Journal of Computational Neuroscience, 2, 45–62 (1995)
© 1995 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

A Multiscale Dynamic Routing Circuit for Forming Size- and Position-Invariant Object Representations

BRUNO A. OLSHAUSEN baol@cornell.edu
Dept. of Anatomy and Neurobiology, Box 8108, Washington University School of Medicine, 660 S. Euclid Avenue, St. Louis, MO 63110; and Computation and Neural Systems Program, California Institute of Technology, Pasadena, CA 91125

CHARLES H. ANDERSON AND DAVID C. VAN ESSEN
Dept. of Anatomy and Neurobiology, Box 8108, Washington University School of Medicine, 660 S. Euclid Avenue, St. Louis, MO 63110

Received July 18, 1994; Revised November 15, 1994; Accepted December 12, 1994

Action Editor: Eve Marder

Abstract. We describe a neural model for forming size- and position-invariant representations of visual objects. The model is based on a previously proposed dynamic routing circuit that remaps selected portions of an input array into an object-centered reference frame. Here, we show how a multiscale representation may be incorporated at the input stage of the model, and we describe the control architecture and dynamics for a hierarchical, multistage routing circuit. Specific neurobiological substrates and mechanisms for the model are proposed, and a number of testable predictions are described.

Keywords: recognition, invariance, routing, multiscale, attention

1. Introduction

The formation of invariant object representations is a problem that confronts both biological and machine vision systems. Since objects may appear over a wide range of different positions and sizes in the visual field, invariant properties of objects must be somehow captured and represented in order to successfully recognize them under different viewing configurations. Most theories for how invariance is achieved have been based on the notion that increasingly complex features are extracted through a series of processing stages, with cells at each stage becoming progressively less specific for position and size via a process of regional summation (e.g., Fukushima, 1980; LeCun et al., 1990; Foldiak, 1991). An inherent problem with these proposals, however, is that information about the spatial relationships of features within an object is lost. Thus, it is unclear how these systems would discriminate between objects containing the same features with different spatial arrangements. In addition, these systems necessarily lose information about the position and size of an object, or other contextual information (e.g., the slant or font of a character) that we readily retain and use to advantage during recognition.

Recently, we proposed a neurobiological model for forming position- and size-invariant representations in which information about spatial relationships is explicitly preserved (Olshausen et al., 1993; Anderson and Van Essen, 1987). The model utilizes a set of control neurons to dynamically change the strengths of intracortical connections so that information from a windowed region in the retina is routed into an object-centered reference frame representation in higher cortical areas. Although this model was designed to be consistent with available neurophysiological, neuroanatomical, and psychophysical data, certain aspects of the early representation in cortical area V1 were ignored as an initial simplification. For example, the fact that cells in the visual cortex are tuned to different spatial-frequencies was not considered in initial versions of the model, nor was the logarithmic transformation of visual space due to non-uniform spatial sampling in the retina taken into account.

In this paper, we expand upon our earlier model (Olshausen et al., 1993) by showing how the multiscale, logarithmic nature of the early cortical representation may be modeled as a stack of sampling lattices with different resolutions and spatial extents. This representation can then be incorporated into the routing circuit advantageously by selectively routing information from high or low resolution lattices depending on whether the window is small or large respectively. This way, much of the image blurring required for rescaling can be accomplished by switching between the different lattices of the stack (where the scaling has been "precomputed"), rather than requiring the routing circuit to dynamically blur over a very large range of spatial scales. We also describe here the control architecture and dynamics for a multistage, hierarchical routing circuit, since the initial version of the model described only the control of a single stage routing circuit.

We begin by describing a scaled-down, model routing circuit that is largely divorced from neurobiological details for the purposes of illustration and simulation. The full-scale model along with its proposed neurobiological substrates and mechanisms is presented in the next section. Predictions arising from the model are then discussed, and the commonalities and differences between this model and related models are described. A Bayesian interpretation of the model is also provided which provides a coherent framework for understanding the operation of the routing circuit as a whole. Finally, we describe some of the shortcomings of the model and suggest future directions for improvement.

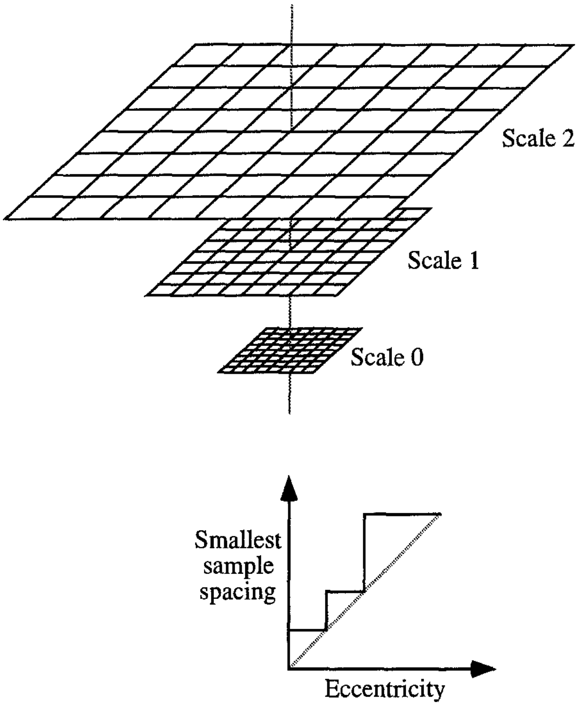

Fig. 1. The multiscale "stack" model of Koenderink and Van Doorn. The input is represented by a stack of sampling lattices at different scales. Each lattice comprises the same number of sample nodes and covers a progressively greater spatial extent, at lower resolution, than the level below it. When combined, the different lattices of the stack provide both a multi-resolution representation of the input image and also a piecewise approximation of the linear dependence of sample spacing on eccentricity found in the retina, as shown below.

2. Model

Multiscale Representation

The input to the routing circuit is composed of a multiscale array of sampling nodes, as in the *stack model* of Koenderink and van Doorn (1978). The basic scheme is illustrated in Fig. 1. A stack of sampling lattices represents the image at different resolutions. Each sampling lattice has the same total number of sample nodes and covers a portion of the visual field with uniform sample spacing. A low resolution lattice (e.g., scale 2) thus covers a proportionally larger region of visual space than a high resolution lattice (e.g., scale 0). It is assumed for the moment that resolution changes by a factor of two from one lattice to the next. The combination of these lattices provides a multi-resolution representation of the input image and also constitutes a piecewise approximation to the linear dependence of sample spacing on eccentricity found in the retina (which results in the logarithmic transformation of visual space in the cortex, Schwartz, 1977). A useful property of this scheme is that when an object changes in size by a factor of two, its representation simply translates from one lattice to another but otherwise remains unchanged. In order to take advantage of this property for forming invariant object representations, though, there must be a mechanism for routing information from the appropriate lattice and spatial location, and also for appropriate interpolation when the scale of the object lies in between that of neighboring lattices (i.e., when the size change of the object is not an integer multiple of an octave).

Routing Circuit Architecture

In devising a neural architecture for routing information from an input array into an output array, it is important to consider constraints on fan-in, since real neurons are limited in the number of inputs they accept. In order to route information from any location within an $N \times N$ input array to an $M \times M$ output array, a fan-in of $(N-M+1)^2$ is required for each

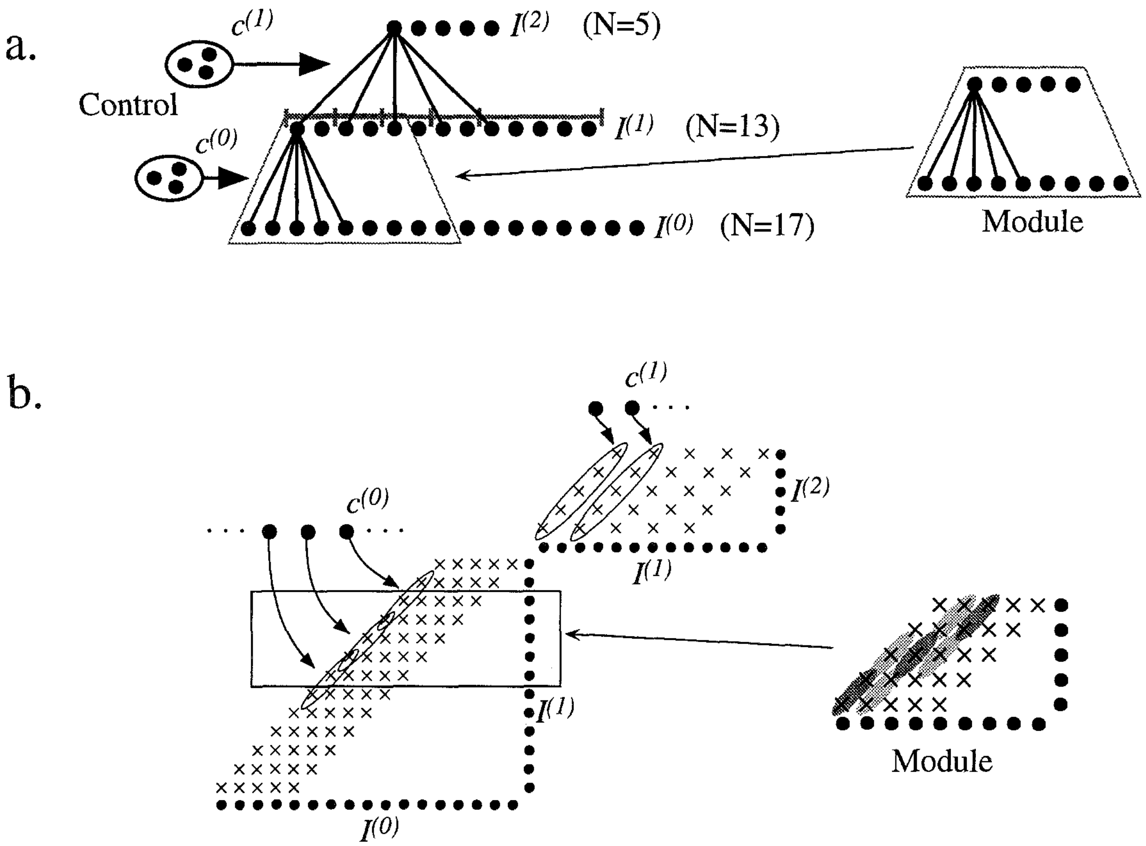

Fig. 2. Routing architecture for a single lattice of the stack. (a) A two-stage routing circuit. Connections are shown only for the leftmost node of each level. The connections for the other nodes are the same but merely shifted. An intermediate level of 13 nodes allows a window of 5–8 nodes located anywhere within a 17-node input array to be remapped into a five-node output array while maintaining a fan-in of five inputs on any node. The first stage is composed of overlapping modules of 9 inputs and 5 outputs each (a module is outlined by the trapezoid), while the second stage links the output of these modules to the top level. The left border of each module's output is denoted by the vertical tick marks in level 1. (b) A connection-space diagram illustrating the shape of the control blocks (Ψ_k). The horizontal axes represent the nodes of an input array, and the vertical axes represent the nodes of an output array. (Since $I^{(1)}$ serves as both an input and output array, it appears depicted both horizontally and vertically.) Each × denotes a physical connection from an input node to an output node. Each control neuron modulates a local block of connections, as outlined by the diagonal ellipses. The connection space of a single module in the first stage is shown magnified at right, with an example of how continuous shifting and scaling may be achieved by interpolating patterns in connection space.

output node. If this amount exceeds the allowable fan-in, then the routing circuit must be broken into multiple stages.

We consider first the routing circuit for a single lattice of the stack in one dimension. For the scaled-down circuit, we shall assume that each input lattice comprises 17 sample nodes across, that the output array comprises five nodes across, and that a maximum of five inputs is allowed on any node. Thus, the routing circuit must be broken into at least two stages in order to satisfy the fan-in constraint. One possible multistage routing architecture that satisfies this constraint is shown in Fig. 2a. It is helpful to think of the first stage of this circuit as composed of overlapping modules of nine inputs and five outputs each. The second stage then connects the output node of each first-stage module (spaced apart by two nodes in the intermediate level) to its corresponding node in the top level.

The activities of the nodes at each level are determined via a linear summation from the level below. That is,

$$I_i^{(l+1)} = \sum_j w_{ij}^{(l)} I_j^{(l)}, \qquad (1)$$

where $I_i^{(l)}$ denotes the activity of node i of level l, and

$w_{ij}^{(l)}$ denotes the weight of the connection from node j of level l to node i of level $l+1$. The weights are set dynamically by a set of control neurons, c, that make multiplicative couplings with the inputs. Each control neuron of level l, $c_k^{(l)}$, modulates a local group of connections so that patterns in weight space may be generated via interpolation:

$$w_{ij}^{(l)} = \sum_k c_k^{(l)} \Psi_k^{(l)}(j,i) + w_{\text{rest}} \tag{2}$$

where $\Psi_k^{(l)}(j,i)$ is an analog function that specifies the extent to which the connection from input j to output i is modulated by control neuron k, and the term w_{rest} denotes the default value of a connection in the absence of any control neuron activity. The region of weight space covered by each $\Psi_k^{(l)}$ (i.e., $\Psi_k^{(l)} > 0$) is denoted a *control block*.

The specific configuration of the control blocks is illustrated in the connection-space diagram of Fig. 2b. At the top stage of the circuit, the $\Psi_k^{(1)}$ are chosen so that each control neuron maps the output of a different module of level 1 into level 2. Thus,

$$\Psi_k^{(1)}(j,i) = \delta(j-i-2k), \quad k = 0,1,\ldots,4 \tag{3}$$

$$\delta(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

There are a total of five control neurons for this stage (one for each module). In the bottom stage, the control blocks have a gaussian taper and are chosen to overlap so as to interpolate patterns in weight space. This allows scaling and warping less than a factor of two within each module. Formally, we specify the shape of these control blocks as

$$\Psi_k^{(0)}(j,i) = \exp\left[-\frac{(j-i-m)^2}{2\sigma^2} - \frac{(i-nW)^2}{W^2/2}\right]$$
$$k = mN_B + n, \quad m = 0,1,\ldots,4$$
$$n = 0,1,\ldots,6 \tag{5}$$

where m denotes the index of translation (i.e., which diagonal in Fig. 2b), n is the index of the control block within a diagonal, N_B is the number of control blocks within a diagonal, and W is the spacing of the control blocks along the diagonal. Here, $W = 2$ (half the window size), and $N_B = 7$, giving a total of 35 (7×5) control neurons for this stage. Choosing the coupling coefficients for the upper and lower stages in this way results in a hierarchical control scheme, in which the top stage control neurons perform "macroshifting" among the five modules in level 1, while the bottom stage control neurons perform "microshifting" and rescaling within a factor of two on the input. (The combination of macro- and micro-shifting was originally proposed for translational shifts by Anderson and Van Essen, 1987.)

The proposed routing architecture incorporating all lattices of the stack is shown in Fig. 3. Each lattice serves as input to a separate routing stream that translates the window of attention within that lattice and rescales within a factor of two. The final stage of the circuit performs rescaling greater than a factor of two by switching between the outputs of the different routing streams for each scale. More generally, multiple scales could be routed into the output simultaneously

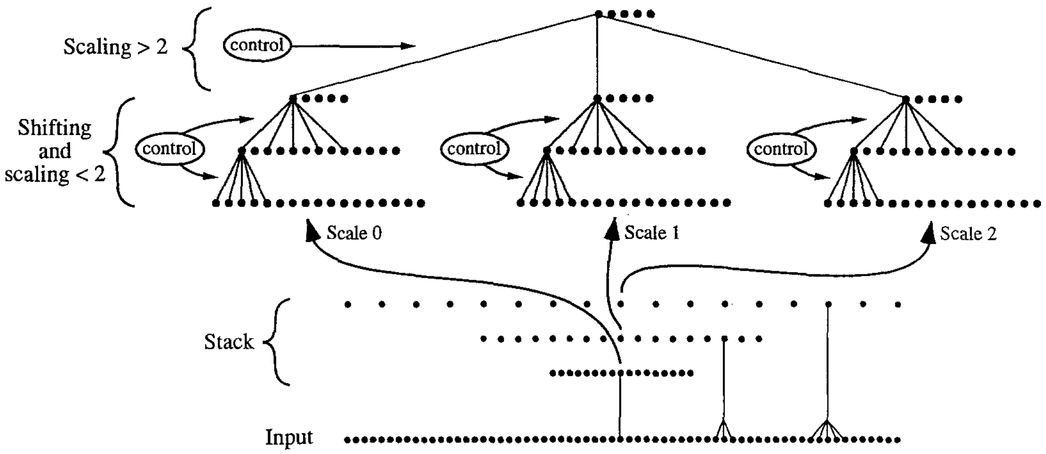

Fig. 3. Routing architecture for the stack. A third and final stage switches between scales by selecting among the outputs of the different routing streams for each scale.

in order to have a multiscale representation within the attentional window (as in the "jet" representation of Buhmann et al., 1990), but for now we shall utilize only one scale at a time.

An alternative means for arranging the routing circuit would be to perform only shifts within the lower stages, leaving the top stage to perform the interpolation between the precomputed scales provided by the stack. However, this would require routing an image twice the window size up to the top stage. On the other hand, rotations and warps would best be performed at the top stage; and since these operations will inevitably involve a moderate amount of rescaling, it may actually be desirable to maintain an extra margin of space surrounding the window as information is routed upward.

Control Dynamics

The purpose of the routing circuit is to focus the neural resources for recognition on a specific region, or object, within a scene. Thus, it would be desirable for the control neurons to automatically steer the attentional window to salient areas, or potential objects in the image. Salient areas can often be defined on the basis of relatively low-level cues—such as local contrast in motion, depth, texture, or color (i.e., "pop-out"; Koch and Ullman, 1985; Anderson et al., 1985; Milanese, 1993). Here, we utilize a very simple measure of salience based on luminance contrast, in which attention is attracted to "blobs"—or contiguous regions of activity—in the image. Once the attentional window has been roughly focused on a blob, the contents of the window are fed to an associative memory, which then acts to refine the position and size during object matching. After this has been accomplished, the current locus is inhibited so that attention may be shifted to other blobs within the input.

In what follows, we derive the control circuitry and dynamics for achieving these autonomous modes of operation. We consider first the control of a routing circuit for a single lattice of the stack (i.e., Fig. 2a) and then the multiscale case.

Focusing Attention on a Blob. In order to focus the attentional window on a blob in the input, the network's "goal" will be to fill the output array with a blob while maintaining a topographic correspondence between the input and output nodes. This goal is formulated as a two part objective function, and the control neuron dynamics are then obtained by performing gradient descent on this function. The first part of the objective function, E_b, provides a measure of how well a blob is focused on the output array. We choose E_b to be defined as

$$E_b = -\sum_i I_i^{(2)} G_i \qquad (6)$$

where the G_i are samples of a blob template centered on the output array, $G_i = e^{-(i-2)^2/4}$. The second part of the objective function, E_c, is designed to favor control states that correspond to translations or scalings of the input-output transformation. We choose E_c to be

$$E_c = -\sum_{l,m,n} c_m^{(l)} U_{mn}^{(l)} c_n^{(l)} \qquad (7)$$

where the constraint matrix for each stage, $\mathbf{U}^{(l)}$, is chosen so as to appropriately couple the control neurons: for the top stage of the circuit, where each control neuron corresponds to a different position of the window of attention, we set $\mathbf{U}^{(1)}$ to a matrix of all -1's except along the diagonal, which has the effect of punishing any state in which two or more control neurons are active simultaneously (winner-take-all); for the bottom stage, where control blocks overlap so as to interpolate patterns in connection space, the constraint matrix is set so that control neurons corresponding to a common translation or scale (those lying along a common diagonal in connection-space) couple positively ($U_{mn}^{(0)} > 0$), while control neurons that are not part of the same transformation couple negatively ($U_{mn}^{(0)} < 0$).[1] These couplings are only necessary for control neurons belonging to the same module, so we can set $U_{mn}^{(0)} = 0$ if $c_m^{(0)}$ and $c_n^{(0)}$ are in different modules.

The differential equation governing the dynamics of the top-stage control neurons is the same as that derived previously for a single-stage circuit (Olshausen et al., 1993). It is based on taking the derivatives of Eqs. 6 and 7 with respect to $c_k^{(1)}$:

$$c_k^{(1)} = \sigma\left(u_k^{(1)}\right) \qquad (8)$$

$$\frac{du_k^{(1)}}{dt} + \frac{u_k^{(1)}}{\tau} = \eta \sum_{i,j} G_i\, \Psi_k^{(1)}(j,i)\, I_j^{(1)} + \eta\beta \sum_m U_{km}^{(1)} c_m^{(1)} \qquad (9)$$

where the constants τ and η determine the rate of convergence of the system, and the constant β determines the contribution of E_c relative to E_b. A sigmoidal squashing function (σ) is used to limit c to the interval [0, 1]. (Equations 8 and 9 are a simultaneous system of equations.) The neural circuitry required for computing Eqs. 8 and 9 is shown in Fig. 4a. The first term on the

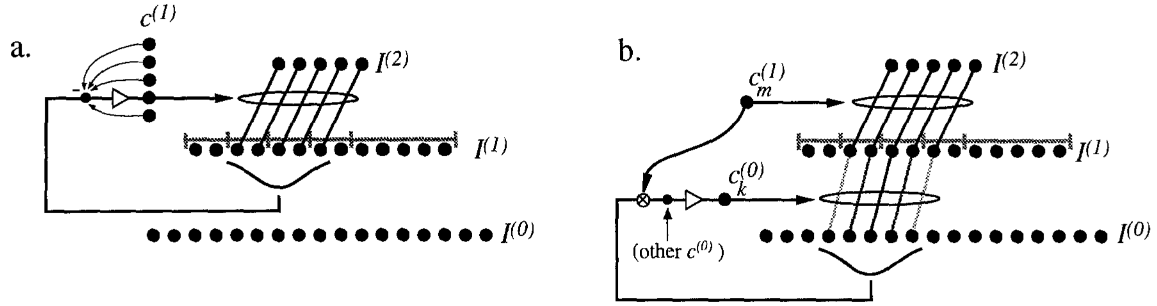

Fig. 4. Autonomous control of a multistage routing circuit. (a) Each control neuron in the top stage has a Gaussian receptive field in level 1 whose position corresponds to the module it gates into $I^{(2)}$. (b) Each control neuron in the first stage, $c_k^{(0)}$, has a Gaussian receptive field in level 0 whose activity is gated by the control neuron in stage 1 that corresponds to the module to which $c_k^{(0)}$ belongs. All five control neurons in the top stage compete among each other, whereas control neurons in the first stage both cooperate and compete in local groups within each module.

right of Eq. 9 is computed by correlating the Gaussian, G, with a shifted version of the intermediate level array, $I^{(1)}$ (the amount of shift depends on the index k). The second term is computed by forming a weighted sum of the activities on the other control neurons. These two results are then summed together and passed through a leaky integrator and squashing function to form the output of the control unit, $c_k^{(1)}$. Thus, each control neuron essentially has a Gaussian receptive field in layer 1, and competition among the control neurons allows only the unit with the strongest input to prevail.

The dynamics for the control neurons in the bottom stage are derived by using the chain rule to take the derivative one step further down, which yields

$$c_k^{(0)} = \sigma\left(u_k^{(0)}\right) \tag{10}$$

$$\frac{du_k^{(0)}}{dt} + \frac{u_k^{(0)}}{\tau} = \eta \sum_{i,j,m,n} G_i c_m^{(1)} \Psi_m^{(1)}(j,i) \Psi_k^{(0)}(n,j) I_n^{(0)} + \eta\beta \sum_m U_{km}^{(0)} c_m^{(0)} \tag{11}$$

This equation essentially states that $c_k^{(0)}$ has a Gaussian receptive field in layer 0, and that the total input from this receptive field is gated by the control neuron in stage 1 that corresponds to the module to which $c_k^{(0)}$ belongs. This is illustrated in Fig. 4b. Control neurons at this stage both compete via inhibitory interactions ($U_{km}^{(0)} < 0$) and cooperate via excitatory interactions ($U_{km}^{(0)} > 0$) locally within a module, as specified above.

Thus, a global selection of position and size is effected through local interactions among the control neurons in a hierarchically organized control circuit. Initially, the activity of the level 1 units is determined by blurring $I^{(0)}$ in the "all connections open" state—that is, with all $c_k^{(0)} = 0$ and $w_{ij}^{(0)} = w_{\text{rest}}$. (Alternatively, without a default connection strength, one could allow each control neuron to have a low, tonically firing resting state.) The stage 1 control neurons will then compete among each other to select the brightest module, or "chunk," of $I^{(1)}$. The winning $c^{(1)}$ will then enable those control neurons in stage 0 belonging to the selected module, and these control neurons will then compete and cooperate locally to position and scale the window of attention within this module. One can alternatively think of the control neurons as being driven by a hierarchical saliency map, with each node in the first level of the saliency map having a Gaussian receptive field in the input, and the second level forming a coarse-grain map by summing and subsampling this map.

Recognition. To derive the control neuron dynamics during recognition, we substitute the blob portion of the objective function, E_b, with a recognition measure, E_m, which we define as

$$E_m = -\sum_{i,j} V_i T_{ij} V_j - \sum_i I_i^{(2)} V_i \tag{12}$$

where the V_i are the activities of associative memory neurons, and the T_{ij} are the coupling coefficients in which the memories are stored (see e.g., Cohen and Grossberg, 1983; Hopfield, 1984). Taking derivatives as before, we find that the activities in the top stage control neurons are determined by

$$c_k^{(1)} = \sigma\left(u_k^{(1)}\right) \tag{13}$$

$$\frac{du_k^{(1)}}{dt} + \frac{u_k^{(1)}}{\tau} = \eta \sum_{i,j} V_i \Psi_k^{(1)}(j,i) I_j^{(1)} + \eta\beta \sum_m U_{km}^{(1)} c_m^{(1)} \tag{14}$$

and the control neurons in the bottom stage are determined by

$$c_k^{(0)} = \sigma\left(u_k^{(0)}\right) \tag{15}$$

$$\frac{du_k^{(0)}}{dt} + \frac{u_k^{(0)}}{\tau} = \eta \sum_{i,j,m,n} V_i c_m^{(1)} \Psi_m^{(1)}(j, i) \Psi_k^{(0)}(n, j) I_n^{(0)} + \eta\beta \sum_m U_{km}^{(0)} c_m^{(0)} \tag{16}$$

These equations are exactly the same as (8, 9) and (10, 11), with the exception that V_i replaces G_i. The significance of this difference is that the V_i are dynamic variables, and so we cannot simply incorporate their multiplicative effect into a fixed weight as we did for the G_i previously. Thus, the top-stage control neurons, $c_k^{(1)}$, will be driven by the correlation between the level 1 nodes, $I_j^{(1)}$, and memory outputs, V_i, that are connected via that control neuron. The bottom-stage control neurons, $c_k^{(0)}$, are driven by correlating the inputs, $I_n^{(0)}$, and memory outputs, V_i, that are connected via $c_k^{(0)}$, and gating the result by the appropriate control neuron in the stage above. Alternatively, we can rewrite the first term on the right of Eq. 16 in a simpler form as $\eta \sum_j \sum_m V_j^{(1)} \Psi_k^{(0)}(m, j)\, I_m^{(0)}$, where $V_j^{(1)} = \sum_i V_i w_{ij}^{(1)}$ is the result of routing the output of the associative memory, V, "backwards" into a separate population of neurons in level 1. In other words, $V^{(1)}$ is a fed-back template of what is expected in the intermediate level. In this case, the activity of $c_k^{(0)}$ would be driven by correlating the inputs, $I_m^{(0)}$, and the fed back signals in level 1, $V_j^{(1)}$, that are in the same position as the level 1 nodes connected to $I_m^{(0)}$ via $c_k^{(0)}$. This then circumvents the need for a long-distance feedback connection from the high-level (V) directly to the low-level control neurons.

Switching between Modes. The main difference between the "blob search" and "recognition" modes is that in the former case the control neurons are driven by fixed, gaussian receptive fields, whereas in the latter case the control neurons are driven by local correlations between groups of associative memory outputs and groups of inputs. Switching from one mode to the other thus requires switching between these two sources of input. This is a global switching process, as opposed to the highly specific and localized switching performed by the routing control neurons, and can be mediated by a global control system that alternates between two modes. For example, there may be two sets of *mode control* neurons that gate the two alternate sources of input feeding into the routing control neurons. The two sets of mode control neurons could then be made to oscillate between alternate states of one set being active and the other set inactive by having delayed, inhibitory connections between the two sets. The delay would need to be long enough to allow the routing control neurons to settle to the steady state in both blob search and recognition modes.

Shifting Attention. The method used for shifting the locus of attention will depend on how the default state of the weights is determined in lower stages of the routing circuit. In the case where the control neurons have a low, tonically active resting state, we can simply inhibit those control neurons in the first stage corresponding to the currently attended locus. This will then prevent any activity from showing up in $I^{(1)}$ and subsequently being used to attract attention. In the case where the control neurons have a default state of zero and $w_{ij} = w_{\text{rest}}$, it will not be feasible to simply inhibit the first stage control neurons corresponding to the attended locus. This is because the saliency in $I^{(1)}$ is being computed independent of the control neuron activities and will still register these locations as interesting. Thus, a node in the first level saliency map must receive a delayed inhibitory signal from the currently active control neuron corresponding to its position. A third alternative is that the top-stage control neurons may be self-inhibited weakly, or with a fast time constant, and the bottom-stage control neurons self-inhibited strongly, or with a slow time constant. This way, attention would more likely be drawn to an object that is far away from (or a different size than) the currently attended object, but would go back to revisit neighboring objects after a sufficient delay period.

Multiscale Case. We now return to the multiscale stack circuit in which there are three different routing streams corresponding to different spatial scales (Fig. 3). The dynamics for the control neurons at the top stage of this circuit can be derived following the same steps as before. Thus, $c_s^{(2)}$ will have a Gaussian receptive field in level 2 of scale s. Or, in terms of the saliency map, $c_s^{(2)}$ is driven by the sum of activity in the top-level saliency map for scale s. Control of attention thus begins at the top stage (stage 2); control neurons here will compete to select the scale with the most salience, and the winning control neuron will enable control neurons below to select the most salient module within that scale, and finally the position and size within that module.

In order to make the comparison between saliencies at different scales meaningful, the shape of the saliency function, G, needs to be modified to be appropriate for selecting a particular scale. As it stands, the saliency nodes for the smallest scale will respond equally well or better to part of a large object as compared to a small object alone. The actual objective we seek during blob search is to just fill the window of attention with a blob that is confined within the bounds of the window. A simple-minded scheme for expressing this objective is to add an inhibitory surround to G, so that objects beyond a certain size are no longer salient at one scale but instead become salient at the next higher scale. In general, more sophisticated forms of saliency detection will be required in order to pre-attentively segment objects of different sizes in a realistic manner, but this topic is beyond the scope of this paper.

It may also be desirable to build-in a precedence for global (low-resolution) over local (high-resolution) information by providing the low-resolution circuits with faster time constants. This would have the effect of "canceling out" the larger objects before attending to the small objects.

Simulation

A 2D version of the above model was implemented in computer simulation. In Fig. 5, the circuit is shown attending to a medium-size 'A' at the middle level of the stack (scale 1). The small 'C' was attended in the previous fixation, and thus has been canceled out in the saliency map for scale 0. (The method for shifting attention was based on delayed inhibition to the level 0 saliency nodes.) Switching from blob search mode to recognition mode and back again was accomplished by automatically switching the source of input to the control neurons after 50 iterations ($\approx 2\tau$). Note that most of the activity in level 1 is the result of blurring $I^{(0)}$ in the all-connections-open state ($w_{ij}^{(0)} = w_{\text{rest}}$), except for the attended region of level 1 where the first-stage connections have been refined by the active control neurons. The circuit as a whole is capable of continuously scaling the attentional window over a factor of eight: Objects of sizes 5×5 to 8×8 are attended at scale 0, sizes 9×9 to 17×17 are attended at scale 1, and sizes 18×18 to 40×40 are attended at scale 2. The interpolating circuit for each first-stage module rescales objects ranging in size from 5×5 to 9×9, and has been discussed and demonstrated at greater length elsewhere (Olshausen, 1994).

3. Neurobiological Substrates and Mechanisms

We now turn to the issue of how the model routing circuit we developed in the previous section may be scaled-up to neurobiological proportions, and we propose specific anatomical substrates and mechanisms in the brain of the macaque monkey. We first describe a multiscale stack model for the representation in area V1, and then the neurobiological substrates for routing and control.

The Multiscale "Stack"

In order to propose a quantitative model for a multiscale stack representation in V1, we need to specify 1) the highest resolution available as a function of eccentricity, and 2) the resolution ratio between adjacent lattices of the stack. For the primate visual system, the highest resolution available at a given eccentricity is approximately

$$S(E) = .01(E + 1.3) \text{ deg} \tag{17}$$

where S denotes the average retinal spacing (in one-dimension) between samples nodes at eccentricity E (Van Essen and Anderson, 1990). In two dimensions, each sample node would cover an area of approximately S^2. To infer the resolution ratio between adjacent lattices, we consider the spatial-frequency bandwidths of V1 cells. An efficient coverage of the spatial frequency domain would require that the spacing in spatial frequency be approximately equal to the bandwidth. Since the physiologically determined bandwidths of V1 cells are generally in the range of 1 to 1.5 octaves (De Valois et al., 1982), we will assume that resolution approximately doubles for each successive lattice of the stack (see also Field, 1989, for computational reasons for octave spacing based upon the statistics of natural scenes).

Given these constraints, a stack comprising approximately 6 lattices would suffice to cover the visual field up to ±50° eccentricity (a total of 100°), as illustrated in Fig. 6. This range of eccentricity corresponds roughly to the region of binocular overlap, or about 90% of the surface area of visual cortex (Van Essen et al., 1984). Beyond this range, retinal ganglion cell sample spacing no longer adheres to the linear relationship of Eq. 17 (Drasdo, 1977), and it is even debatable whether objects beyond this size[2] are recognizable as a whole. The highest resolution lattice of the stack has a mean sample spacing of about .015°, while the lowest resolution

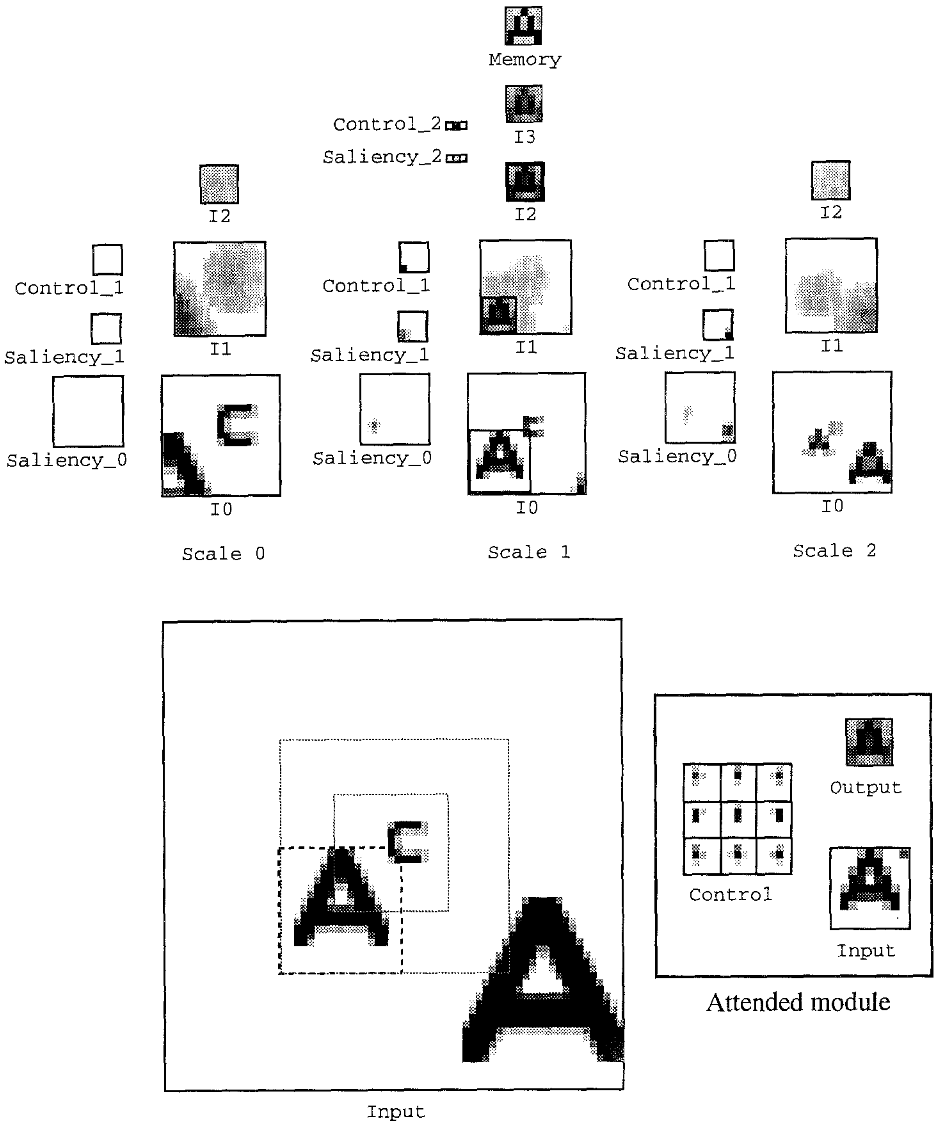

Fig. 5. Simulation of the stack circuit. The input array consists of an array of 68 × 68 sample nodes. There are three lattices in the stack, each comprising an array of 17 × 17 sample nodes and separated in resolution by octaves. The thin dotted lines in the input denote the borders of each lattice. Above the input array are shown the routing circuits for each lattice. The output of the entire routing circuit is shown above this (I3), and at the top is shown the output of the associative memory (Memory). The saliency map and control neurons are shown to the left of each level, with the exception of the first-stage control neurons which are shown only for the attended module (inset to right of the input). The circuit is shown attending to a medium-size (14 × 14) 'A'. The object is rescaled by a factor of two by the sampling lattice of scale 1, reducing it to a size of 7 × 7 by the time it reaches the input to the routing circuit (I0). The object is then remapped by the attended first-stage module into a size of 5 × 5 in I1, where it is then passed up 1:1 to the top level (I3) and successfully recognized by the associative memory. The saliency function (G) used in the simulation was chosen to have a value of +1 within a 7 × 7 window (with a gaussian taper) and −2 within a two-pixel wide perimeter. This choice allowed objects of size 5 × 5 to 8 × 8 within a particular scale to be registered as salient within that scale. If the net input to a saliency node was negative, its output was set to zero.

lattice has a mean spacing of about 0.5°. The number of sample nodes in 1D for each lattice is given by

$$N = \frac{2E}{S(E)} = \frac{2E}{.01(E + 1.3)} \qquad (18)$$

which equals approximately 200 for $E \gg 1°$. At eccentricities near or below one degree the number of nodes within a lattice will be fewer. The total number of sample nodes for the entire stack will thus be on the order of $6 \times 200^2 = 240{,}000$, which is in rough

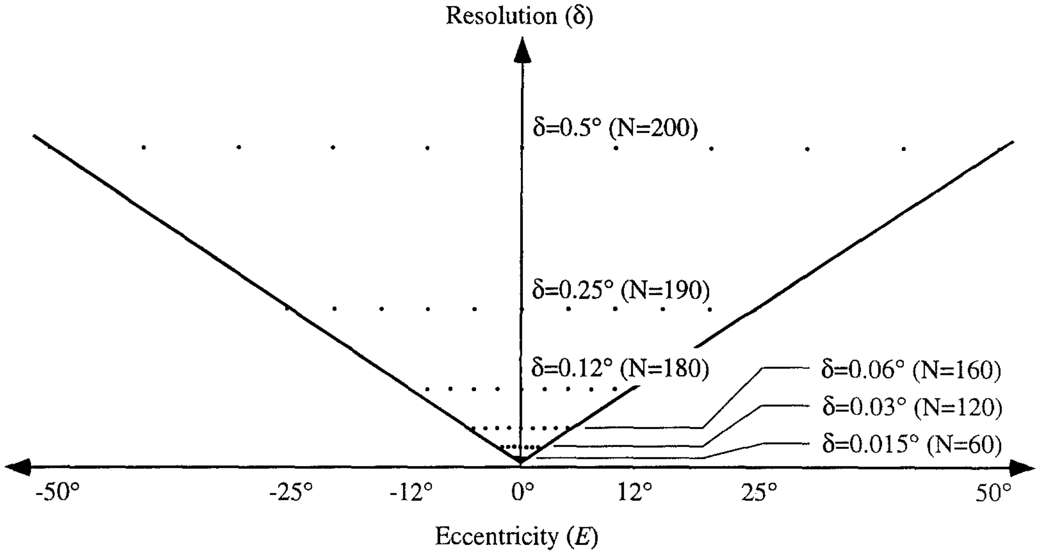

Fig. 6. A six-level "stack" model for V1. Resolution, or sample spacing (δ), is represented along the vertical axis and eccentricity (E) along the horizontal axis. The function $S(E)$ is plotted by the solid lines. The six levels of the stack are represented by the sampling arrays. N denotes the number of nodes in each level in one-dimension. The sample nodes are shown separated by 20 times their actual spacing in order that they remain distinguishable for most levels of the stack.

agreement with the total number of sample nodes delivered by the optic nerve for the central 50° (ca. 80% of the total) when one takes into account the fact that information is divided into on- and off-channels and different spectral bands within the dominant parvo stream (Van Essen and Anderson, 1990). (The number of sample nodes in the stack would actually be expected to be slightly larger than the original number of sample nodes supplied by the optic nerve, due to the addition of multiple scales.)

Within foveal V1 cortex, the highest resolution nodes would have a spacing of about 200 μm (cortical distance), based on physiologically determined estimates of cortical magnification factor (ca. 10–20 mm/deg in the foveal region; Van Essen et al., 1984). The lowest frequency nodes would have a spacing of about 6 mm. With increasing eccentricity, the cortical spacing between low resolution nodes will decrease, and the number of lattices will decrease as well until only the lowest resolution lattice remains. At the largest eccentricity (50°), the spacing between the lowest resolution nodes will be equal to the spacing between the highest resolution nodes in the fovea ($\approx$200 μm). Since the density of sampling nodes decreases by a factor of four (in 2D) for each octave decrease in resolution, the total density of sample nodes in the cortex does not vary greatly with eccentricity, even though there are many more lattices present in the fovea than in the periphery (i.e., there will be a modest over-representation in the fovea.)

It is not necessary that the nodes in the stack model be arranged in a highly uniform, crystalline lattice as depicted in Fig. 6, since the coupling coefficients of the routing circuit (Ψ) can be computed, or learned, once the node positions in scale-space (δ, E) have been specified. Thus, the actual positions of the sample nodes may be scattered about the lattice shown, with an average density in scale space of

$$D(\delta, E) = \begin{cases} \delta^{-2} & \text{if } \delta > S(E) \\ 0 & \text{otherwise.} \end{cases} \qquad (19)$$

where D denotes the number of sample nodes (in one spatial dimension) per square degree of scale-space (δ, E).

Routing Circuit Substrates

Since we are primarily interested in how invariant representations are formed for object recognition, we will focus on the "form" pathway from V1 to IT as the main substrate for routing.[3] The major intermediate visual areas in the form pathway (i.e., V2, V4) are proposed to serve as intermediate stages for routing visual information from the multiscale stack in V1 into representations that are progressively more position and scale invariant at higher stages. The size of each area would be roughly proportional to the number of "sample nodes" in each area, where each node now represents a vector of feature analyses extracted in that area (e.g., a full set of orientation tuning curves in V1), rather than simple luminance values. Thus, a combination of feature analysis, shifting, and rescaling would occur in the intermediate visual areas, leading to a canonical object representation that preserves spatial relationships among features within the object. This canonical representation is hypothesized to exist at early or intermediate stages of the inferotemporal complex (area PIT or CIT). Cells at higher stages (i.e., AIT, or in the STS) would then perform their analyses on the contents of the window of attention, with variations in position and scale removed (e.g., face cells). Since the nature of form processing occurring in the intermediate visual areas is not yet fully understood, we will leave this aspect as an unknown for now and attempt to deal with the issues of routing independently.

The number and sizes of intermediate stages of routing required depend on the total input-output convergence and the maximum allowable fan-in. The input to the routing circuit for each scale will be a 2D array comprising approximately 200 × 200 nodes, as described above. The output of the routing circuit is hypothesized to be a relatively small array, comprising on the order of 30 × 30 sample nodes (an estimate based on spatial acuity and recognition studies that provide hints about the resolution of the window of attention—see Van Essen et al., 1991). Thus, the total convergence for the routing circuit for each scale needs to be about 40,000:1. Since the maximum allowable fan-in is on the order of 10^3–10^4 inputs per neuron (Cherniak, 1990; Douglas and Martin, 1990a), the routing circuit for each scale must be broken into several stages. A nominal configuration would be for each routing circuit to be broken into two stages, as shown in Fig. 7a. This circuit is simply a scaled-up version of the circuit described in the previous section (Fig. 2a), where the first stage is now composed of modules comprising 60 inputs and 30 outputs each. On the right, the circuit is pictured in terms of its fan-out, which is more neurobiologically relevant. Although this architecture does not appear to pose any anatomical problems as shown, it must be kept in mind that the routing circuits for each scale will most likely be superimposed in register in the cortex. Thus, the low resolution nodes will need to have a very great divergence in terms of cortical distance in order to span 30 nodes in the next cortical area. We can reduce the fan-in/fan-out at each stage by breaking the routing circuit into more stages, as shown in Fig. 7b, c. In each of these circuits, the fan-out at each stage is the same but spread over a progressively larger area at higher levels. This is generally consistent with neuroanatomical evidence showing increasing divergence and patchiness in the intracortical connections of higher visual areas (Van Essen and DeYoe 1994; Rockland 1992; Felleman and McClendon 1991; De Yoe et al., 1994). Also, it is interesting to note that direct projections from V1 to V4 have been reported mainly for central visual fields (Yukie and Iwai, 1985; Van Essen et al., 1986), which is consistent with the view that this direct access would be mainly used for routing the high resolution information from the fovea.

Control and Gating Mechanisms

The pulvinar is hypothesized to serve as a major source of the control signals required for routing information, due to its massive interconnectivity with all visual areas, and the lesion and physiological studies that point toward a role in visual attention. Bottom-up sources for driving the control neurons during blob search are hypothesized to arise from a saliency map in either the superior colliculus or the posterior parietal areas, while top-down sources for driving the control neurons during recognition are hypothesized to originate from IT. A variety of biophysical mechanisms provide plausible candidates for gating intracortical connection strengths, such as shunting inhibition, NMDA channels, or voltage gated spikes in dendrites. (See Olshausen et al. 1993 for further discussion of these issues.)

The minimum number of control neurons required for any stage of the routing circuit is given by $(N_B \times F)^2$, where N_B is the number of control blocks along a diagonal in connection space (as in Fig. 2b) and F is the fan-in, in one-dimension. For the first stage of the routing circuit of Fig. 7a, $N_B = 12$ (nominally) and $F = 30$, thus requiring about 130,000 control neurons. For the second stage, approximately $(1 \times 11)^2 = 121$ control neurons would be required. The total number of control neurons required for the other circuits (Fig. 7b, c) would be similar. Thus, the minimal number of control neurons required for routing from all levels of the stack would be on the order of $6 \times 130{,}000 \cong 800{,}000$, which is well within a plausible estimate for the number of neurons in the pulvinar. It should be noted, however, that since there will be a multitude (100–1000) of neurons for each sample node in the routing circuit (corresponding to the

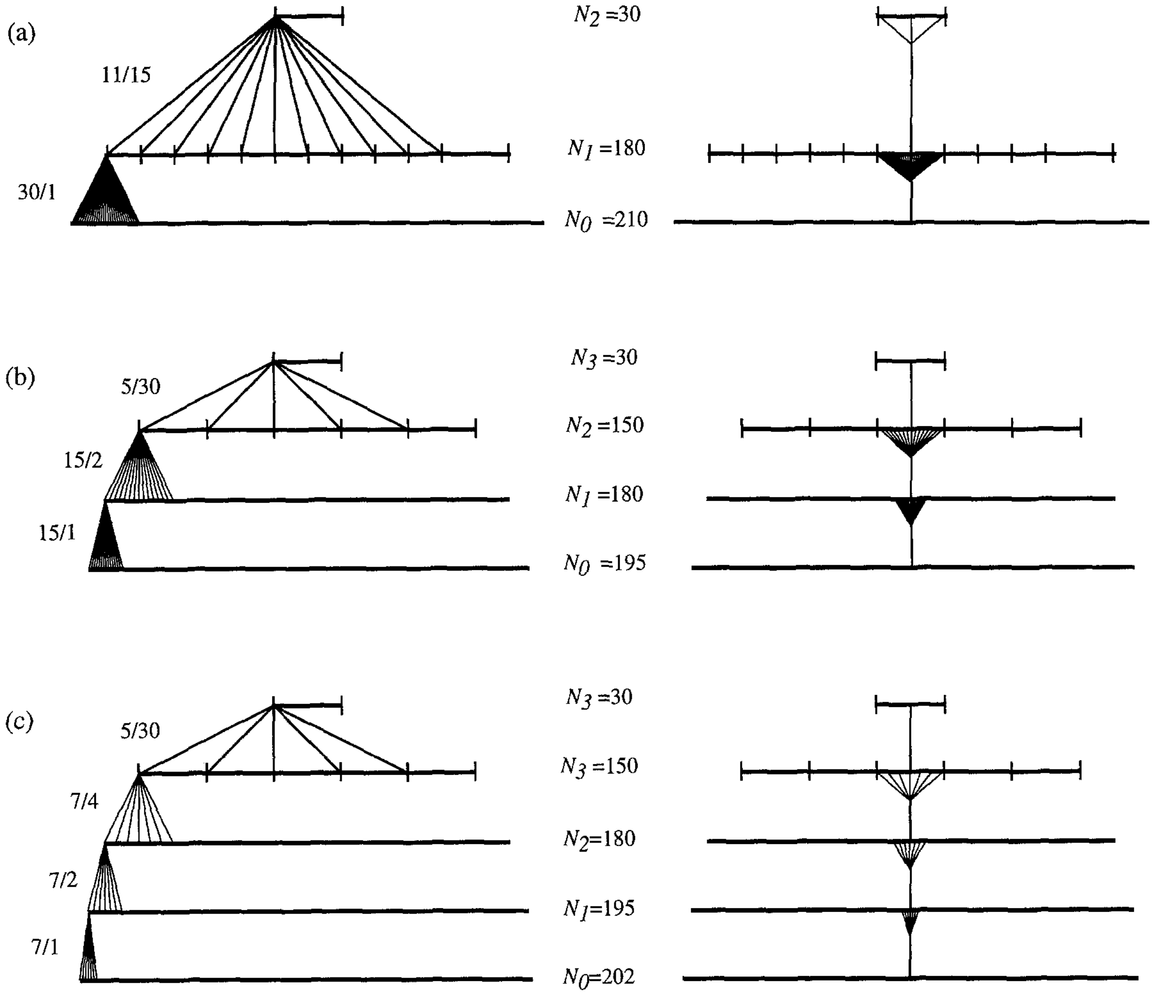

Fig. 7. Some possible multistage routing circuits. The numbers to the left of each stage (x/y) denote the fan-in (x) and the spacing between inputs (y) for the stage. N_l denotes the number of sample nodes in 1D for each layer. The vertical tick marks in the next to last layer denote the borders of the modules of size 30 (equal to the size of the window of attention). Each circuit is redrawn to the right in terms of fan-out. (a) A nominal configuration, corresponding to the circuit of Fig. 2a. The modules here are spaced apart by 15 nodes (half the window size). (b, c) The fan-in/fan-out in the first stage can be reduced by adding more intervening stages.

particular features extracted in each area), the fan-out of each pulvinar control neuron will need to be amplified by additional neurons. This function may possibly be subserved by additional control neurons residing in the deeper layers (5 and 6) of the cortex (Olshausen, 1993; Van Essen and Anderson 1990). Control may thus be distributed between the pulvinar and cortex, with the pulvinar control neurons selecting a general region and scale of the visual field to attend to, and the cortical control neurons acting within this context to take care of the details of routing information between nodes and maintaining spatial relationships within the window of attention.

4. Discussion

Predictions

Stack Model. The stack model predicts that there should exist a large range of spatial-frequency bands represented within foveal V1. Assuming that the optimum spatial-frequency represented on a given lattice is somewhat less than half the sample node frequency, then the highest spatial-frequency band would be expected to be around 15–30 cy/deg (corresponding to the nodes spaced at .015°) and the lowest spatial-frequency band should be around 0.5–1.0 cy/deg (corresponding to the nodes spaced at 0.5°). This predicted range is in reasonable agreement with physiologically determined estimates for foveal V1. For example, De Valois et al. (1982) report that the peak spatial-frequency tunings of foveal V1 cells range from 0.5 to 16 cy/deg. In addition, Tootell et al. (1988) show that the range of spatial frequency bands represented is greatest in the foveal region, and drops off with increasing eccentricity. On the other hand, there is a major discrepancy in the relative numbers of cells predicted in different spatial-frequency bands. In our model, the number of cells at each spatial-frequency band decreases by a factor of four for each octave decrease in resolution. However, the data of De Valois et al. (1982) show that most foveal V1 cells are tuned to peak spatial-frequencies in the range of 2–4 cy/deg, with extremely few cells at 15 cy/deg or above. The data of Parker and Hawken (1988) are somewhat more encouraging, showing that the majority of foveal V1 cells can be fit by a difference-of-difference-of-gaussians function with a central, excitatory zone of about 2–4 minutes in diameter. However, these cells would still be tuned about an octave or so lower than the highest expected spatial-frequency band, and they also do not exhibit anywhere near a factor of four decrease in number for each octave decrease in resolution.

If these data are correct, they imply that the representation of spatial structure at 8 cy/deg and above is very incomplete, and that rather sophisticated deblurring processes must be at work at higher stages of cortical processing in order to allow primates to perceive the world as clearly as they do (Olshausen and Anderson, 1994). On the other hand, there are several reasons to suspect that the incidence of cells tuned for high spatial-frequencies may have been significantly underestimated. Most of the relevant data reported to date have been from anesthetized monkeys (including the studies mentioned above), and it is possible that anesthesia and also sub-optimal optics degrades high spatial-frequency tuning. In addition, short-term adaptation effects, such as those demonstrated by Pettet and Gilbert (1992), may cause cells that are normally very small to become larger (via their horizontal connections) and thereby exhibit low-pass filtering characteristics when they are not being stimulated in a more "natural" environment. Finally, a sampling bias toward low frequency cells might occur because the low spatial-frequency cells will outcompete the high spatial-frequency cells in terms of their duration of firing in response to the bar stimuli typically used as probes during isolation. A possible strategy for overcoming these effects would be to use multi-neuronal recording techniques (e.g., Wilson and McNaughton, 1993) that are capable of isolating many cells simultaneously (i.e., not just those that the experimenter happened to be drawn to) and to record from these cells while the animal is awake and performing a natural and challenging visual task. If these more stringent tests confirm existing physiological reports, then the current theory is wrong and must be reformulated. In particular, it will be necessary to formulate theories for how high spatial resolution information is preserved despite an overwhelming dominance of low spatial-frequency cells.

Routing Circuit. The most obvious prediction of the routing circuit model is that the receptive fields of cortical neurons should change their position or size as attention is shifted or rescaled. This prediction is thus far consistent with the neurophysiological findings of Moran and Desimone (1985) and Connor et al. (1993; 1994a, b) in area V4. In particular, Connor et al. (1994a, b) have mapped the receptive fields of V4 cells while an animal attends to different regions within its visual field, and they have found evidence in many cells for substantial shifts in the receptive field position in the direction of attention. For a third of the cells tested, the maximal response region within the receptive field shifted by an average of half the classical receptive field (CRF) diameter. An important aspect of these data is that the cell's responsiveness cannot be described in terms of a simple function of proximity of the probe stimulus to the attended stimulus. Thus, it provides direct evidence of an actual receptive field shift, as opposed to a simple halo of enhancement surrounding the attended stimulus as would be predicted by "spotlight" models of attention (see below). A preliminary investigation of the effect of size changes revealed no major changes in the receptive field profile for a factor of four change in the size of the attended region (Connor et al., 1993); however, according to our multiscale model, one would expect no more than a factor of two change for a cell in an intermediate stage of the circuit; for size changes beyond this it is difficult to predict what exactly the cell would do (e.g., it may simply relax into the all-connections-open state).

A prediction that arises from the proposed modular structure at higher stages of the routing circuit is that one would expect to see the beginnings of an object-centered reference frame formed in the higher intermediate areas (i.e., V4). If the modules do not overlap, as in Fig. 7b, c, then the receptive field of any particular cell would be expected to maintain a constant position with respect to the window of attention. If the modules do overlap, as in Fig. 7a, then a cell's receptive field may have several distinct positions relative to the window. Interestingly, Connor et al. (1994a, b) have observed that some cells respond preferentially to a bar probe when presented to one side of the attended stimulus, which is consistent with a cell that happens to reside to one side of the center of a non-overlapping module within V4. It will be important to ascertain whether this effect also is evident in inferotemporal cortex.

Since the control neurons of the routing circuit are instrumental for forming invariant object representations, one would expect damage to hypothesized control structures such as the pulvinar to yield deficits in the recognition of objects independent of position and size. Previous lesion studies of the pulvinar have shown that it plays a role in mediating attention, but its role in invariant object recognition has not been properly explored (see Olshausen et al., 1993, for further discussion of this issue).

Finally, a prediction that arises from having a fixed number of sample nodes in the output of the routing circuit is that the resolution within the window of attention should be limited to a fixed number of "pixels" at any size. Thus, increasing the size of the attentional window should result in the same number of pixels being spread over a larger area, with decreased overall resolution (Van Essen et al., 1991). Previous experiments have shown a general decrease in processing efficiency as attentional window size increases (e.g., Eriksen and St. James, 1986; Verghese and Pelli, 1992), but the issue of resolution has not yet been investigated. This aspect of the model could be tested by controlling attentional window size and measuring performance on a task requiring judgments of spatial relationships (e.g., spatial interval discrimination).

Relation to Other Models

The idea of gating connection strengths in order to translate and rescale sensory information is not new. In fact, the general notion has its origin in the early modeling work of Pitts and McCulloch (1947). Several models since then have also proposed gating-type circuits for shifting and rescaling information (e.g., Trehub, 1977; Hinton, 1981; Hinton and Lang, 1985; Baron, 1987; Sandon and Uhr, 1988), but without the requisite ties to neurobiology that allow them to form detailed predictions. On the other hand, neurobiological models of attention, such as those of Niebur and Koch (1994), Desimone (1992), Tsotsos (1994), Ahmad (1992), and LaBerge et al. (1992), have largely been concerned with the general problem of how a select portion of the input is enhanced, as in a spotlight, and do not make clear provisions for preserving information about spatial relationships within the window, which we consider crucial for object recognition. Von der Malsburg and Bienenstock (1986) have proposed methods for remapping information from one array to another using a synchronicity-based short-term binding mechanism; however, in our view it is doubtful that temporal fluctuations of neuronal activity could be coordinated in a neurobiologically plausible fashion in order to establish point-to-point mappings of the type discussed here. Presumably, a unique frequency or phase of bursting would need to be attached to each input node within the attentional window in order to map its activity onto a unique node in the output array, and this would seem to require even more elaborate and precise circuitry than our routing circuit.

Postma (1992; 1994) has proposed a neurobiological model for shifting and rescaling visual information that shares many similarities to our model. The main difference lies in the control structure, in that Postma's network does not utilize control blocks. Rather, control neurons for each synapse are interlinked in a "gating lattice," and the dynamics are formulated within a statistical mechanical framework. Also, Postma's network utilizes a hierarchical control network in which control is propagated bottom-up, instead of top-down as proposed here. The notion of bottom-up control is an interesting and plausible alternative worth exploring further.

A Bayesian Interpretation of the Objective Function

The total objective function of the network, in its most general form, can be written as

$$E_{\text{total}} = -\beta_1 \mathbf{I}^{\text{out}}\mathbf{V} - \beta_2 \mathbf{V}\mathbf{T}\mathbf{V} - \beta_3 \mathbf{c}\mathbf{U}\mathbf{c} \qquad (20)$$

where the boldface variables denote the same quantities as in Section 2 except that they have been converted to vectors and matrices to eliminate the bulky summation signs and indices. We can write $\mathbf{I}^{\text{out}}$ in terms of the input image, $\mathbf{I}^{\text{in}}$, and the control neurons, $\mathbf{c}$, as $\mathbf{I}^{\text{out}} = \mathbf{c}\boldsymbol{\Gamma}\mathbf{I}^{\text{in}}$, where $\boldsymbol{\Gamma}$ is a concatenation of the coupling coefficients, $\Psi_k^l(j, i)$, at all stages of the routing circuit.

The goal of the network is to infer the most probable identities of the objects (WHAT) and their positions and sizes (WHERE) within the image. In other words, we wish the network to maximize $P(\text{WHAT},\text{WHERE} \mid \text{IMAGE})$. Expanding this function according to Bayes' rule gives us

$$\begin{aligned} &P(\text{WHAT, WHERE} \mid \text{IMAGE}) \\ &\quad \propto P(\text{IMAGE} \mid \text{WHAT,WHERE})\, P(\text{WHAT, WHERE}) \\ &\quad = P(\text{IMAGE} \mid \text{WHAT, WHERE})\, P(\text{WHAT})\, P(\text{WHERE}) \end{aligned} \tag{21}$$

where the last step can be taken assuming that WHAT and WHERE are statistically independent (i.e., that any given object is equally likely to appear at any location and size). In the network, the IMAGE is represented by $\mathbf{I}^{\text{in}}$, WHAT is expressed in the ensemble of activity in $\mathbf{V}$, and WHERE is expressed in the ensemble of activity in $\mathbf{c}$. If we make the following equivalences using the Gibb's distribution

$$P(\text{IMAGE} \mid \text{WHAT, WHERE}) \propto e^{\beta_1 V c \Gamma I^{\text{in}}} \tag{22}$$

$$P(\text{WHAT}) \propto e^{\beta_2 V T V} \tag{23}$$

$$P(\text{WHERE}) \propto e^{\beta_3 c U c} \tag{24}$$

then it follows that E_{total} is equivalent to $-\log P(\text{WHAT, WHERE} \mid \text{IMAGE})$ plus a constant. Thus, performing gradient descent on E_{total} will tend to hill-climb on $P(\text{WHAT, WHERE} \mid \text{IMAGE})$. In Bayesian terms, $e^{\beta_1 V c \Gamma I^{\text{in}}}$ acts as the likelihood function for the image (i.e., it measures how well $\mathbf{c}$ and $\mathbf{V}$ can "explain" $\mathbf{I}^{\text{in}}$ by taking the inner product $\mathbf{V c \Gamma} \cdot \mathbf{I}^{\text{in}}$), while the terms $e^{\beta_2 V T V}$ and $e^{\beta_3 c U c}$ act as priors on $\mathbf{V}$ and $\mathbf{c}$.

Finding the WHAT (V) and WHERE (c) that maximize the posterior, $P(\text{WHAT, WHERE} \mid \text{IMAGE})$, requires optimizing over a huge search space. The strategy we adopt in the routing circuit is to first set $\beta_2 = 0$ and let $\mathbf{c}$ evolve while holding $\mathbf{V} = \mathbf{G}$ (blob search). Then we turn on β_2 and let $\mathbf{V}$ evolve. The reason we can do this is that the statistical independence of WHAT and WHERE allows the use of rather primitive, pre-attentive measures to guess the WHERE without knowing WHAT. More generally, though, the pre-attentive measures can also help guess WHAT as well. For example, measures such as color, texture, or primitive shape statistics may be able to narrow down the class of objects to be considered as potential matches. (See Lowe, 1985, for related ideas in machine vision.) In this framework, then, "attention" may be understood as a heuristic that exploits the statistical independence of WHAT and WHERE in order to make an extremely computationally intensive problem tractable with limited resources.

Future Directions

Our eventual goal is to provide a physically realizable and neurobiologically plausible model of visual object recognition. The model presented here represents an incremental step in pursuit of that goal. Some of the more important problems that remain to be solved are outlined here.

Learning. The degree of specificity required in the connectivity of the routing circuit is most likely beyond what can be specified genetically. One must therefore design learning algorithms that could develop or fine tune such an architecture from visual experience. One promising approach has been described by Foldiak (1991), who has shown how a complex cell can learn translation invariance by assuming that the presence of an object (or feature) is stable over time. In our model, stability would be desired in the representation in IT, and the control neurons would need to learn how to configure themselves to maintain the stability of this representation as an object moves or changes size on the retina. The key to doing this, we believe, lies in formulating learning rules for networks with control-like structures, or three-way interactions among units, rather than simple perceptron-type networks with two-way interactions only. Recent work in this direction has shown that networks in which inputs interact in a local, non-linear fashion are capable of learning higher-order regularities—such as disparity—using a local Hebb rule (Lee and Olshausen, 1994). We are currently working on extending this type of learning to the control networks proposed here.

Pre-attentive Segmentation. Clearly, something more sophisticated than "blob detection" will be required in order to indicate the presence of objects within more realistic and natural visual environments. Presumably, other salience measures—such as motion or texture gradients—could provide a more robust method for pre-attentively guessing the size and position of potential objects, as evidenced by other work in this area (Anderson, 1985; Milanese, 1993). In addition, it may be necessary to utilize grouping rules based on local shape cues to provide hints as to which image features should be bundled together within the window of attention (e.g., Freeman, 1992; Sajda and Finkel, 1994; Thau, 1994). It will also be important to explore how low-level feature measures can be used to position the attentional window within object space so as to pre-attentively narrow down the number of objects to be considered as potential candidates during recognition.

Recognition. The simple model of recognition employed here utilizes only second-order correlations between image pixels for storing objects. More realistically, higher-order correlations among features within the window could be used to improve memory capacity (e.g., combinatorial codes). It will also be necessary to provide a richer description of spatial structure, beyond mere luminance values. Work on formulating efficient visual coding strategies based on the redundancies that naturally occur in the environment provides one promising approach (e.g., Field, 1994; Atick, 1992).

5. Conclusions

We have shown in this paper how a multiscale representation may be incorporated advantageously into a routing circuit architecture for forming position- and scale-invariant representations of visual objects, and we have described a solution for the autonomous control of a multistage circuit using a hierarchical control circuit. The proposed neurobiological substrates and mechanisms lead to a number of predictions, such as shifting receptive fields and object-centered representations, that are amenable to experimental tests. In addition, the stack model presented here advances upon the earlier version of Koenderink and Van Doorn (1978) by proposing a specific neurobiological substrate, area V1, for the multiscale representation of visual information. The model exposes certain gaps in our understanding of the spatial frequency organization in primary visual cortex that motivate the need for additional data in this area.

Acknowledgments

Conversations with Chris Lee were helpful in developing the Bayesian interpretation of the objective function. We also thank David Field, Ken Miller, and two anonymous reviewers for helpful comments on the manuscript. Work was supported by ONR grant N00014089-J-1192, and postdoctoral NIH grant T32-MH19389 (BAO).

Notes

1. This is similar to the constraint matrix utilized in the Marr/Poggio stereo algorithm (Marr and Poggio, 1976).
2. 100° would be the extent of a fully stretched hand when held 3 inches from the eye.
3. More generally, one can conceive of dynamic routing taking place in the other visual processing streams as well—for example, in the motion pathway for making fine discriminations of motion (Nowlan and Sejnowski, 1993; Van Essen and Anderson, 1990).

References

Ahmad S (1992) VISIT: A neural model of covert visual attention. In: JE Moody, SJ Hanson, and RP Lippman, eds. Advances in Neural Information Processing Systems 4. Kaufmann, San Mateo, CA, pp. 420–427.

Anderson CH, Burt PJ, and van der Wall GS (1985) Change detection and tracking. In: *SPIE Vol. 579—Intelligent Robots and Computer Vision*, pp. 72–78.

Anderson CH and Van Essen DC (1987) Shifter circuits: A computational strategy for dynamic aspects of visual processing. *Proceedings of the National Academy of Sciences*, USA 84:6297–6301.

Atick JJ (1992) Could information theory provide an ecological theory of sensory processing? *Network* 3:213–251.

Baron RJ (1987) The Cerebral Computer. Erlbaum.

Buhmann J, Lades M, and von der Malsburg C (1990) Size and distortion invariant object recognition by hierarchical graph matching. In: *Proceedings of the International Joint Conference on Neural Networks*, San Diego pp. 411–416.

Cherniak C (1990) The bounded brain: Toward quantitative neuroanatomy. *Journal of Cognitive Neuroscience* 2(1):58–68.

Cohen MA and Grossberg S (1983) Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems, Man and Cybernetics* 13(5):815–826.

Connor CE, Gallant JL, and Van Essen DC (1993) Effects of focal attention on receptive field profiles in area V4. *Soc. Neurosci. Abstr.* 19, p. 974.

Connor CE, Gallant JL, and Van Essen DC (1994a) Modulation of receptive field profiles in area V4 by shifts in focal attention. *Invest. Opthal. Vis. Sci.* 35, p. 2147.

Connor CE, Gallant JL, and Van Essen DC (1994b) Dynamic modulation of receptive field profiles in area V4. Submitted for publication.

Desimone R (1992) Neural circuits for visual attention in the primate brain. In: GA Carpenter and S Grossberg, eds. Neural Networks for Vision and Image Processing. MIT Press, Cambridge, Mass, pp. 343–364.

De Valois RL, Albrecht DG, and Thorell LG (1982) Spatial frequency selectivity of cells in macaque visual cortex. *Vision Res.* 22: 545–559.

DeYoe EA, Felleman DJ, Van Essen DC, and McClendon E (1994) Multiple processing streams in occipito-temporal visual cortex. *Nature* 371:151–154.

Douglas RJ and Martin KAC (1990a) Neocortex. In: GM Shepard, ed. Synaptic Organization of the Brain. Oxford UP, New York, pp. 389–438.

Drasdo N (1977) The neural representation of visual space. *Nature* 266:554–556.

Eriksen CW and St. James JD (1986) Visual attention within and around the field of focal attention: A zoom lens model. *Perception and Psychophysics* 40(4):225–240.

Felleman DJ and McClendon E (1991) Modular connections between area V4 and temporal lobe area PITv in macaque monkeys. *Society for Neuroscience Abstracts* 17:1282.

Field DJ (1989) What the statistics of natural images tell us about visual coding. SPIE Vol. 1077 Human Vision, Visual Processing, and Digital Displays, pp. 269–273.

Field DJ (1994) What is the goal of sensory coding? *Neural Computation* 6:559–601.

Foldiak P (1991) Learning invariance from transformation sequences. *Neural Computation* 3:194–200.

Freeman WT (1992) Steerable filters and local analysis of image structure, Ph.D. thesis, MIT Media Laboratory.

Fukushima K (1980) Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* 36:193–202.

Hinton GE (1981) A parallel computation that assigns canonical object-based frames of reference. In: *Proceedings of the Seventh International Joint Conference on Artificial Intelligence* 2, Vancouver B.C., Canada.

Hinton GE and Lang KJ (1985) Shape recognition and illusory conjunctions. In: *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, Los Angeles.

Hopfield JJ (1984) Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Natl. Acad. Sci.* USA 81:3088–3092.

Koch C and Ullman S (1985) Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology* 4:219–227.

Koenderink JJ and van Doorn AJ (1978) Visual detection of spatial contrast; Influence of location in the visual field, target extent and illuminance level. *Biological Cybernetics* 30:157–167.

LaBerge D, Carter M, and Brown V (1992) A network simulation of thalamic circuit operations in selective attention. *Neural Computation* 4:318–331.

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, and Jackel LD (1990) Backpropagation applied to handwritten Zip code recognition. *Neural Computation* 1: 541–551.

Lee CW and Olshausen BA (1994) A nonlinear hebbian network that learns to detect disparity in random-dot stereograms. Technical Report 94-20, Philosophy, Neuroscience and Psychology program, Washington University, St. Louis, MO. In review.

Lowe DG (1985) Perceptual organization and visual recognition. Boston: Kluwer.

Marr D and Poggio T (1976) Cooperative computation of stereo disparity. *Science* 194:283–287.

Milanese R (1993) Detecting salient regions in an image: From biological evidence to computer implementation. Ph.D. thesis, Computer Science Dept., University of Geneva.

Moran J and Desimone R (1985) Selective attention gates visual processing in the extrastriate cortex. *Science* 229:782–784.

Niebur E and Koch C (1994) A model for the neuronal implementation of selective visual attention based on temporal correlation among neurons. *Journal of Computational Neuroscience* 1: 141–158.

Nowlan SJ and Sejnowski TJ (1993) Filter selection model for generating visual motion signals. In: SJ Hanson, JD Cowan, and CL Giles, eds. *Advances in Neural Information Processing Systems*, 5. Morgan-Kaufmann, San Mateo, CA, pp. 369–376.

Olshausen BA (1994) Neural routing circuits for forming invariant representations of visual objects. Ph.D. thesis, Computation and Neural Systems Program, California Institute of Technology.

Olshausen BA, Anderson CH, and Van Essen DC (1993) A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience* 13:4700–4719.

Olshausen BA and Anderson CH (1994) A model of the spatial frequency organization in primate visual cortex. Paper presented at CNS*94, Monterey, CA, (proceedings in press).

Parker AJ and Hawken MJ (1988) Two-dimensional spatial structure of receptive fields in monkey striate cortex. *Journal of the Optical Society of America* A 5:598–605.

Pettet MW and Gilbert CD (1992) Dynamic changes in receptive-field size in cat primary visual cortex. *Proc. Natl. Acad. Sci.* USA 89:8366–8370.

Pitts W and McCulloch WS (1947) How we know universals: The perception of auditory and visual forms. *Bulletin of Mathematical Biophysics* 9:127–147.

Postma EO, van den Herik HJ, and Hudson PTW (1992) The gating lattice: A neural substrate for dynamic gating. In: *CNS*92 Proceedings*, July 26–29, San Francisco, California. Kluwer Academic Publishers.

Postma EO (1994) SCAN: A neural model of covert attention. Ph.D. thesis, Computer Science Dept., University of Limburg, Maastricht, The Netherlands.

Rockland KS (1992) Configuration, in serial reconstruction, of individual axons projecting from area V2 to V4 in the macaque monkey. *Cerebral Cortex* 2:353–374.

Sajda P and Finkel LH (1994) Dual mechanisms for neural binding and segmentation. In: JD Cowan, G Tesauro, and J Alspector, eds. *Advances in Neural Information Processing Systems*, 6. Morgan-Kaufmann, San Francisco, CA, pp. 993–1000.

Sandon PA and Uhr LM (1988) An adaptive model for viewpoint-invariant object recognition. *Proceedings of the 10th Annual Conference of the Cognitive Science Society*, Montreal, Canada, August, pp. 209–215.

Schwartz EL (1977) Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception. *Biological Cybernetics* 25:181–194.

Thau R (1994) Visual segmentation and feature binding without synchronization. Paper presented at CNS*94, Monterey, CA, (proceedings in press).

Tootell BH, Silverman MS, Hamilton SL, Switkes E, and De Valois RL (1988) Functional anatomy of macaque striate cortex. V. Spatial Frequency. *The Journal of Neuroscience* 8:1610–1624.

Trehub A (1977) Neuronal models for cognitive processes: Networks for learning, perception and imagination. *J Theor. Biol.* 65: 141–169.

Tsotsos JK (1994) Towards a computational model of visual attention. In: T Papathomas, ed. Early Vision and Beyond. MIT Press, Cambridge, Mass.

Van Essen DC and Anderson CH (1990) Information processing strategies and pathways in the primate retina and visual cortex. In: SF Zornetzer, JL Davis, C Lau, ed. An Introduction to Neural and Electronic Networks. Academic, New York, pp. 43–72. (2nd Edition in press).

Van Essen DC and DeYoe EA (1994) Concurrent processing in the primate visual cortex. In: MS Gazzaniga, ed. The Cognitive Neurosciences. MIT Press, Cambridge, MA, pp. 383–400.

Van Essen DC, Newsome WT, and Maunsell JHR (1984) The visual field representation in striate cortex of the macaque monkey: Asymmetries, anisotropies, and individual variability. *Vision Research* 24(5):429–448.

Van Essen DC, Newsome WT, Maunsell JHR, and Bixby JL (1986) The projections from striate cortex (V1) to areas V2 and V3 in the macaque monkey: asymmetries, areal boundaries, and patchy connections. *J. Comp. Neurol.* 244:451–480.

Van Essen DC, Olshausen B, Anderson CH, and Gallant JL (1991) Pattern recognition, attention, and information bottlenecks in the primate visual system. In: BP Mathur, C Koch, eds. Proc. SPIE Conf. on Visual Information Processing: From Neurons to Chips, Vol. 1473. SPIE, Bellingham, WA, pp. 17–28.

Verghese P and Pelli DG (1992) The information capacity of visual attention. *Vision Research* 32(5):983–995.

von der Malsburg C and Bienenstock E (1986) Statistical coding and short-term synaptic plasticity: A scheme for knowledge representation in the brain. In: E Bienenstock, Soulie F Fogelman , G Weisbuch, eds. Disordered Systems and Biological Organization (NATO ASI Series, Vol. F20). Springer, Berlin, pp. 247–272.

Wilson MA and McNaughton BL (1993) Dynamics of the hippocampal ensemble code for space. *Science* 261:1055–1058.

Yukie M and Iwai E (1985) Laminar origin of direct projection from cortex area V1 to V4 in the rhesus monkey. *Brain Research* 346:383–386.