative results in other related network architectures [Peter Dayan, personal communication]. Simulations are currently underway to evaluate the effectiveness of the above learning rule and the role of temporally smoothing the state estimates.

Intuitively, Equation 55 may be regarded as a form of quasi-Hebbian adaptation involving the presynaptic activity $\frac{\partial f}{\partial \mathbf{v}}$ (which is simply the response $\widehat{R}(t)$ at time t for linear neurons) and the postsynaptic activity $(\widehat{\mathbf{r}}(t+1|N) - \overline{\mathbf{r}}(t+1))$ which indicates the error in prediction. Note that computing the difference in the prediction error term requires the presence of some form of inhibitory interneurons. Inhibitory interneurons are also required by some of the other estimation algorithms derived in this paper. Interestingly, a wide variety of such interneurons have been discovered in the primate visual cortex [Jones, 1981].

References

[Allman *et al.*, 1985] J. Allman, F. Miezin, and E. McGuinness. Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local-global comparisons in visual neurons. *Annual Review of Neuroscience*, 8:407–429, 1985.

[Andersen and Van Essen, 1987] C.H. Andersen and D.C. Van Essen. Shifter circuits: A computational strategy for dynamic aspects of visual processing. *Proceedings of the National Academy of Sciences*, 84:6297–6301, 1987.

[Andersen *et al.*, 1985] R.A. Andersen, G.K. Essick, and R.M. Siegel. Encoding of spatial location by posterior parietal neurons. *Science*, 230:456–458, 1985.

[Ayache and Faugeras, 1986] N. Ayache and O.D. Faugeras. HYPER: A new approach for the recognition and positioning of two-dimensional objects. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8(1):44–54, 1986.

[Baddeley, 1996] R. Baddeley. Searching for filters with "interesting" output distributions: an uninteresting direction to explore? *Network*, 7(2):409–421, 1996.

[Baizer *et al.*, 1977] J.S. Baizer, D.L. Robinson, and B.M. Dow. Visual responses of area 18 neurons in awake, behaving monkey. *Journal of Neurophysiology*, 40:1024–1037, 1977.

[Ballard *et al.*, 1996] D.H. Ballard, M.M. Hayhoe, P.K. Pook, and R.P.N. Rao. Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 1996. In review.

[Barrow, 1987] H.G. Barrow. Learning receptive fields. In *Proceedings of the IEEE Int. Conf. on Neural Networks*, pages 115–121, 1987.

[Baum *et al.*, 1970] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41:164–171, 1970.

[Bell and Sejnowski, 1996] A.J. Bell and T.J. Sejnowski. The 'independent components' of natural scenes are edge filters. Submitted to *Vision Research*, 1996.

[Blake and Yuille, 1992] A. Blake and A. Yuille, editors. *Active Vision*. Cambridge, MA: MIT Press, 1992.

[Bolz and Gilbert, 1986] J. Bolz and C.D. Gilbert. Generation of end-inhibition in the visual cortex via interlaminar connections. *Nature*, 320:362–365, 1986.

[Broida and Chellappa, 1986] T.J. Broida and R. Chellappa. Estimation of object motion parameters from noisy images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8(1):90–99, 1986.

[Bryson and Ho, 1975] A.E. Bryson and Y.-C. Ho. *Applied Optimal Control*. New York: John Wiley and Sons, 1975.

[Carpenter and Grossberg, 1987] G.A. Carpenter and S. Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, 37:54–115, 1987.

[Chatfield and Collins, 1980] C. Chatfield and A.J. Collins. *Introduction to Multivariate Analysis*. New York: Chapman and Hall, 1980.

[Chou *et al.*, 1994a] K.C. Chou, A.S. Willsky, and A. Benveniste. Multiscale recursive estimation, data fusion, and regularization. *IEEE Transactions on Automatic Control*, 39(3):464–478, March 1994.

[Chou *et al.*, 1994b] K.C. Chou, A.S. Willsky, and R. Nikoukhah. Multiscale systems, Kalman filters, and Riccati equations. *IEEE Transactions on Automatic Control*, 39(3):479–492, March 1994.

[Churchland and Sejnowski, 1992] P. Churchland and T. Sejnowski. *The Computational Brain*. Cambridge, MA: MIT Press, 1992.

[Churchland *et al.*, 1994] P.S. Churchland, V.S. Ramachandran, and T.J. Sejnowski. A critique of pure vision. In C. Koch and J.L. Davis, editors, *Large-Scale Neuronal Theories of the Brain*, pages 23–60. Cambridge, MA: MIT Press, 1994.

[Creutzfeldt, 1977] O.D. Creutzfeldt. Generality of the functional structure of the neocortex. *Naturwissenschaften*, 64:507–517, 1977.

[Daugman, 1980] J.G. Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20:847–856, 1980.

[Daugman, 1988] J.G. Daugman. Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Trans. Acoustics, Speech, and Signal Proc.*, 36(7):1169–1179, 1988.

[Dayan and Hinton, 1996] P. Dayan and G.E. Hinton. Varieties of Helmholtz machine. *Neural Networks*, 1996. (In press).

[Dayan and Zemel, 1995] P. Dayan and R.S. Zemel. Competition and multiple cause models. *Neural Computation*, 7:565–579, 1995.

[Dayan *et al.*, 1995] P. Dayan, G.E. Hinton, R.M. Neal, and R.S. Zemel. The Helmholtz machine. *Neural Computation*, 7:889–904, 1995.

[Deacon, 1989] T. Deacon. Holism and associationism in neuropsychology: An anatomical synthesis. In E. Perecman, editor, *Integrating Theory and Practice in Clinical Neuropsychology*, pages 1–47. Hillsdale, NJ: Lawrence Erlbaum, 1989.

[DeAngelis *et al.*, 1992] G.C. DeAngelis, J.G. Robson, I. Ohzawa, and R.D. Freeman. Organization of suppression in receptive fields of neurons in cat visual cortex. *Journal of Neurophysiology*, 68(1):144–163, 1992.

[Dempster *et al.*, 1977] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society Series B*, 39:1–38, 1977.

[Desimone and Schein, 1987] R. Desimone and S.J. Schein. Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. *Journal of Neurophysiology*, 57:835–868, 1987.

[Desimone and Ungerleider, 1989] R. Desimone and L.G. Ungerleider. Neural mechanisms of visual processing in monkeys. In F. Boller and J. Grafman, editors, *Handbook of Neuropsychology*, volume 2, chapter 14, pages 267–299. New York: Elsevier, 1989.

[Dickmanns and Mysliwetz, 1992] E.D. Dickmanns and B.D. Mysliwetz. Recursive 3D road and relative ego-state recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 14(2):199–213, 1992.

[Douglas *et al.*, 1989] R.J. Douglas, K.A.C. Martin, and D. Whitteridge. A canonical microcircuit for neocortex. *Neural Computation*, 1:480–488, 1989.

[Duhamel *et al.*, 1992] J. Duhamel, L. Colby, and M.E. Goldberg. The updating of the representation of visual space in parietal cortex by intended eye movements. *Science*, 255:90–92, 1992.

[Edelman, 1978] G.M. Edelman. Group selection and phasic re-entrant signaling: a theory of higher brain function. In V.B. Mountcastle and G.M. Edelman, editors, *The Mindful Brain*, pages 51–100. Cambridge, MA: MIT Press, 1978.

[Felleman and Van Essen, 1991] D.J. Felleman and D.C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1:1–47, 1991.

[Feller, 1968] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, New York, 1968.

[Fukushima, 1988] K. Fukushima. A neural network for visual pattern recognition. *Computer*, 21(3):65–75, 1988.

[Gallant *et al.*, 1994] J.L. Gallant, C.E. Connor, and D.C. Van Essen. Responses of visual cortical neurons in a monkey freely viewing natural scenes. *Soc. Neuro. Abstr.*, 20:838, 1994.

[Gallant *et al.*, 1995] J.L. Gallant, C.E. Connor, H. Drury, and D.C. Van Essen. Neural responses in monkey visual cortex during free viewing of natural scenes: Mechanisms of response suppression. *Invest. Ophthalmol. Vis. Sci.*, 36:1052, 1995.

[Gallant, 1996] J.L. Gallant. Cortical responses to natural scenes under controlled and free viewing conditions. *Invest. Ophthalmol. Vis. Sci.*, 37(3):674, 1996.

[Gilbert and Wiesel, 1981] C.D. Gilbert and T.N. Wiesel. Laminar specialization and intracortical connections in cat primary visual cortex. In *The Organization of the Cerebral Cortex*, pages 163–191. Cambridge, MA: MIT Press, 1981.

[Girosi *et al.*, 1995] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269, 1995.

[Grenander, 1976-81] U. Grenander. *Lectures in Pattern Theory I, II and III: Pattern Analysis, Pattern Synthesis and Regular Structures*. Berlin: Springer-Verlag, 1976-81.

[Grimes and McConkie, 1995] J. Grimes and G. McConkie. On the insensitivity of the human visual system to image changes made during saccades. In K. Akins, editor, *Problems in Perception*. Oxford, UK: Oxford University Press, 1995.

[Grosof *et al.*, 1992] D.H. Grosof, R.M. Shapley, and M.J. Hawken. Macaque striate responses to anomalous contours? *Invest. Ophthalmol. Vis. Sci.*, 33:1257, 1992.

[Gulyas *et al.*, 1987] B. Gulyas, G.A. Orban, J. Duysens, and H. Maes. The suppressive influence of moving textured backgrounds on responses of cat striate neurons to moving bars. *Journal of Neurophysiology*, 57(6):1767–1791, 1987.

[Hallam, 1983] J. Hallam. Resolving observer motion by object tracking. In *Proc. of 8th International Joint Conf. on Artificial Intelligence*, volume 2, pages 792–798, 1983.

[Hancock *et al.*, 1992] P.J.B. Hancock, R.J. Baddeley, and L.S. Smith. The principal components of natural images. *Network*, 3:61–70, 1992.

[Harpur and Prager, 1996] G.F. Harpur and R.W. Prager. Development of low-entropy coding in a recurrent network. *Network*, 7:277–284, 1996.

[Harth *et al.*, 1987] E. Harth, K.P. Unnikrishnan, and A.S. Pandya. The inversion of sensory processing by feedback pathways: A model of visual cognitive functions. *Science*, 237:184–187, 1987.

[Heeger *et al.*, 1996] D.J. Heeger, E.P. Simoncelli, and J.A. Movshon. Computational models of cortical visual processing. *Proc. National Acad. Sciences*, 93:623–627, 1996.

[Hinton *et al.*, 1995] G.E. Hinton, P. Dayan, B.J. Frey, and R.M. Neal. The wake-sleep algorithm for unsupervised neural networks. *Science*, 268:1158–1161, 1995.

[Hubel and Wiesel, 1962] D.H. Hubel and T.N. Wiesel. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology (London)*, 160:106–154, 1962.

[Hubel and Wiesel, 1965] D.H. Hubel and T.N. Wiesel. Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, 28:229–289, 1965.

[Hubel and Wiesel, 1968] D.H. Hubel and T.N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology (London)*, 195:215–243, 1968.

[Hubel, 1988] D.H. Hubel. *Eye, brain, and vision*. New York: Scientific American Library, 1988.

[Irwin *et al.*, 1994] D.E. Irwin, G.W. McConkie, L. Carlson-Radvansky, and C. Currie. A localist evaluation solution for visual stability across saccades. *Behavioral and Brain Sciences*, 17:265–266, 1994.

[Jaakkola *et al.*, 1996] T. Jaakkola, L.K. Saul, and M.I. Jordan. Fast learning by bounding likelihoods in sigmoid type belief networks. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 528–534. Cambridge, MA: MIT Press, 1996.

[Jones, 1981] E.G. Jones. Anatomy of cerebral cortex: Columnar input-output organization. In *The Organization of the Cerebral Cortex*, pages 199–235. Cambridge, MA: MIT Press, 1981.

[Jordan and Rumelhart, 1992] M.I. Jordan and D.E. Rumelhart. Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16:307–354, 1992.

[Kalman and Bucy, 1961] R.E. Kalman and R.S. Bucy. New results in linear filtering and prediction theory. *Trans. ASME J. Basic Eng.*, 83:95–108, 1961.

[Kalman, 1960] R.E. Kalman. A new approach to linear filtering and prediction theory. *Trans. ASME J. Basic Eng.*, 82:35–45, 1960.

[Kanizsa, 1990] G. Kanizsa. Subjective contours. In I. Rock, editor, *The Perceptual World: Readings from Scientific American Magazine*, pages 155–163. New York: W.H. Freeman and Company, 1990.

[Kawato *et al.*, 1993] M. Kawato, H. Hayakawa, and T. Inui. A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network*, 4:415–422, 1993.

[Koch, 1996] C. Koch. Towards the neuronal substrate of visual consciousness. In S.R. Hameroff, A.W. Kaszniak, and A.C. Scott, editors, *Towards a Science of Consciousness: The First Tucson Discussions and Debates*. Cambridge, MA: MIT Press, 1996.

[Kosslyn *et al.*, 1995] S.M. Kosslyn, W.L. Thompson, I.J. Kim, and N.M. Alpert. Topographical representations of mental images in primary visual cortex. *Nature*, 378:496–498, 1995.

[Li and Vitanyi, 1993] M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. New York: Springer-Verlag, 1993.

[Linsker, 1988] R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

[Luettgen and Willsky, 1995] M.R. Luettgen and A.S. Willsky. Likelihood calculation for a class of multi-scale stochastic models, with application to texture discrimination. *IEEE Transactions on Image Processing*, 4(2):194–207, February 1995.

[Lund, 1981] J.S. Lund. Intrinsic organization of the primate visual cortex, area 17, as seen in Golgi preparations. In *The Organization of the Cerebral Cortex*, pages 105–124. Cambridge, MA: MIT Press, 1981.

[Luttrell, 1992] S.P. Luttrell. Self-supervised adaptive networks. *IEE Proceedings Part F*, 139:371–377, 1992.

[Luttrell, 1995] S.P. Luttrell. A componential self-organising neural network. Technical Report DRA/CIS(SE1)/651/11/RP/3.1, Defence Research Agency, Malvern, UK, November 1995.

[MacKay, 1956] D.M. MacKay. The epistemological problem for automata. In *Automata Studies*, pages 235–251. Princeton, NJ: Princeton University Press, 1956.

[Marcelja, 1980] S. Marcelja. Mathematical description of the responses of simple cortical cells. *Journal of the Optical Society of America*, 70:1297–1300, 1980.

[Matthies *et al.*, 1989] L. Matthies, T. Kanade, and R. Szeliski. Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3:209–236, 1989.

[Maunsell and Newsome, 1987] J.H.R. Maunsell and W.T. Newsome. Visual processing in monkey extrastriate cortex. *Annual Review of Neuroscience*, 10:363–401, 1987.

[Maybeck, 1979] P.S. Maybeck. *Stochastic Models, Estimation, and Control (Vols. I and II)*. New York: Academic Press, 1979.

[McConkie, 1991] G.W. McConkie. Perceiving a stable visual world. In *Proceedings of the Sixth European Conference on Eye Movements*, pages 5–7. Leuven, Belgium: Laboratory of Experimental Psychology, 1991.

[Mignard and Malpeli, 1991] M. Mignard and J.G. Malpeli. Paths of information flow through visual cortex. *Science*, 251:1249–1251, 1991.

[Miller *et al.*, 1991] E.K. Miller, L. Li, and R. Desimone. A neural mechanism for working and recognition memory in inferior temporal cortex. *Science*, 254:1377–1379, 1991.

[Muller *et al.*, 1996] J.R. Muller, B. Singer, J. Krauskopf, and P. Lennie. Center-surround contrast effects in macaque cortex. *Invest. Ophthalmol. Vis. Sci.*, 37(3):904, 1996.

[Mumford, 1994] D. Mumford. Neuronal architectures for pattern-theoretic problems. In C. Koch and J.L. Davis, editors, *Large-Scale Neuronal Theories of the Brain*, pages 125–152. Cambridge, MA: MIT Press, 1994.

[Murphy and Sillito, 1987] P.C. Murphy and A.M. Sillito. Corticofugal feedback influences the generation of length tuning in the visual pathway. *Nature*, 329:727–729, 1987.

[Nelson and Frost, 1978] J.I. Nelson and B.J. Frost. Orientation-selective inhibition from beyond the classic visual receptive field. *Brain Research*, 139:359–365, 1978.

[Nowlan and Hinton, 1992] S.J. Nowlan and G.E. Hinton. Simplifying neural networks by soft weight-sharing. *Neural Computation*, 4(4):473–493, 1992.

[Oja, 1989] E. Oja. Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, 1:61–68, 1989.

[Olshausen and Field, 1996] B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[Oram and Perrett, 1992] M.W. Oram and D.I. Perrett. Time course of neural responses discriminating different views of the face and head. *Journal of Neurophysiology*, 68(1):70–84, 1992.

[Palmer *et al.*, 1991] L.A. Palmer, J.P. Jones, and R.A. Stepnoski. Striate receptive fields as linear filters: Characterization in two dimensions of space. In A.G. Leventhal, editor, *The Neural Basis of Visual Function*, pages 246–265. Boca Raton: CRC Press, 1991.

[Pece, 1992] A.E.C. Pece. Redundancy reduction of a Gabor representation: a possible computational role for feedback from primary visual cortex to lateral geniculate nucleus. In I. Aleksander and J. Taylor, editors, *Artificial Neural Networks 2*, pages 865–868. Amsterdam: Elsevier Science, 1992.

[Pentland, 1992] A.P. Pentland. Dynamic vision. In G.A. Carpenter and S. Grossberg, editors, *Neural Networks for Vision and Image Processing*, pages 133–159. Cambridge, MA: MIT Press, 1992.

[Pitts and McCulloch, 1947] W. Pitts and W.S. McCulloch. How we know universals: the perception of auditory and visual forms. *Bulletin of Mathematical Biophysics*, 9:127–147, 1947.

[Poggio *et al.*, 1985] T. Poggio, V. Torre, and C. Koch. Computational vision and regularization theory. *Nature*, 317:314–319, 1985.

[Poggio *et al.*, 1992] T. Poggio, M. Fahle, and S. Edelman. Fast perceptual learning in visual hyperacuity. *Science*, 256:1018–1021, May 1992.

[Poggio, 1990] T. Poggio. A theory of how the brain might work. In *Cold Spring Harbor Symposia on Quantitative Biology*, pages 899–910. Cold Spring Harbor Laboratory Press, 1990.

[Rabiner and Juang, 1986] L.R. Rabiner and B.H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3:4–16, January 1986.

[Rao and Ballard, 1995a] R.P.N. Rao and D.H. Ballard. An active vision architecture based on iconic representations. *Artificial Intelligence (Special Issue on Vision)*, 78:461–505, 1995.

[Rao and Ballard, 1995b] R.P.N. Rao and D.H. Ballard. Dynamic model of visual memory predicts neural response properties in the visual cortex. Technical Report 95.4, National Resource Laboratory for the Study of Brain and Behavior, Department of Computer Science, University of Rochester, November 1995.

[Rao and Ballard, 1996] R.P.N. Rao and D.H. Ballard. A class of stochastic models for invariant recognition, motion, and stereo. Technical Report 96.1, National Resource Laboratory for the Study of Brain and Behavior, Department of Computer Science, University of Rochester, June 1996.

[Rao *et al.*, 1996] R.P.N. Rao, G.J. Zelinsky, M.M. Hayhoe, and D.H. Ballard. Modeling saccadic targeting in visual search. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 830–836. Cambridge, MA: MIT Press, 1996.

[Rissanen, 1989] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific, 1989.

[Rockland and Pandya, 1979] K.S. Rockland and D.N. Pandya. Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Research*, 179:3–20, 1979.

[Rolls, 1989] E. Rolls. The representation and storage of information in neuronal networks in the primate cerebral cortex and hippocampus. In R. Durbin, C. Miall, and G. Mitchison, editors, *The Computing Neuron*, chapter 8, pages 125–159. Addison-Wesley, 1989.

[Saul *et al.*, 1996] L.K. Saul, T. Jaakkola, and M.I. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76, 1996.

[Saund, 1995] E. Saund. A multiple cause mixture model for unsupervised learning. *Neural Computation*, 7:51–71, 1995.

[Shannon, 1948] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.

[Softky, 1996] W.R. Softky. Unsupervised pixel-prediction. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 809–815. Cambridge, MA: MIT Press, 1996.

[Thorpe and Imbert, 1989] S.J. Thorpe and M. Imbert. Biological contraints on connectionist modelling. In R. Pfeifer, Z. Schreter, F. Fogelman-Soulie, and L.Steels, editors, *Connectionism in Perspective*, pages 63–92. Amsterdam: Elsevier, 1989.

[Tovee *et al.*, 1993] M.J. Tovee, E.T. Rolls, A. Treves, and R.P. Bellis. Information encoding and the responses of single neurons in the primate temporal visual cortex. *Journal of Neurophysiology*, 70(2):640–654, 1993.

[Ullman, 1994] S. Ullman. Sequence seeking and counterstreams: A model for bidirectional information flow in the cortex. In C. Koch and J.L. Davis, editors, *Large-Scale Neuronal Theories of the Brain*, pages 257–270. Cambridge, MA: MIT Press, 1994.

[Van Essen and Maunsell, 1983] D.C. Van Essen and J.H.R. Maunsell. Hierarchical organization and functional streams in the visual cortex. *Trends in Neuroscience*, 6:370–375, 1983.

[Van Essen *et al.*, 1994] D.C. Van Essen, C.H. Anderson, and B.A. Olshausen. Dynamic routing strategies in sensory, motor, and cognitive processing. In C. Koch and J.L. Davis, editors, *Large-Scale Neuronal Theories of the Brain*, pages 271–299. Cambridge, MA: MIT Press, 1994.

[Van Essen, 1985] D.C. Van Essen. Functional organization of primate visual cortex. In A. Peters and E.G. Jones, editors, *Cerebral Cortex*, volume 3, pages 259–329. Plenum, 1985.

[von der Heydt *et al.*, 1984] R. von der Heydt, E. Peterhans, and G. Baumgartner. Illusory contours and cortical neuron responses. *Science*, 224:1260–1262, 1984.

[Williams, 1985] R.J. Williams. Feature discovery through error-correction learning. Technical Report 8501, Institute for Cognitive Science, University of California at San Diego, 1985.

[Wolpert *et al.*, 1995] D.M. Wolpert, Z. Ghahramani, and M.I. Jordan. An internal model for sensorimotor integration. *Science*, 269:1880–1882, 1995.

[Young, 1985] R.A. Young. The Gaussian derivative theory of spatial vision: Analysis of cortical cell receptive field line-weighting profiles. *General Motors Research Publication GMR-4920*, 1985.

[Zeki, 1976] S.M. Zeki. The functional organization of projections from striate to prestriate visual cortex in the rhesus monkey. *Cold Spring Harbor Symp. Quant. Biology*, 40:591–600, 1976.

[Zeki, 1978] S.M. Zeki. Uniformity and diversity of structure and function in rhesus monkey prestriate visual cortex. *Journal of Physiology (London)*, 277:273–290, 1978.

[Zeki, 1983] S.M. Zeki. Colour coding in the cerebral cortex: the reaction of cells in monkey visual cortex to wavelengths and colours. *Neuroscience*, 9:741–765, 1983.

[Zemel, 1994] R.S. Zemel. *A Minimum Description Length Framework for Unsupervised Learning*. PhD thesis, Department of Computer Science, University of Toronto, 1994.