COMPUTER VISION AND IMAGE UNDERSTANDING
Vol. 69, No. 1, January, pp. 1–22, 1998
ARTICLE NO. IV970561

The What-and-Where Filter

A Spatial Mapping Neural Network for Object Recognition and Image Understanding

Gail A. Carpenter,* Stephen Grossberg,† and Gregory W. Lesher‡

Center for Adaptive Systems and Department of Cognitive and Neural Systems, Boston University, 677 Beacon Street, Boston, Massachusetts 02215

Received August 20, 1993; accepted October 8, 1996

The What-and-Where filter forms part of a neural network architecture for spatial mapping, object recognition, and image understanding. The Where filter responds to an image figure that has been separated from its background. It generates a spatial map whose cell activations simultaneously represent the position, orientation, and size of all the figures in a scene (where they are). This spatial map may be used to direct spatially localized attention to these image features. A multiscale array of oriented detectors, followed by competitive and interpolative interactions between position, orientation, and size scales, is used to define the Where filter. This analysis discloses several issues that need to be dealt with by a spatial mapping system that is based upon oriented filters, such as the role of cliff filters with and without normalization, the double peak problem of maximum orientation across size scale, and the different self-similar interpolation properties across orientation than across size scale. Several computationally efficient Where filters are proposed. The Where filter may be used for parallel transformation of multiple image figures into invariant representations that are insensitive to the figures' original position, orientation, and size. These invariant figural representations form part of a system devoted to attentive object learning and recognition (what it is). Unlike some alternative models where serial search for a target occurs, a What and Where representation can be used to rapidly search in parallel for a desired target in a scene. Such a representation can also be used to learn multidimensional representations of objects and their spatial relationships for purposes of image understanding. The What-and-Where filter is inspired by neurobiological data showing that a Where processing stream in the cerebral cortex is used for attentive spatial localization and orientation, whereas a What processing stream is used for attentive object learning and recognition. © 1998 Academic Press

* Supported in part by DARPA (ONR N00014-92-J-4015), British Petroleum (BP 89-A-1204), the National Science Foundation (NSF IRI-90-00530), and the Office of Naval Research (ONR N00014-91-J-4100, ONR N00014-95-1-0409, and ONR N00014-95-1-0657).

† Supported in part by the Air Force Office of Scientific Research (AFOSR F49620-92-J-0499), DARPA (ONR N00014-92-J-4015), and the Office of Naval Research (ONR N00014-91-J-4100, ONR N00014-95-1-0409, and ONR N00014-95-1-0657).

‡ Supported in part by the Air Force Office of Scientific Research (AFOSR F49620-92-J-0334), DARPA (AFOSR 90-0083), a National Science Foundation Graduate Fellowship, and the Office of Naval Research (ONR N00014-91-J-4100).

1. INVARIANT FILTERING FOR OBJECT RECOGNITION AND IMAGE UNDERSTANDING

A typical pattern recognition problem requires that an object be identifiable at various positions, sizes, and orientations. A representation of the object that is invariant with respect to these properties is often computed at a preprocessing stage. For example, a combination of Fourier and log–polar transforms has been used to provide translation, scale, and rotation invariance [1, 2]. The output of log–polar-Fourier preprocessing is an invariant representation, but one that has lost information about the form of the object, as well as about the object's place in a larger scene. This article introduces a filter-based invariant transform system in which information about the position, size, and orientation of the object is retained, and no form information is lost.

The strategy leading to this system is suggested by the brain's use of parallel streams in the visual cortex to compute Where an object is and What the object is [3, 4]. Goodale and Milner [5] have proposed, moreover, that the Where processing stream sets the stage for commanding motor actions towards targets. The What processing stream includes such brain regions as the visual cortical area V4 and inferotemporal cortex. The Where processing stream includes visual cortical area MT and parietal cortex.

The neural network defined below consists of a Where channel that simultaneously computes the position, orientation, and size of all target figures, and a What channel that uses the information provided by the Where channel to encode invariant object representations of all target figures. Subsequent recognition of individual objects is based upon output from the What channel. More global scenic interpretations, or context-sensitive recognition of ambiguous objects, may be achieved by parallel fusion of data about multiple objects and their spatial relationships from both the What and Where representations.

1077-3142/98 $25.00

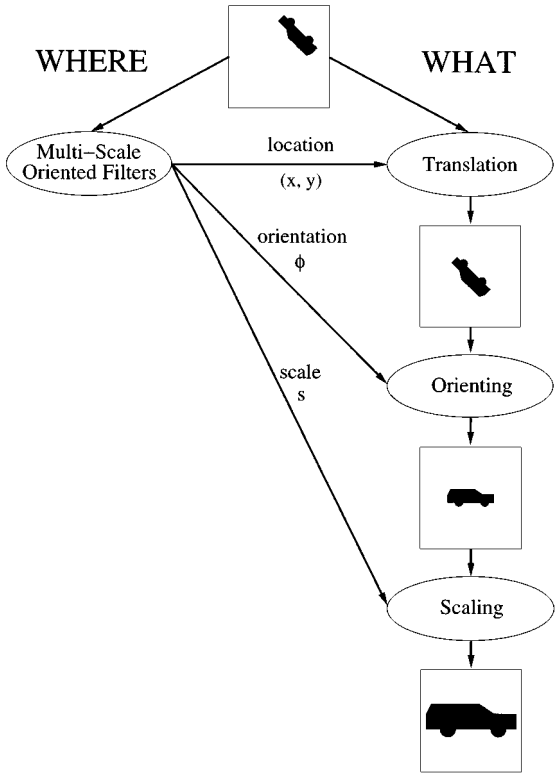

FIG. 1. A What-and-Where filter. The Where filter generates a multiplexed spatial map of a figure's position, orientation, and size. This spatial map is used by the What processing stream to generate an invariant figural representation. This representation is used to learn a recognition category for all figures that are sufficiently similar to one another in their form, at all possible positions, orientations, and sizes.

The Where channel includes banks of spatial filters of varying sizes and orientations. Suitably defined competition between filters yields a spatial map whose cell activations multiplex a representation of the position, orientation, and size of all target figures. The What-and-Where filter may thus be realized as a one-pass algorithm that preattentively generates information about all the objects in a scene. Such a one-pass algorithm can rapidly prepare image data for attentive recognition and search processes that interact reciprocally with the What-and-Where representations. In particular, the Where representation for each object is used to transform the representation of that object within the What stream so that it is centered at the origin with canonical size and horizontal orientation. Figure 1 illustrates the main computations of the What-and-Where filter.

The What-and-Where filter is one processing stage in a family of multistage architectures that are designed to accomplish automatic visual pattern recognition and image understanding. Six stages of such an architecture are depicted in Fig. 2. The first stage compensates for variable illumination in a scene. The second stage generates a boundary segmentation of the image that completes and regularizes incomplete figural boundaries while suppressing image noise. The third stage separates the figures of the image from each other and from the image background onto slabs on which individual figures are isolated. The fourth stage is the Where filter. Here each slab contributes to a spatial map of its figure's position, orientation, and size. This spatial map is used to generate an input figure to the fifth stage, the What representation, that is invariant under two-dimensional changes of position, orientation, and size. This stage also consists of multiple channels, one for each slab, that interact with a self-organizing pattern recognition system at stage six. This system learns to categorize the 2D invariant figures in its channel. In particular, 2D view categories of each object can be learned and fused into an invariant 3D object representation. The last stage carries out more complex predictions of image understanding by combining information about what the objects are from stage six, with information from stage four about their spatial relations with respect to each other.

Neural networks that realize the functional requirements of stages one, two, three, and six have previously been developed. These networks use a consistent computational format that will enable them to be combined into a larger system architecture. Each of these stages has been derived from an analysis of perceptual and neural data aimed at discovering how the brain accomplishes similar computational goals. The present article describes a network for stage four that will provide a foundation for combining stage four, five, and six computations into a global scenic interpretation at stage seven.

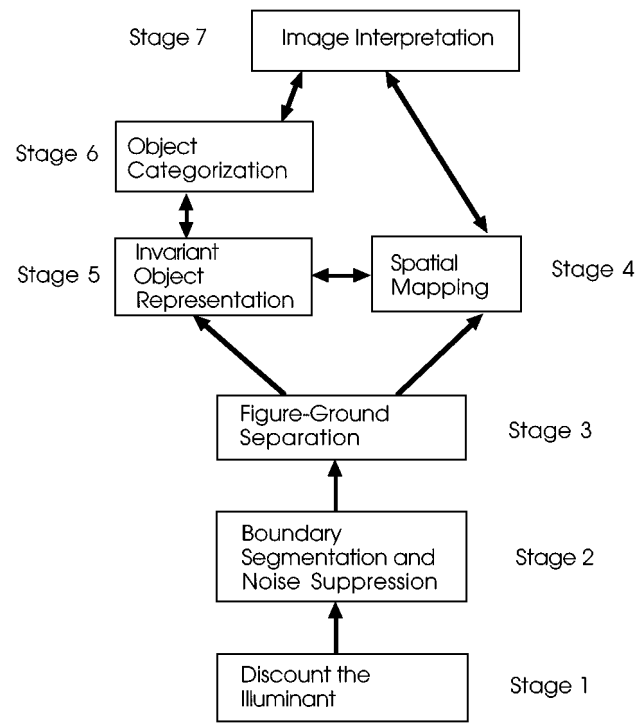

FIG. 2. Processing stages of a multistage architecture for pattern recognition and image understanding.

In particular, compensating for variable illumination (also called discounting the illuminant) can be carried out at stage one by a shunting on-center off-surround network [6, 7]. The Grossberg–Todorović model has been realized as a VLSI chip by Andreou and Boahken [8] using a retinal circuit like the one described in Grossberg [9, Section 25]. Coherent boundary segmentation and noise suppression can be accomplished at stage two by a boundary contour system or BCS [10–14]. A simplified version of the BCS, called the CORT-X filter, accomplishes boundary segmentation and noise suppression using only fast feedforward operations [15, 16].

Figure–ground separation of the figures in a 2D image can be accomplished by a model that is called an FBF filter because it combines boundary segmentation and noise suppression operations (B) with illumination compensation and surface filling-in operations (F) in the order FBF [16, 17]. An FBF model is capable of simultaneously separating all figures with connected boundaries from one another and the background. Figure 3 provides an example of FBF separation applied to a laser radar image. Such a model is often sufficient to carry out figure–ground separation in scenes wherein important targets are not partially occluded by other targets. In cases wherein partial occlusions do occur, a more general FACADE model of 3D pop-out of figures from their backgrounds, and completion of partially occluded targets, in response to both 2D images and 3D scenes may be used [18–21].

The operations at stages one, two, and three may all be called *preattentive* visual mechanisms because they are applied in parallel to all image data, whether familiar or unfamiliar. Attentional mechanisms select among, and bind together, various of these image representations. Attentive object learning and categorization can be accomplished at stage six by adaptive resonance theory, or ART, networks that may operate either in an unsupervised mode, as with ART 1, ART 2, and Fuzzy ART [22–25], or a supervised mode, as with ARTMAP, Fuzzy ARTMAP, Fusion ARTMAP, and Gaussian ARTMAP [26–30]. It has also been shown how Fuzzy ARTMAP can be used to automatically learn invariant 3D representations of objects from their 2D views, as in the ART-EMAP [31, 32] and VIEWNET [33, 34] architectures.

Several properties of a What-and-Where filter make it an appealing candidate for a stage four invariant filter. For one, the Where filter uses oriented receptive fields of multiple sizes that compete across position, size, and orientation. Such multiscale competitive interactions are also used in the BCS, CORT-X, and FBF networks. The Where filter operations that determine *spatial* properties of image figures are thus variations of operations used at earlier processing stages to determine *visual* properties of image figures. This computational homolog facilitates the choice of consistent parameters in the full multistage architecture. It also highlights the research question of how replication of a shared set of competitive modules may be realized in applications and *in vivo* to carry out both visual and spatial computations.

A second useful property is that the filter may be designed to operate in a one-pass mode, or an efficient serial algorithm; hence, it is capable of fast response in image processing applications. Various other recent approaches to generating spatially invariant representations use multistage concurrent bottom-up and top-down operations, or complex relaxation methods, that are more computationally expensive and time-consuming. (See Section 14 for further discussion.) Such approaches also typically attempt to focus attention upon a single target at a time using the same operations that put it into an invariant representation. These approaches have difficulty explaining how an important target may be quickly recognized if the initially chosen targets are the wrong ones.

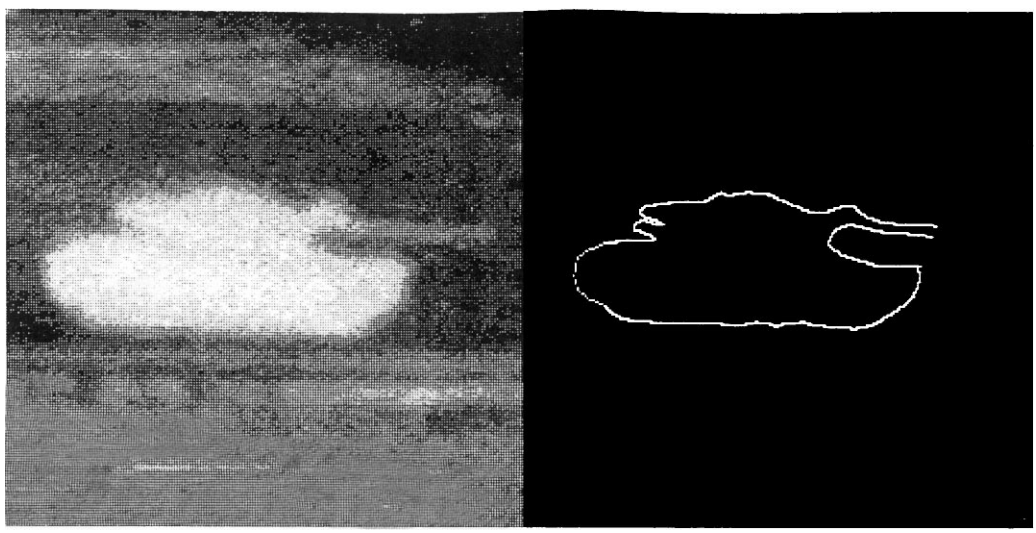

FIG. 3. An example of figure-ground separation of a target from a laser radar image using and FBF filter in stages one to three of Fig. 2. (Reprinted with permission from Grossberg and Wyse [17].)

The present model generates spatial representations and invariant representations preattentively for all targets. Attentive mechanisms can thus begin at once to search for any of them using higher-level knowledge. In particular, an ART architecture can be primed to rapidly recognize a desired target on any of the slabs. Grossberg [19] and Grossberg, Mingolla, and Ross [35] have shown how a What-and-Where representation of the type described here can be used in a visual search algorithm, called the SOS, or Spatial Object Search, model that has been used to quantitatively simulate psychophysical data from human visual search experiments. These experiments show that humans exhibit properties of parallel search in many more viewing conditions than previously realized (see Section 14).

A third useful property of a What-and-Where filter can be exploited in image understanding. The Where filter defines a spatial map whose nodes multiplex information about the position, size, and orientation of every figure in an image. In particular, activation of a node, or cell population, in this map implicitly represents all three spatial properties of the corresponding image figure. The Where filter nodes are thus distinct channels that each process at most one figure. Each channel, in turn, inputs to its own What invariant filter and recognition network. Thus the Where map of each figure is linked, or bound, to the corresponding What recognition of the figure, even though the What recognition strips the figure of its spatially variant properties. Due to this linkage, the Where spatial map and the What recognition categories can be combined into a total input vector to the stage seven image interpretation network (Fig. 4). Such a network can learn to combine information about the identities of each object with information about the objects' spatial relationships to derive a more global interpretation of scenic meaning.

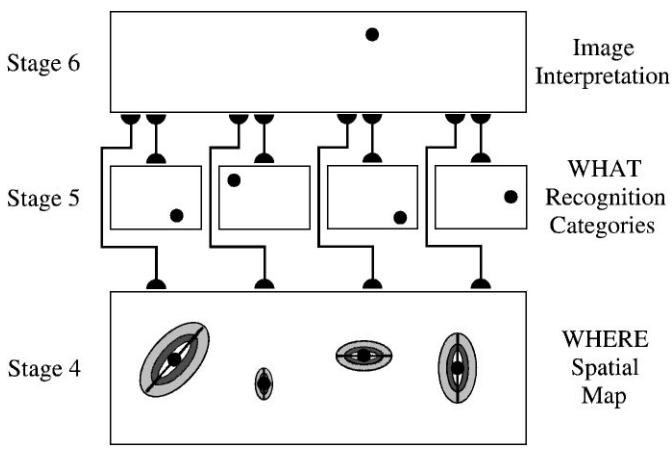

FIG. 4. Reciprocal interactions of a Where spatial map and What recognition categories with an image interpretation network can learn scenic interpretations that combine information about multiple objects and their spatial relations. Fusion ARTMAP (see text) can be used for supervised learning of those combinations of object categories and spatial relations that reliably predict a prescribed scenic interpretation.

A supervised recognition and prediction network, called Fusion ARTMAP, has been designed to handle such problems of multidimensional data fusion, classification, and prediction [26]. Fusion ARTMAP autonomously searches for and learns those combinations of input features that provide the best prediction. In an image understanding application, Fusion ARTMAP may be applied to learn those combinations of spatial and visual information that predict a desired image interpretation.

In the present article, the Where filter is used to generate an invariant What representation in Cartesian coordinates. Variations on this design can also readily be used that include, say, complex logarithmic processing to achieve partial invariance and data compression [36] in much the same way as the cortical magnification factor works in the mammalian visual system [37–41]. Cartesian coordinates are used herein to demonstrate how well the Where filter, operating alone, can create a fully invariant What representation.

Section 2 describes the oriented filter components of the What-and-Where system. Section 3 presents the simplest form of the What-and-Where filter, employing normalized filter elements, or receptive fields, to determine position, orientation, and size using a cascade of competitive parallel operations. The number of receptive fields in the parallel system can be greatly reduced via Gaussian interpolation across coarsely coded orientations, as shown in Section 4. Section 5 presents the equations for the parallel What-and-Where filter algorithm with orientation interpolation, and Section 6 describes a computer simulation of this system employing vehicle silhouette images. The role of receptive field normalization is discussed in Section 7, along with the value of dissociating determination of figure position from that of orientation and size when using unnormalized receptive fields. A parallel What-and-Where filter, modified to employ unnormalized receptive fields, is presented in Section 8. Section 9 shows how a hybrid serial-parallel system, which first calculates orientation and then size, can dramatically reduce the computational load on the filter. Further reduction in the number of filters can be achieved by orientation interpolation, as shown in Section 10. Such a hybrid What-and-Where filter algorithm is mathematically defined in Section 11. In Section 12, system responses to elliptical test images of variable elongation are used to calibrate system parameters. Simulations demonstrating performance of the hybrid system in response to the vehicle silhouette images from an MIT Lincoln Laboratory database are summarized in Section 13. Section 14 compares the What-and-Where approach with alternatives. Section 15 provides concluding remarks and open problems.

2. THE ORIENTED CLIFF FILTER

The Where computations of the What-and-Where filter employ a spatial array of oriented receptive fields with different sizes and orientations that are convolved with the input image. Computation of orientation is based upon receptive fields within an oblong excitatory region. As shown below, the winning

orientation provides a stable measure of an object's net orientation in response to objects of variable shape.

Reliable computation of both orientation and size depend upon the use of a strongly inhibitory surround region around the oblong excitatory region. Such an oriented receptive field, centered at the origin with orientation ϕ degrees, size s pixels, and elongation a, is defined by the kernel

$$K(x, y, \phi, s) = (1 - r^6) \exp\left(-\frac{r^4}{1 + r^2}\right), \quad (1)$$

$$r^2 = \left(\frac{x'}{as}\right)^2 + \left(\frac{y'}{s}\right)^2, \quad (2)$$

$$x' = x \cos\phi + y \sin\phi, \quad (3)$$

and

$$y' = y \cos\phi - x \sin\phi. \quad (4)$$

Figure 5a depicts the geometry of an oriented receptive field and a normalized cross section of kernel values as a function of r. Four examples of receptive fields (with elongation $a = 2$) are shown in Fig. 5b, where white signifies large positive values of $K(x, y, \phi, s)$ and black signifies large negative values. The black ellipse in the left-hand receptive field indicates the set of points where $r^2 = 1$ and $K = 0$. As specified by Eqs. (1)–(4), each receptive field includes a positively weighted elliptical center area bordered by a sharp drop-off to a negatively weighted surround field. It cannot be overemphasized that this cliff-like surround is essential to deriving our results.

Given receptive fields such as those in Fig. 5, the greatest response obviously results from an elliptical input oriented and sized such that it fits perfectly within the central region ($r^2 \leq 1$) of the receptive field. On the other hand, a sizable response will be observed for any anisotropic input that is oriented and scaled such that it stays primarily within the central region. The steep drop to negative values in the region where $r^2 \approx 1$ causes the response of the receptive field to drop sharply when parts of an image fall outside the central region. Throughout the article, the elongation parameter a is set equal to 2. The system still performs well, however, on images whose ratio of width to height is far from 2, as shown in Section 12.

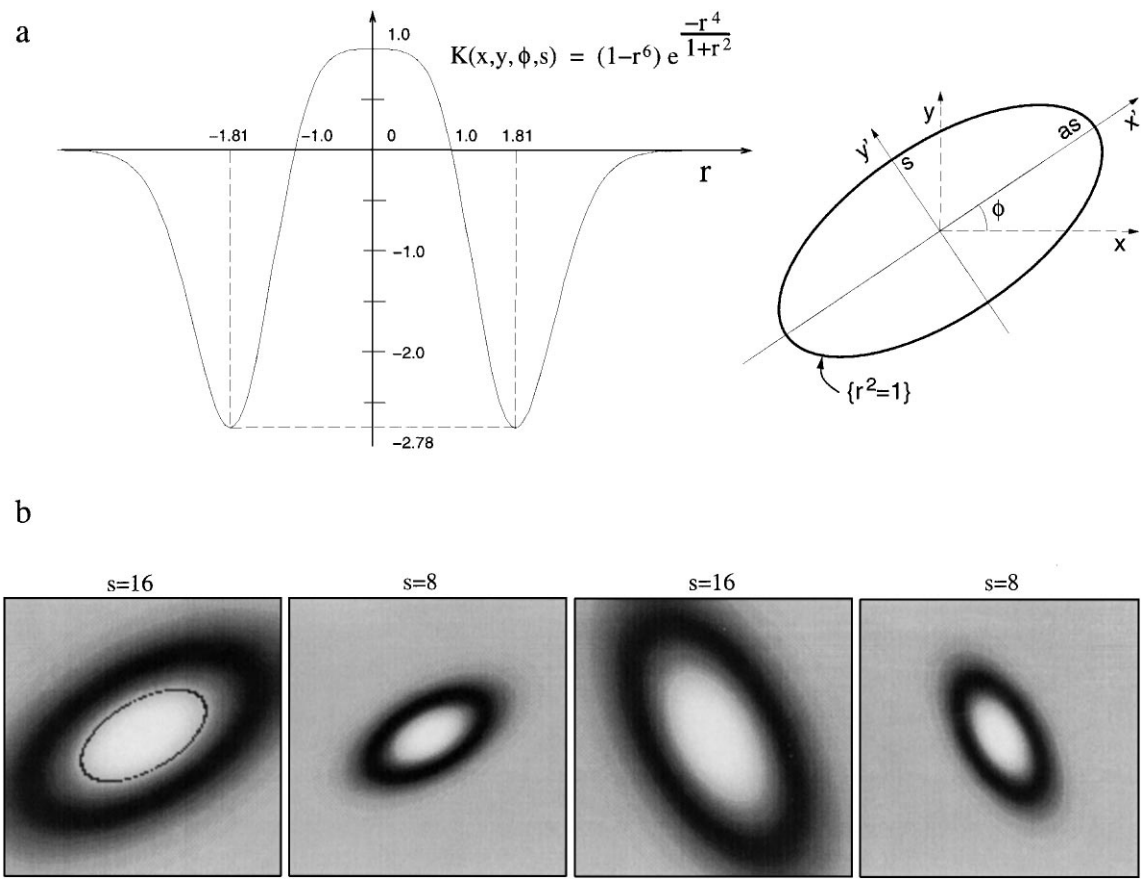

FIG. 5. An oriented on-center off-surround receptive field with a steep cliff-like on–off border can be used as the building block of a Where filter. (a) A region of positive K values, in the ellipse where $r^2 \leq 1$, is surrounded by a region of negative values, as this cross section of the filter depicts. The elongation parameter a is the ratio of the x' axis to the y' axis of the ellipse. (b) Oriented filters at size scales $s = 16$ pixels and $s = 8$ pixels and orientations $\phi = 30^\circ$ and $\phi = 120^\circ$, each in a 128×128 pixel square and with elongation $a = 2$. The ellipse in the left-hand filter shows the set of points, where $r^2 = 1$ for that filter.

3. A ONE-PASS PARALLEL WHERE FILTER

Within the Where channel, the input figure is convolved with each oriented filter. Each filter element is identified with a point (x, y) on a two-dimensional grid of neuron nodes, and its receptive field is centered at (x, y). The activity A of a node located at (x, y) whose receptive field K has orientation ϕ and size s is given by the discrete convolution

$$A(x, y, \phi, s) = \sum_p \sum_q K(p, q, \phi, s) I(p - x, q - y), \quad (5)$$

in response to an input pattern $I(x, y)$.

The convolution between input image and filters of different orientation and size yields a four-dimensional array of neuron nodes: two dimensions correspond to the x and y coordinates of the receptive field center; and one dimension each corresponds to orientation ϕ and size s of the filter. Each of these nodes provides a measure of the degree of match between the input and the four-dimensional vector which characterizes all the spatial characteristics of the node's receptive field. A good match implies that the node's receptive field shares similar position, orientation, and size with the input image. The most active of all nodes indicates the best estimate of these spatial parameters. If the nodes are spaced finely enough across position, orientation, and size, the Where information of the input may be accurately assessed by finding this maximal activity. This general strategy for defining a multiplexed Where map is quite elementary. Its interest lies in its simplicity and in the analysis that is required to make sure that it works in an efficient way.

Determination of the optimal node may be achieved via a winner-take-all competition between all nodes; that is, by a competition across position, orientation, and size. Perhaps the first winner-take-all, or WTA, network to be mathematically characterized is a competitive network whose cells undergo shunting, or divisive, inhibition and which communicate via faster-than-linear feedback signals [42]. A number of related WTA schemes have since been proposed; e.g., by Feldman and Ballard [43], Hadeler [44], Koch and Ullman [45], and Tsotsos *et al.* [46]. The one node which remains active after the competition carries the necessary Where information via its receptive field geometry. Figure location (x_I, y_I), orientation (ϕ_I), and size (s_I) are provided to the What channel to achieve invariance prior to object recognition.

Before such a parallel Where filter can function well, the filters given by Eqs. (1)–(4) must be modified so that they provide an unbiased measure of the size of the input figure. Consider, for example, a small elliptical input, oriented at $\phi = 120^\circ$, that fits snugly within the central region ($r^2 \leq 1$) of a filter of size $s = 8$ (Fig. 5b). The corresponding map cell will react strongly to this input, with $A = 250.82$. However, a nonoptimally sized cell receptive field of larger size $s = 16$ will respond even more strongly ($A = 384.11$), since the input ellipse lies in the region where $K \approx 1$ (Fig. 5). All other factors being equal, filter response increases with filter size. There are two mechanisms whereby this bias can be removed: normalization of the filter weights and unbiasing of filter output via spatial competition. The former approach is taken in this section, while the latter is developed in Section 5. Normalization of the filter weights and spatial competition can also be realized by the same shunting, or divisive, competition that is used for WTA selection, if the shunt is restricted to a single filter's receptive field weights. These two variants of the What-and-Where filter thus illustrate how different orderings of a single mechanism of shunting competition can be used to accomplish both functional tasks.

The scale bias arises as a straightforward consequence of the fact that larger filters have the same maximal (excitatory) value as smaller filters, but greatly increased excitatory receptive field area. In order to achieve unbiased scale estimation, the increase in excitatory receptive field area of large filters can be compensated for by reduced weights within this area. That is, filter weights are normalized by the area of the excitatory receptive field

$$N(s) = \int_{\Re} K(x, y, s, 0) dx dy, \quad (6)$$

where $\Re$ represents the excitatory region of the filter, at which $r < 1$. With a normalized kernel

$$K_N(x, y, s, \phi) = \frac{K(x, y, s, \phi)}{N(s)}, \quad (7)$$

the level of response of each filter to its optimal elliptical stimulus is equal to 1.0. The activity of the normalized filter output nodes is now given by

$$A(x, y, \phi, s) = \sum_p \sum_q K_N(p, q, \phi, s) I(p - x, q - y), \quad (8)$$

as in Eq. (5). The maximally activated node accurately codes the figural position, orientation, and size. A winner-take-all competition among all the nodes thus chooses the node whose spatial position in the Where map encodes figural position, orientation, and size. In summary, three competitive operations, acting in parallel across the network, are competent to generate a Where map: a cliff off-surround at each receptive field, normalization of each kernel by the integral of the excitatory on-center of each receptive field, and global winner-take-all competition across all receptive fields.

Figure 6 illustrates the result of convolving a small set of cliff-normalized filters and a vehicle silhouette image, depicted in the lower right-hand corner of the image. Each frame represents the output of a given filter convolution, with filter scale increasing from top to bottom and filter orientation changing in a counterclockwise fashion from left to right. The receptive field of the node with the maximal activity across position, orientation, and size corresponds to the Where information of the input image. In this case, the maximal activity occurs at the node marked with an "X" in the figure, and indeed represents the correct position, orientation, and size of the input figure, modulo the coarse filter

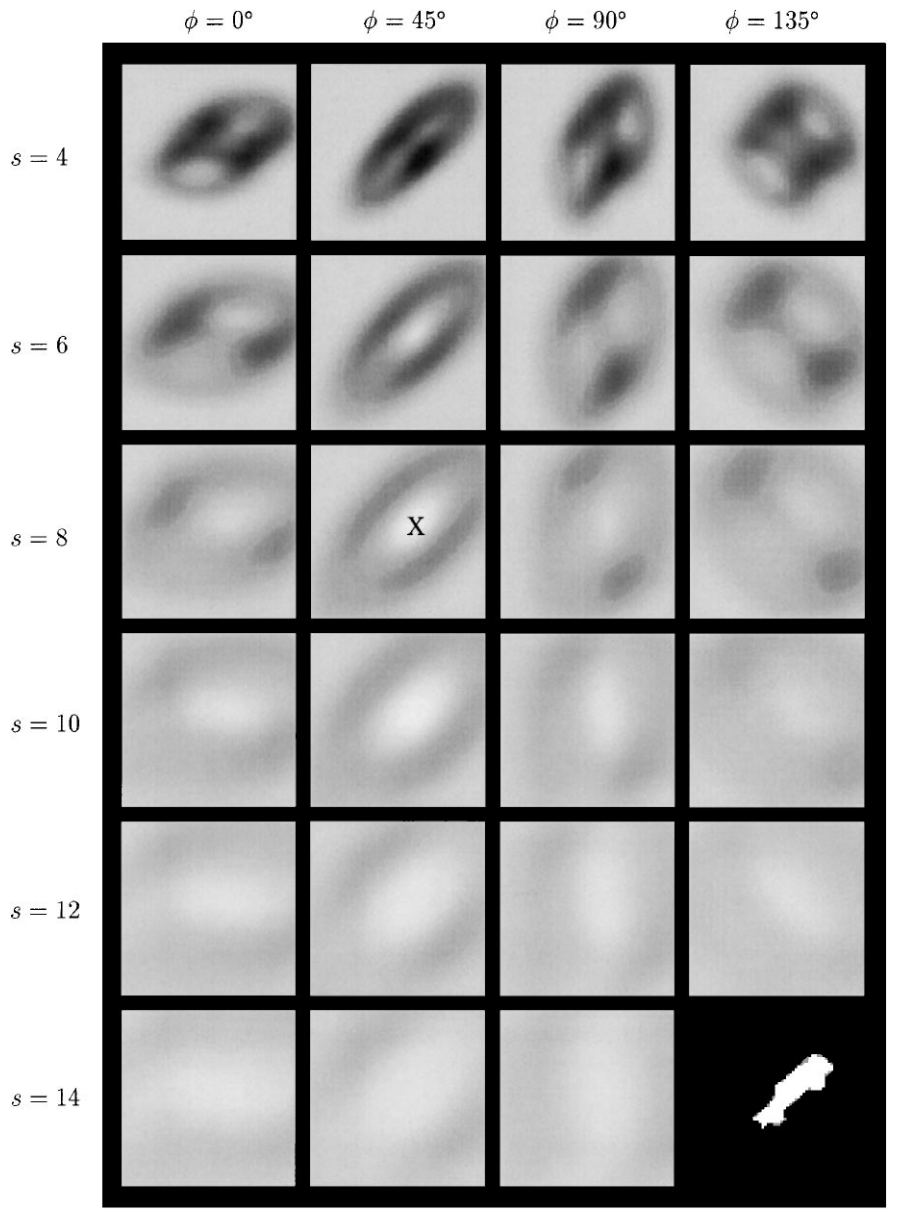

FIG. 6. A parallel Where filter that uses a cliff off-surround and a normalized on-center in each receptive field, followed by winner-take-all competition across receptive fields. Receptive field size increases from down columns ($s = 4, 6, 8, 10, 12, 14$) and orientation varies across rows ($\phi = 0, 45, 90, 135$). The output of the parallel Where filter to the figure given in the lower right-hand corner is given as a four-dimensional array of data. Each 2D subframe image represents the output of the convolution of a single normalized filter. The "X" marks the maximal activity across position, orientation, and size and correctly provides the Where information associated with the input image.

spacing. Tests employing finer filter spacing gave excellent qualitative results, but it soon became clear that a one-pass Where filter with fine filter spacing across orientation is needlessly expensive from a computational point of view.

4. COARSE CODING AND INTERPOLATION OF RECEPTIVE FIELD ORIENTATIONS

In order to achieve invariant image representation in the What channel, both orientation and size determination in the Where channel must meet high accuracy criteria. In a parallel system, high accuracy can be achieved by fine spacing of filters across receptive field position, orientation, and size, but at the cost of maintaining a multitude of filters. Orientation accuracy to within 1° would require 90 different filter orientations, with a spacing of 2° between filters at each size and at each position. Size accuracy to within 2% across scales from $s = 4$ to $s = 32$ (object length of from about $2as = 16$ to 128 pixels) would require 29 different scales. Thus 2610 (90×29) different filters at each node of the spatial grid would be required for accurate scale and

orientation determination. This large number of filters could be computed in real-time in a neural tissue or a parallel computer chip, but might prove cumbersome in applications that depend upon a serial computer. Orientational tuning in the primate striate visual cortex is typically much coarser, yet orientation changes can be detected with high accuracy [47]. We now show how coarse orientational tuning followed by an orientation interpolation mechanism can achieve fine orientational discrimination while reducing the computational load by more than an order of magnitude.

Within each set of filters of a given size s at a given position (x, y), interpolation across a sparse set of six orientations that are calibrated in degrees ($\theta = 0, 30, \ldots, 150$) can reduce the total number of filters from 2610 to just 174 (6×29) while maintaining orientation accuracy. Convolving the coarse filter activities $A(x, y, \theta, s)$ with a one-dimensional Gaussian kernel across orientation only, and resampling across a finer set of orientations ϕ, accomplishes the interpolation. The interpolated distribution of activity $A_G(x, y, \phi, s)$ more accurately reflects the actual orientation of the image than do the coarse filter activities $A(x, y, \theta, s)$.

This interpolation is realized as follows. For each orientation θ, let $A(x, y, \theta, s)$ be the output of the coarse filter bank, as in Eq. (8). Then for any $\phi \in [0, 180)$, the interpolated activity $A_G(x, y, \phi, s)$ obeys the equation

$$A_G(x, y, \phi, s) = \sum_{\theta} A(x, y, \theta, s) G(\theta - \phi), \tag{9}$$

where $G(\psi)$ is the Gaussian kernel

$$G(\psi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\psi^2/2\sigma^2}. \tag{10}$$

The standard deviation (σ) of the Gaussian is taken to be a fixed fraction of the coarse filter spacing being employed. Using this self-similarity constraint, the wider the spacing, the greater the standard deviation is chosen to ensure smooth interpolation. In practice, setting σ equal to 0.7 times the coarse filter spacing works well. Following interpolation, global winner-take-all competition between all activities $A_G(x, y, \phi, s)$ yields the position, size, and orientation of the figure, as described in Section 5.

Interpolation could also be implemented across position and size, leading to a four-dimensional Gaussian interpolation kernel and appropriate convolution. In practice, however, even with size interpolation, accuracy decreases rapidly as filter scale spacing increases. When combined with the complexities arising in implementation, interpolation across domains other than orientation was not deemed cost effective in this parallel Where filter.

5. ONE-PASS WHAT-AND-WHERE FILTER ALGORITHM

The one-pass parallel What-and-Where filter outlined in the last three sections will now be summarized mathematically. The input image $I = I(x, y)$ is convolved with each of the filters within the oriented filter bank. Interpolation across orientation provides a more accurate estimation of orientation. Figure position, orientation, and size are determined simultaneously via a winner-take-all competition between all output nodes, and the input image is centered, oriented, and scaled to form the invariant image I_{COS}. The filter operations are defined as follows.

Oriented cliff filter For orientation $\phi \in [0, 180)$ and size $s = 4, \ldots, 32$ pixels, the unnormalized oriented filter with orientation ϕ, size s, and elongation a is defined in terms of the cliff kernel K:

$$K(x, y, \phi, s) = (1 - r^6) \exp\left(-\frac{r^4}{1 + r^2}\right), \tag{11}$$

where

$$r^2 = \left(\frac{x'}{as}\right)^2 + \left(\frac{y'}{s}\right)^2, \tag{12}$$

$$x' = x \cos\phi + y \sin\phi, \tag{13}$$

$$y' = y \cos\phi - x \sin\phi. \tag{14}$$

The normalized filter is then given by

$$K_N(x, y, \phi, s) = \frac{K(x, y, \phi, s)}{N(s)}, \tag{15}$$

where

$$N(s) = \int_{\Re} K(x, y, 0, s) dx dy \tag{16}$$

and $\Re$ represents the excitatory region of the filter.

A coarse set of filter orientations and a fine set of filter sizes that respectively span the orientation and scale range are selected. In simulations, 174 filters, with 6 orientations $\theta = 0, 30, \ldots, 150$ (degrees), and 29 sizes $s = 4, 5, \ldots, 32$ (pixels) were used.

Filter Each normalized filter is convolved with the input image:

$$A(x, y, \theta, s) = \sum_{p} \sum_{q} K_N(p, q, \theta, s) I(p - x, q - y). \tag{17}$$

Interpolation Gaussian interpolation provides accurate orientation estimation, using

$$A_G(x, y, \phi, s) = \sum_{\theta} A(x, y, \theta, s) G(\theta - \phi), \tag{18}$$

where $G(\psi)$ is the Gaussian kernel

$$G(\psi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\psi^2/2\sigma^2}. \tag{19}$$

Parameter σ was set equal to 0.7 times the angle between successive orientations θ. In simulations, $\sigma = 0.7 \times 30 = 21$ degrees. Interpolated values of ϕ were computed in steps of 0.1 degrees.

Winner-take-all competition Global competition between all nodes identifies the position $(x_{\max}, y_{\max})$, orientation $\phi_{\max}$, and size $s_{\max}$ that maximize $A(x, y, \phi, s)$. Thus,

$$A(x_{\max}, y_{\max}, \phi_{\max}, s_{\max}) \geq A(x, y, \phi, s). \tag{20}$$

These values provide all the Where information; that is, position $(x_I, y_I) = (x_{\max}, y_{\max})$, orientation $\phi_I = \phi_{\max}$, and scale $s_I = s_{\max}$.

Center, orient, and scale invariant figure The figure I was translated by $(-x_I, -y_I)$, rotated through an angle of $-\phi_I$, and magnified by a factor of $24/s_I$ to obtain the invariant figure I_{COS}.

6. WHAT-AND-WHERE VEHICLE SIMULATIONS

The one-pass What-and-Where filter was tested quantitatively on vehicle input images. The four prototype vehicle inputs are shown in Fig. 7 in general position, with orientation 0° and scale 24 pixels. Although the What-and-Where system indicates that each of these images in horizontal ($\phi_I = 0$) the chunkier vehicles appear slightly tilted due to asymmetries about both the horizontal and vertical axes. The goal of the orientational measure is not to determine a veridical horizontal, but rather to generate a stable measure of canonical orientation. The actual orientation chosen will depend upon the shape of the figure, notably its anisotropy. The image squares were 128 × 128 pixels, and the prototype vehicle images ranged from 80 to 100 pixels in length and from 30 to 60 pixels in height. To create the test input set, prototype images were randomly rotated through angles of 0 to 180°, magnified by random factors ranging from 0.2 to 1.2 and placed at random positions in a 30 × 30 pixel area at the center of the image square. Each of the four prototype images generated 250 such random representations. System performance was judged by the accuracy of orientation and size determination.

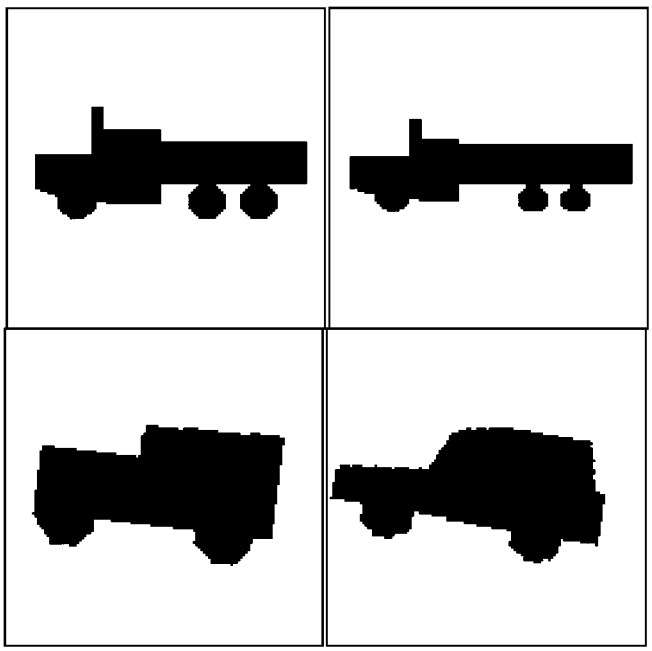

FIG. 7. Prototype vehicle images, with orientation $\phi_I = 0°$ and size $s_I = 24$ pixels.

Employing just 174 filters in the Where channel, as described in Section 5, the system easily met the goal of recovering orientation to within 1° and scale to within 2%. The mean orientation error was 0.43° and the mean scale error was 1.97%. Object localization was likewise very good. The mean error across x positions or y positions, taken independently, was less than 1 pixel. The mean Euclidean error across all (x, y) pairs was 1.12 pixels. The subsampling distortion caused by reduction of scale was the rate-limiting factor on system accuracy, as is described in Section 12. What-and-Where filter simulations are illustrated in Fig. 8. Column (a) shows the input image, column (b) shows the translated image I_C, column (c) shows the translated and horizontally oriented image I_{CO}, and column (d) shows the translated, oriented, and scaled image I_{COS}. I_{COS} is the What channel output figure (see Fig. 1). An XOR in column (e) between the output image I_{COS} and the prototype image indicates where errors occur.

7. NORMALIZED VERSUS UNNORMALIZED FILTERS

Filter normalization provides unbiased estimates of image scale, but relies upon accurate calculation of filter coefficients based on information about the entire excitatory region of the filter. Any inaccuracies or "drift" in filter coefficients could result in scale biases which could, in turn, lead to position and orientation biases as well. For this reason, an alternate, but related, model that provides a method of eliminating scale estimation bias is also presented. This method allows the use of unnormalized filters by first translating the image so that its center-of-mass lies at the origin. Together, these models provide a broader insight into the variety of operations whereby What-and-Where filters may be designed.

When convolving an image with unnormalized filters, it is possible for the maximal value across all filters to occur at a position that is distant from the actual position of the figure. If the filter scale is much smaller than the figure, then the response at the position of the figure will be weak due to the strong inhibitory troughs around the center of the figure. The response at the periphery of the figure position, where inhibition is weak, will be relatively strong. This can be seen in the relative responses of the smallest normalized filters of Fig. 6 (top row), where the maximal values lie outside the area occupied by the input figure. The same relative activations within a filter occur

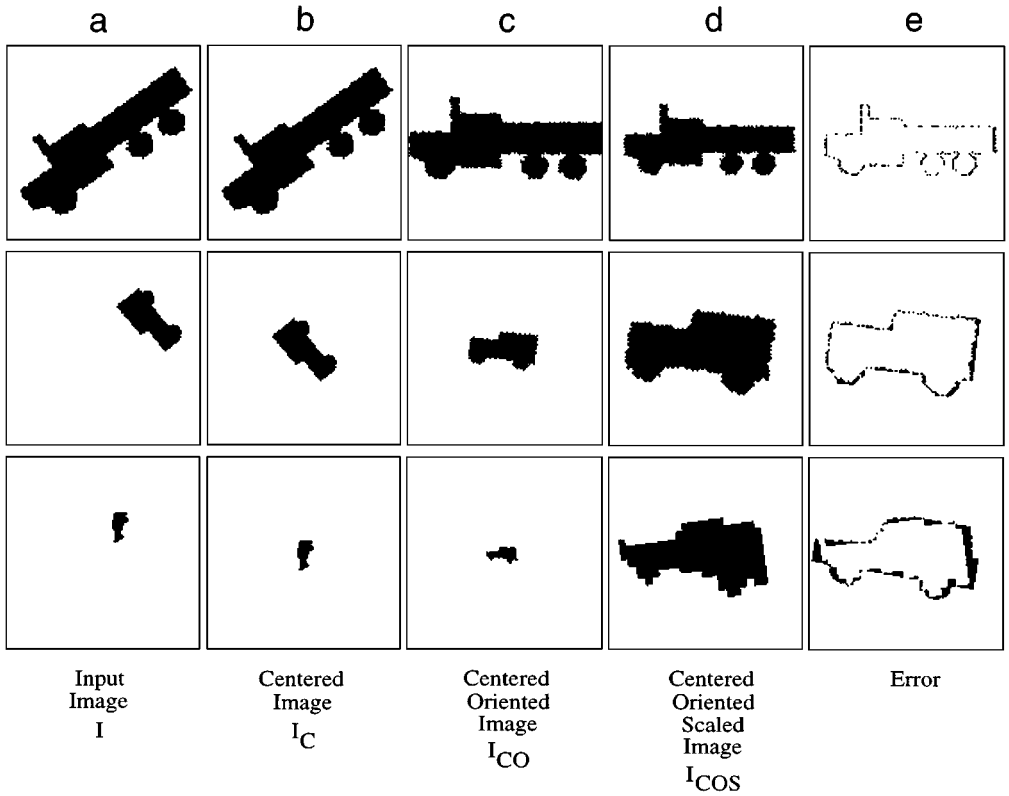

FIG. 8. Three examples of What-and-Where filter simulation results: (a) The input figure (I). (b) The image translated to its neutral position (I_C). (c) The image translated and oriented to the horizontal (I_{CO}). (d) The figure translated, oriented, and scaled to $s = 24$ pixels (I_{COS}). The last column indicates the degree of match (XOR) between the output of the What channel and the corresponding prototype figure.

whether or not it is normalized. Normalization ensures that positionally mismatched activations do not win the interfilter competition. Due to this problem, the next What-and-Where filter first determines figure position, using a method that is robust in noise, before filtering to determine orientation and size. After the target figure has already been separated from noise and background clutter, say by a CORT-X filter [15] or by an FBF network [16], a center-of-mass computation provides convenient and sufficiently accurate localization. The center of mass (x_I, y_I) may be computed by

$$T = \sum_x \sum_y I(x, y), \tag{21}$$

$$x_I = \frac{1}{T} \sum_x \sum_y x I(x, y), \tag{22}$$

$$y_I = \frac{1}{T} \sum_x \sum_y y I(x, y). \tag{23}$$

The figure is translated so that the new, centered, image has its center of mass at the origin:

$$I_C(x, y) = I(x - x_I, y - y_I). \tag{24}$$

When a clean image cannot be assumed, a noise-tolerant method, such as the diffusion-enhancement bilayer of Seibert and Waxman [48], can be used for target localization.

The unnormalized receptive fields of Eqs. (1)–(4) can be employed in an algorithmic manner to robustly determine figure position. Since localization is a problem only with small scale filters, large scale filters can be used to determine a coarse estimate of figure position. That is, the position of maximal activity within large scale filters yields figure position, but not with a high degree of accuracy. The positional estimate can be refined by finding the maximal activity at progressively smaller scales. This estimate slowly varies, becoming more accurate as filters have greater positional sensitivity due to their snugger fit around the input, until the scale becomes significantly smaller than the input figure, at which point the maximal activity occurs at some distant point. The correct figure position is the last positional estimate before this discontinuous jump.

8. AN ALTERNATIVE PARALLEL WHAT-AND-WHERE FILTER: THE DOUBLE PEAK PROBLEM

The centered figure I_C is next presented to a bank of filters, centered at the origin, that span all orientations ϕ and sizes s.

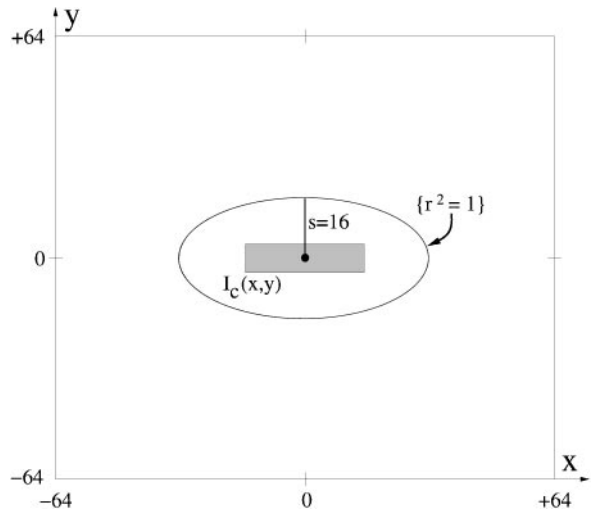

FIG. 9. A rectangular input image is defined by setting $I_C(x, y)$ equal to 1 inside the rectangle, and 0 elsewhere. A node centered at (0, 0) with receptive field orientation $\phi = 0^\circ$ and size $s = 16$ pixels has activity $A(\phi, s)$ equal to the sum of kernel values K over all points (x, y) in the rectangle. Figures that fall in the region where $r^2 > 1$ tend to generate large negative responses.

Convolution of the input vector with each filter bank at each spatial position (x, y) is no longer required, as the figure is already centered. Now, a point by point multiplication between the input image and each filter suffices; namely,

$$A(\phi, s) = \sum_x \sum_y K(x, y, \phi, s) I_C(x, y), \qquad (25)$$

as depicted in Fig. 9. If normalized filters are employed, then finding the maximal activity across orientation and size would determine ϕ_I and s_I. When employing the unnormalized filters of equations (1)–(4), it is not sufficient simply to select the filter that gives the maximal response across all $A(\phi, s)$, since outputs at all filter sizes larger than the figural size will tend to be greater than the activity at the correct filter size.

This problem can be solved by utilizing the *pattern* of activity across orientations at each given filter size. If the size s of the filter is too large, then the image tends to remain within the excitatory central region at all filter orientations, leading to a flat distribution of high activity. If the filter size s is too small, a flat distribution of low activity is observed, since then both the excitatory and the inhibitory regions of the filter intersect the image at all orientations. Only near the optimal size will activation levels vary rapidly with orientation.

These observations suggest a two-stage competitive mechanism that determines the optimal orientation and size of the input. At the first stage, nodes at each fixed size compete among orientations. This first competitive stage, which contrast-enhances responses at each size, emphasizes variations in activity, as in the BCS boundary segmentation network of Grossberg and Mingolla [11, 12]. Activities $A(\phi, s)$ that are flat across several orientations are inhibited, as when filter sizes are too large or too small. Peaks of activity in near-optimal scales are emphasized (Fig. 10a). After first-stage competition, the new activity profile $B(\phi, s)$ at each size measures how well both filter orientation and size fit the figure. A second competitive stage acts between all output nodes of the first competitive stage to select the node with maximum activity B (Fig. 10b). The image orientation ϕ_I and size s_I are estimated by the orientation and size of the selected node.

Figure 11 shows the simulation output of the first competitive stage (Fig. 10a). Function $B(\theta(s), s)$ is plotted to graph the output at the maximally activated orientation as a function of scale. Input $I(x, y)$ represents a car that is 112 pixels long and 38 pixels high, oriented in a horizontal direction and centered at the origin. Peak activity (Fig. 11a) occurs at scale $s = 25$. The second competitive stage thus estimates the height to be about $2s = 50$ pixels and the width to be about $2as = 100$ pixels. The filter corresponding to this peak has $\phi = 0$, the correct image orientation. A second peak in the maximal activity profile occurs in small-scale filters with orientation $\phi_I^\perp$ perpendicular to the correct orientation ϕ_I. In fact, the scale of the lower peak is approximately equal to the scale of the first peak divided by the filter elongation parameter $a = 2$ (Fig. 5a). The lower peak occurs where the major axis of a small-scale filter senses the height, rather than the width, of the input image (Fig. 11b). Since the peak occurring at the larger scale is always the "correct" one, this double peak effect does not cause errors after competition acts at the second competitive stage of the parallel network. As seen below, however, double peak uncertainty leads to a design constraint on a more efficient serial system.

9. PARALLEL VERSUS SERIAL FILTERING STAGES

As discussed in Section 4, a fully parallel system without interpolation requires a large number of filters to estimate figural orientation and size accurately. While interpolation similar to that of Section 5 could be employed with unnormalized filters to greatly reduce the number of these filters, a more fundamental alteration in Where channel structure may also be used to further reduce computational load in applications wherein algorithmic operations can be performed serially. In particular, the total number of filters is reduced by using two serial filtering stages, one to determine orientation and one to determine size. Such a serial system can determine orientation ϕ_I, using 90 oriented filters, and can then determine size s_I, using 29 additional scaled filters, rather than the 90×29 filters that would be needed by a parallel system with comparable performance. In the serial system, the image is first centered, say by (21)–(24). The orientation of the image is then determined via competition among the 90 oriented filters. This information is used to orient the image to a canonical (horizontal) direction, after which the second bank of 29 filters determine the image scale. Figure 12 depicts a Where filter that computes position, orientation, and size in serial stages.

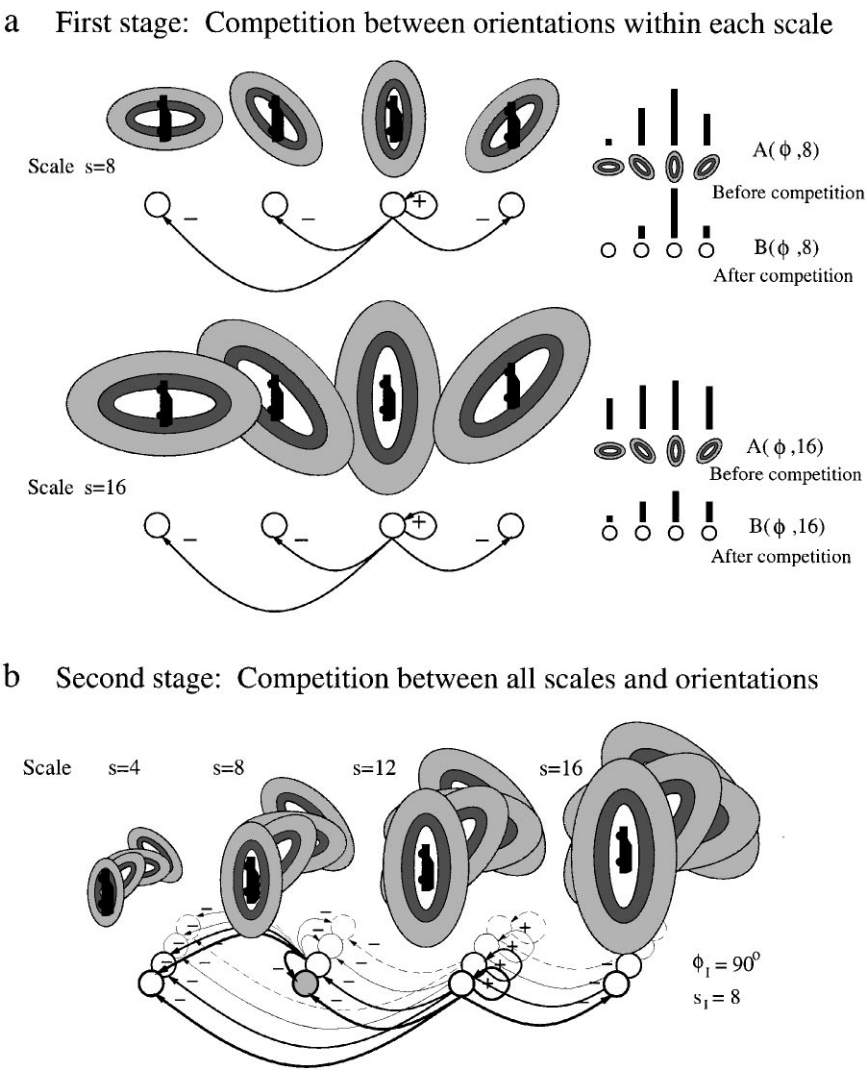

FIG. 10. Fully parallel filter. (a) At the first competitive stage, for each size scale s, competition contrast enhances node responses across orientations. Competition for network activity B enhances the peak response at the optimal scale ($s = 8$). (b) The second competitive stage, among all output neurons of the first stage, selects a filter whose orientation ϕ_I and scale s_I fit the input image.

Whether unnormalized or normalized filters are employed within the Where channel, a great computational savings can be realized by serializing orientation and size determination. The following discussion outlines some problems and solutions inherent in the general case of unnormalized filters. At the end of the discussion, the system simplifications resulting from the use of normalized filters will be presented.

The theoretical ideal of a complete serialization of Where filter modules with unnormalized filters, with a single size scale in the orientation filter bank and a single orientation in the size scale filter bank would not produce accurate results. The double peak in maximal first-stage output $B(\phi(s), s)$ across scales (Fig. 11a) implies that a range of sizes must be employed for accurate orientation determination. If only a single size were employed, the system would tend to choose the orthogonal orientation for images much larger than this size. For example, if the fixed size scale were $s = 10$, the system would incorrectly predict a vertical orientation for the large car shown in Fig. 11b. Employing a single large size scale solves this problem, but creates a new one of small-image inaccuracy. Namely, since small images of all orientations would fit in the central region of all the large filters, orientation would be indeterminate. A compromise can be achieved by using a sparse set of approximately eight scales which span the range of possible figure sizes. The corresponding oriented filters provide enough information to disambiguate the double peaks without sacrificing accuracy. In this hybrid serial-parallel configuration, the orientation determination module is essentially the same as in the parallel system described in Section 8, but with fewer scales. As in Fig. 10, competition between orientations within each size scale (first competitive stage) is followed by competition between orientations and sizes (second competitive stage). The second competitive stage hereby

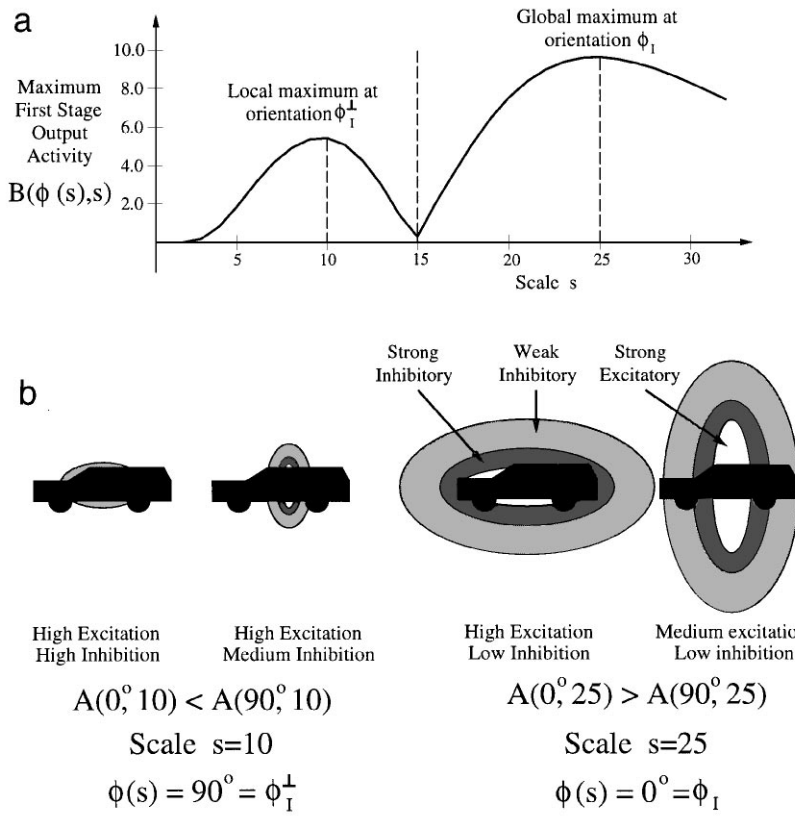

FIG. 11. Double peak effect. (a) Maximum output activity $B(\phi(s), s)$ of the first competitive stage for each size scale s. The global maximum across all scales occurs at $s = 25$, where the orientation $\phi(s) = 0^\circ$ gives the maximal response $B(\phi, s)$ across orientations. A second peak in the graph of $B(\phi(s), s)$ occurs at scale $s = 10$, where $\phi(s) = 90^\circ$. (b) At small scales competition between orientations tends to select the orientation that matches the height of the image, yielding the orientation $\phi_I^{\perp}$ that is perpendicular to ϕ_I. At larger scales the filter that best matches the full test image yields the correct orientation.

yields the orientation of the image (ϕ_I) and a rough estimate of its size. The former is applied to orient the centered image ($I_C \rightarrow I_{CO}$) prior to subsequent filtering for more accurate size determination. The coarsely determined scale *could* be used to narrow down the range of filter scales in the next stage. However, the likelihood that the computed scale is erroneous, due to the double peak effect (Fig. 11a), makes it prudent to ignore this information.

Instead, the horizontal image I_{CO} next becomes the input to the size determination module, where the Where filter again employs a reduced parallel system. Although the orientation I_{CO} is known, a bank of filters of many scales but only horizontal orientation is still inadequate. As in Fig. 10a, it is the variation in activity across orientations that indicates how well a particular filter scale matches that of the image, rather than the absolute size of a node's response. In practice, just two orientations (horizontal and vertical) provide enough information about activity variations to give an accurate estimate of scale. If a scale matches that of the image, the difference between the horizontal and vertical filter responses will be large at that scale (Fig. 10a). In contrast, if the filter scale is too large, both horizontal and vertical responses will be large; if the filter scale is too small, both horizontal and vertical responses will be small; and in either case the *difference* between the horizontal and vertical responses will be small. In summary, the hybrid serial–parallel system uses eight scales at the orientation determination stage and two orientations at the size determination stage. The total number of filters is hereby reduced to a total of 778 ($90 \times 8 + 2 \times 29$), compared to the 2610 (90×29) filters in the fully parallel system.

The use of normalized filters eliminates the need for competition between orientations at each scale (the first competitive stage). Either competition between orientations or filter normalization can be used to derive a goodness-of-fit measure that works across filter sizes. Thus orientation determination that uses normalized filters consists *solely* of winner-take-all competition between orientations and sizes (the second competitive stage). After reorienting of the input image ($I_C \rightarrow I_{CO}$), size determination in a normalized filter system progresses in the same manner as above, with the first competitive stage again being skipped. Note that, as above, the filter bank can be reduced to different sizes at a *single* orientation.

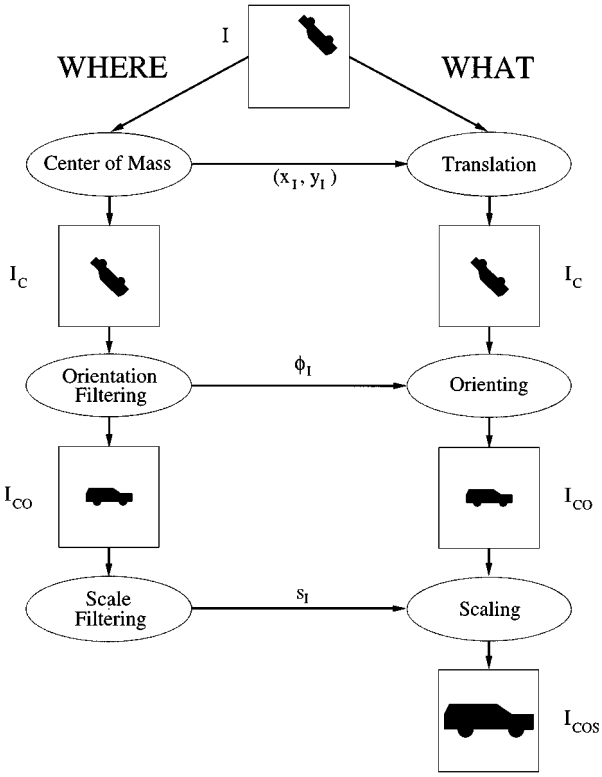

FIG. 12. The serial or serial-parallel What-and-Where system with dissociated orientation and scale filtering.

10. ORIENTATION INTERPOLATION IN A SERIAL ALGORITHM

Interpolation within the orientation determination module can further decrease the number of filters required within the Where channel, just as it did for the parallel system in Section 4. Interpolation across a sparse set of six orientations ($\theta = 0, 30, \ldots, 150$) can reduce the total number of filters to just 106 ($6 \times 8 + 2 \times 29$), rather than the 2610 filters of a fully parallel system or the 778 filters of the serial algorithm without interpolation, while maintaining orientation accuracy. Again, this process can be applied to the system regardless of the type of filter (normalized or unnormalized) being employed. Initial discussion below of how this is done assumes unnormalized filters.

In the orientation determination module, for each orientation θ, let $A(\theta, s)$ be the output of the coarse filter bank, as in (25). Then for any $\phi \epsilon [0, 180)$, the interpolated activity $A_G(\phi, s)$ is given by

$$A_G(\phi, s) = \sum_{\theta} A(\theta, s) G(\theta - \phi), \tag{26}$$

where $G(\psi)$ is the Gaussian kernel given by Eq. (10). As before, a self-similar selection or σ is made to be a fixed fraction of the filter spacing. Following interpolation, the activities $A_G(\phi, s)$ compete across scales and orientations, as in the second competitive stage of Fig. 10.

A similar interpolation can also be performed in the size determination module. Because interscale size comparisons are not made until after the first competitive stage (Fig. 10), interpolation across scales is performed on first competitive stage output B. For a sparse set of size scales t, $B(0, t)$ is the first competitive stage output for the canonical orientation $\phi = 0$. Then for any scale s, the interpolated activity $B_G(0, s)$ is given by

$$B_G(0, s) = \sum_{t} B(0, t) G(t - s), \tag{27}$$

as in Eq. (26). Following interpolation, global competition among all outputs $B_G(0, s)$ yields the image scale s_I. In summary, interpolation in the orientation stage typically permits a coarse filter spacing of up to 30° (rather than 2°), while maintaining a mean orientation error of less than 1%. Size interpolation proved less valuable, still requiring scale (t) filter spacing of one pixel to maintain accuracy.

Employing normalized filters removes the first competitive stage. Interpolation within both orientation and size determination modules occurs before the second competitive stage, but is otherwise the same as above.

11. SERIAL WHAT-AND-WHERE FILTER ALGORITHM

The serial What-and-Where filter implementation algorithm will now be summarized mathematically using unnormalized filters. The algorithm using normalized filters will then be given. The algorithm first computes the position (x_I, y_I) of an input figure $I = I(x, y)$. After translation, the new figure I_C is centered at the origin. A bank of oriented filters then determines the image orientation to be ϕ_I degrees. After rotation, the centered and oriented image I_{CO} is horizontal. Finally, a second bank of filters determines the image scale to be s_I pixels.

Step 1: Determining Position (x_I, y_I)

For pixel values $x, y = -64, \ldots, +64$, a gray-scale figure I is described by input $I(x, y) \in [0, 1]$. In a noise-free setting, figure position (x_I, y_I) corresponds to the center of mass:

$$T = \sum_{x} \sum_{y} I(x, y), \tag{28}$$

$$x_I = \frac{1}{T} \sum_{x} \sum_{y} x I(x, y), \tag{29}$$

$$y_I = \frac{1}{T} \sum_{x} \sum_{y} y I(x, y). \tag{30}$$

Centered image I_C

$$I_C(x, y) = I(x - x_I, y - y_I). \tag{31}$$

Oriented receptive fields For orientation $\phi \in [0, 180)$ and scale $s = 4, \ldots, 32$ pixels, the unnormalized oriented receptive field with size s, orientation ϕ, and elongation a is defined in terms of the kernel K:

$$K(x, y, \phi, s) = (1 - r^6) \exp\left(-\frac{r^4}{1 + r^2}\right), \tag{32}$$

where

$$r^2 = \left(\frac{x'}{as}\right)^2 + \left(\frac{y'}{s}\right)^2, \tag{33}$$

$$x' = x \cos\phi + y \sin\phi, \tag{34}$$

$$y' = y \cos\phi - x \sin\phi. \tag{35}$$

Step 2: Determining Orientation ϕ_I

First stage (Competition across orientations) A coarse set of scales that span the range of input scale values is selected. In simulations, eight scales, $s = 4, 8, \ldots, 32$, were used. Similarly a coarse set of orientations θ that span the range [0, 180) is selected. In simulations, 10 orientations, $\theta = 0, 18, \ldots, 162^\circ$, were used.

Filter.

$$A(\theta, s) = \sum_x \sum_y K(x, y, \theta, s) I_C(x, y). \tag{36}$$

Interpolation.

$$A_G(\phi, s) = \sum_\theta A(\theta, s) G(\theta - \phi), \tag{37}$$

where

$$G(\psi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\psi^2/2\sigma^2}. \tag{38}$$

Parameter σ was set equal to 0.7 times the angle between successive orientations θ. Thus, in simulations, $\sigma = 0.7 \times 18 = 12.6^\circ$. Interpolated values of ϕ were computed in steps of 0.1°.

Competition. Normalization is achieved using competition across orthogonal orientations, as in

$$B(\phi, s) = A_G(\phi, s) - A_G(\phi^\perp, s), \tag{39}$$

where $\phi^\perp$ is the orientation perpendicular to ϕ.

Second stage (Competition across scales and orientations) Competition across size scales and orientations identifies the orientation $\phi_{\max}$ and size $s_{\max}$ that maximize $B(\phi, s)$. That is,

Maximum B.

$$B(\phi_{\max}, s_{\max}) \geq B(\phi, s) \tag{40}$$

across all interpolated orientations $\phi \in [0, 180)$ and scales $s = 4, 8, \ldots, 32$.

Optimal orientation. Let the optimal orientation be

$$\phi_I = \phi_{\max}. \tag{41}$$

Center and orient figure I_{CO}. Figure I_C was rotated through an angle of $-\phi_I$ degrees to obtain the centered and oriented figure I_{CO}.

Step 3: Determining Size Scale s_I

First stage (Competition across orientations) A set of size scales was again selected. In simulations, 29 scales $t = 4, 5, \ldots, 32$ pixels were used.

Filter. For orientations $\phi = 0 = \phi_I$ and $\phi = 90 = \phi_I^\perp$, first-stage output was computed as

$$A(\phi, t) = \sum_x \sum_y K(x, y, \phi, t) I_{CO}(x, y), \tag{42}$$

Competition. Normalization is achieved using competition across orthogonal orientations, as in

$$B(0, t) = A(0, t) - A(90, t). \tag{43}$$

Interpolation.

$$B_G(0, s) = \sum_t B(0, t) G(t - s). \tag{44}$$

The Gaussian G was defined as in (38) and the standard deviation σ was set equal to 0.7 times the number of pixels between successive scales t. In simulations $\sigma = 0.7(1) = 0.7$ pixels, and interpolated values of s were computed in steps of 0.1 pixels.

Second stage (Competition across scales) Competition identifies the scale $s_{\max}$ that maximizes $B_G(0, s)$. That is,

Maximum B_G.

$$B_G(0, s_{\max}) \geq B_G(0, s) \tag{45}$$

for all interpolated $s \in [4, 32]$.

Optimal size. Denote the optimal size scale by

$$s_I = s_{\max}. \tag{46}$$

Center, orient, and scale figure I_{COS}. The figure I_{CO} was magnified by a factor of $24/s_I$ to obtain I_{COS}.

Using normalized filters, the algorithm is modified as follows. Replace K in (32) by

$$K_N(x, y, \phi, s) = \frac{K(x, y, \phi, s)}{N(s)}, \tag{47}$$

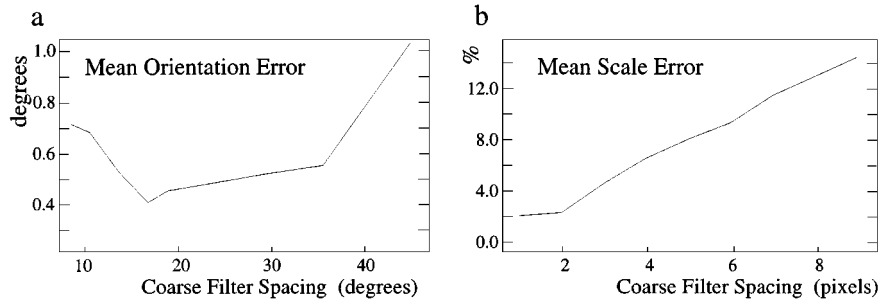

FIG. 13. Coarse filter spacing tests with elliptical input images: (a) Mean size of orientation error as a function of orientation spacing in the first filter bank, before interpolation. (b) Mean scale error $|s_I - s_{\text{actual}}|s_I^{-1}$ as a function of scale spacing in the second filter bank.

where

$$N(s) = \int_{\Re} K(x, y, \phi, s)dxdy. \tag{48}$$

The competitive interaction B across orientations in (39) is unnecessary. It is replaced by

$$B(\phi, s) = A_G(\phi, s). \tag{49}$$

The competitive interaction B in (43) is likewise replaced by

$$B(0, t) = A(0, t). \tag{50}$$

Note that, when determining size, only a single orientation $\phi = 0 = \phi_I$ is needed.

12. PARAMETER DETERMINATION

Serial What-and-Where filter parameters were selected through studies of system response to a simple elliptical input image. In this way, the number of orientations and the number of scales for each of the two Where filter banks (Fig. 12) were chosen. Preliminary testing had fixed the value of the standard deviation σ of the Gaussian interpolation kernel in (44) at 0.7 times the distance between coarse orientation or scale values (Section 4). The effects of figure elongation and scale were also examined. Parameters were determined using unnormalized filters, but carried over well to tests employing normalized filters.

A prototype elliptical test image, within a 128 × 128 pixel square, was defined by the inequality:

$$1 \geq \left(\frac{x}{a^* \times 24}\right)^2 + \left(\frac{y}{24}\right)^2. \tag{51}$$

With $a^* = 2$, this ellipse fits exactly within the central excitatory region of a filter of size $s = 24$, orientation $\phi = 0$, and elongation $a = 2$, as defined in Eqs. (32)–(35).

Orientation (θ) filter spacing of 18° (10 filters) was found to achieve an optimal balance between system accuracy and number of filters, although good performance was maintained at least to spacings of 30° (6 filters). For the simulations, 500 elliptical images were randomly generated from the prototype ellipse (51). Orientations ranged from 0 to 180°, while magnifications ranged from 0.2 to 1.2. Thus the scale factor, which was 24 pixels in the prototype, ranged from 4.8 to 28.8 pixels. These inputs were presented to different What-and-Where filters, each with a different orientation filter spacing, resulting in the mean orientation error plot of Fig. 13a. Eight scales were employed at each orientation to prevent errors due to the double peak problem (Fig. 11a). In Fig. 13a, the best performance occurred at a filter spacing of 18°. Both the exact spacing at which this optimal performance occurs and the minimum error level can be shifted by altering the interpolation σ. However, an orientation spacing of 18°, with $\sigma = 0.7(18) = 12.6$ degrees, provided excellent results.

Within the size determination module, horizontal test images with magnifications from 0.2 to 1.2 were used to determine mean scale error as a function of the coarse scale (t) filter spacing. Accuracy steadily decreases with increased filter scale spacing (Fig. 13b). This illustrates that scale interpolation does not compensate for missing scales in the way orientation interpolation compensates for missing orientations.

The use of a fixed eccentricity ($a = 2$) in Eq. (33) for all What-and-Where filter elements raises the question of how well the system would perform with inputs that do not fit well within any central excitatory region. This question was examined by varying the elongation a^* in (51) of the prototype elliptical input, with results depicted in Fig. 14. The "optimal" coarse filter spacings of 18° and 1 pixel of scale were employed for a total of 138 ($10 \times 8 + 2 \times 29$) filters. As the image approaches circularity ($a^* \to 1$) the accuracy decreases, as expected from the reduced degree of orientation information in the input itself. Accuracy increases monotonically with increasing elongation, even though very elongated elliptical inputs do not fit any of the filters well.

Although the error rates for orientation and size determination do not depend upon the initial orientation of the image, both orientation and size accuracy deteriorate for small inputs. This is due to subsampling effects in the input and the filters, a problem inherent in invariant preprocessing of digital images. This was demonstrated in a What-and-Where filter simulation test of 5000 elliptical inputs, with orientation and size ranges and optimal

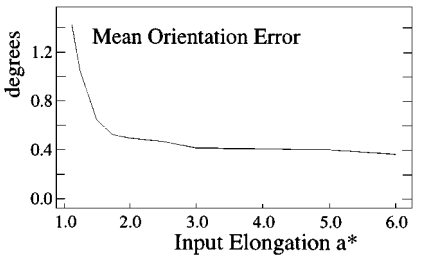

FIG. 14. Mean orientation error as a function of elliptical input elongation a^*.

filter spacing as above. Figure 15 shows that orientation and size errors become significant when the input image is small.

13. SERIAL WHAT-AND-WHERE VEHICLE SIMULATIONS

The What-and-Where system was tested extensively on the vehicle input images (Fig. 7), using parameters derived from elliptical input image studies (Section 12) and using both unnormalized and normalized filters. For the parallel system, prototype images were randomly rotated through angles of 0 to 180°, magnified by random factors ranging from 0.2 to 1.2 and placed at random positions in the square to create a test set. Each of the four prototype images generated 1000 such random representations.

For both unnormalized and normalized filters or unnormalized filters, In all cases, the network recovered orientation to within 1° and size to within 2%. The mean orientation error was 0.42° and the mean size error was 1.8%. The subsampling distortion caused by reduction of size was the limiting factor on system accuracy, as in Fig. 15. Increasing the orientation filter spacing from 18 to 30° increases the orientation error mean to 1.00° while maintaining the same size error mean of 1.8%. The equivalent test was performed with normalized filters, yielding a mean orientation error of 0.51° and a magnitude error of 1.97%. These systems, with only 106 ($6 \times 8 + 2 \times 29$) unnormalized filters or 77 ($6 \times 8 + 29$) normalized filters, meet the original performance criteria with far fewer than the 2610 filters required by the parallel system of Section 3, or the 274 filters required by the parallel system with interpolation of Section 5.

14. ALTERNATIVE WHAT-AND-WHERE MODELS

The present model differs in several notable ways from alternative approaches to the What-and-Where problem. An early model, that of Koch and Ullman [45], includes: (1) an early parallel representation of several stimuli and their featural characteristics; (2) a mapping from these representations into a non-topographic representation which contains properties of only one stimulus at a time; (3) a winner-take-all, or WTA, network that implements stimulus selection based on salience of each location; (4) inhibition of the selected location that causes a shift to the next most conspicuous location. Properties (1), (3), and (4) were introduced in Grossberg [9, 49] as part of a biological model of working memory, wherein multiple items are simultaneously stored in a spatial map, as in property (1). In this model, items are rehearsed, as in property (2), from the most to the least active, and use a self-inhibiting feedback, as in property (4), to prevent perseverative performance of the most active item. A Where map is a type of attentive working memory whose activities happen to code object properties (e.g., size and orientation) at prescribed spatial locations. A neural model of how such a Where map may be used to control sequences of saccadic eye movements was described in Grossberg and Kuperstein [50]. This model clarifies how Where properties can give rise to actions, or How properties, as proposed by Goodale and Milner [5].

WTA circuits are ubiquitous in models of this type. A rigorously characterized WTA neural network based upon competitive feedback between nodes or cells was described in Grossberg [42]. Hadeler [44] proposed a related network. An iterative formulation of a competitive WTA was provided by Feldman and Ballard [43]. By now, there are many variants of such circuits in use; e.g., by Cohen and Grossberg [51, 52], Coultrip, Granger, and Lynch [53], Ellias and Grossberg [54], Ermentrout [55], Grossberg [56], Grossberg and Levine [57], and Tsotsos *et al.* [46]. For purposes of biological modeling, the Grossberg [42] model and later elaborations thereof use cells that obey membrane, or shunting, equations and recurrent on-center

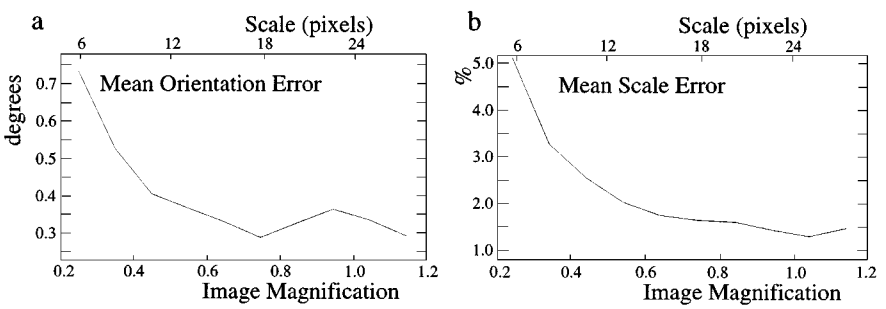

FIG. 15. What-and-Where filter output error as a function of input figure scale. (a) Mean size of orientation error. (b) Mean scale error. System performance deteriorates for small figures.

off-surround interactions. For image processing applications, any efficient algorithm will do.

The Koch and Ullman [45] algorithm differs from the present one in its processing stages and operations. The present model focuses upon *preattentive* mechanisms whereby all the figures in a scene can be transformed, in parallel, into invariant representations that are suitable for pattern recognition. Attentive mechanisms modulate these preattentive stages via nonlinear feedback. For example, Carpenter and Grossberg [58, 27, 59] have described how ART modules can autonomously learn object categories that fit their size, shape, and number to the statistics of a nonstationary environment. An ART module can sample each What representation for this purpose. In ART, activation of a top-down learned prototype primes its target cells so that they respond only when a target that matches the prototype well enough is presented. In this way, such a system can perform a fast parallel search for desired targets across a scene. In contrast, the Koch and Ullman [45] algorithm relies upon primitive object properties, such as object brightness, to select a target via a form of serial processing in which no high-level target priming is possible. The same limitations hold for all the models that are surveyed below.

Several alternative models comment upon how the brain may accomplish search tasks. Explaining human psychophysical data and animal neurobiological data is the ultimate test of such a proposal. The ART model has elsewhere been shown to have properties that qualitatively match neurophysiological recordings from cells in monkey inferotemporal cortex during behavioral tasks [59, 60]. These results concern how recognition categories of variable generality are learned, matched, and reset. In particular, the same ART top-down expectations that prime the system to respond selectively to desired targets also generate a matching rule whereby irrelevant target features are suppressed and primed target features are supported. Several investigators have reported neurophysiological evidence for such a matching rule in extrastriate and temporal cortex (e.g., [61, 62]). Reynolds *et al.* [63, 64] have performed experiments that support the simplest version of this ART matching rule as a substrate for spatial attention in extrastriate visual cortex, a version that seems also to occur in several other visual, auditory, and motor representations [65].

These models of What processing stream recognition are mentioned for three reasons. First, they illustrate the utility of a preattentive What-and-Where model of the type that we have described as a front end for fast parallel search of desired targets. Second, they support the biological relevance of the ART learning and categorization modules that are proposed to help carry out these tasks. Third, they illustrate a key difference with models that carry out serial search based on low-level features.

Grossberg, Mingolla, and Ross [35] and Grossberg [19] have combined What and Where properties in an algorithm that clarifies a large search database, including recent data concerning how humans carry out fast parallel search for complex 3D object properties, as in the work of Bravo and Blake [66], Cohen and Ivry [67], Enns and Resnick [68], He and Nakayama [69], Mordkoff, Yantis, and Egeth [70], Wolfe, Cave, and Franzel [71], and Wolfe and Friedman-Hill [72]. This SOS, or spatial object search, algorithm suggests how 3D boundary and surface representations of a scene interact reciprocally with learned object categories (What stream) and spatial maps (Where stream) to focus attention upon desired objects in a 3D scene and to search for targets amid distractors. Alternative What-and-Where models have typically ignored 3D boundary and surface properties and have not analysed how object categories can be autonomously learned in real time. Earlier work from our group proposed neural models of 3D boundary and surface representation (e.g., [19]) and object category learning (e.g., [58, 59]). The present model analyses some of the computational problems that the Where filter needs to solve when it is embedded into a larger visual recognition and search architecture like SOS.

Olshausen, Anderson, and van Essen [73] have proposed a shifter circuit model whereby an attended object can be transformed into a representation that is invariant under translation and size, but not rotation. As in the Koch and Ullman [45] model, targets are selected based on low-level features such as brightness or size, and these objects can be searched one at a time in a serial manner. The model assumes that each figure to be recognized at the lowest level is matched by an invariant representation of itself at the top level through a clever, but complex, multistage routing circuit. The model does not propose how this invariant exemplar is generated at the top level, and thus faces the challenge that it cannot self-organize its object recognition codes. Indeed, the authors admit that "it remains to be seen whether such a system can self-organize ... with experience." If this is so, then the model cannot operate in an unsupervised way because it has no way to generate the invariant representation on which the algorithm feeds.

The shifter circuit connections are derived by using a Liapunov, or energy, method of the type proposed by Cohen and Grossberg [51] and Hopfield [74] to link the lowest and top layers via selected pathways. It is not stated how the proposed energy function could be implemented by the brain. It is also unclear how such a mechanism, being a relaxation algorithm, could work in real time to recognize an object with the speed that is needed in realistic human or technological image processing applications. Thus the Olshausen *et al.* [73] model, despite the ingenuity of its bottom-up and top-down interactions, faces a serious challenge from the present approach, wherein slow relaxation algorithms are replaced by fast competitive and interpolation operations, and a theory is developed of how attentive recognition categories are self-organized.

Tsotos *et al.* [46] have elaborated a "selective tuning" model of visual attention. Their spatial selection (Where stream) is realized by inhibition of irrelevant connections within a visual pyramid. Their feature selection (What stream) is realized by inhibition of units that compute irrelevant features. A search process operates recursively using WTA operations that move from the globally winning unit in the top layer downwards. The

search process inhibits all the connections that do not contribute to the winner. After recursive processing, the cause of the maximal response at the top layer is isolated at the bottom layer. The algorithm operates through two traversals of the pyramid. The search is set up by a bottom-up sweep through the pyramid to select the global winner which, in turn, drives the top-down search.

This algorithm bypasses the problem of using a relaxation algorithm, but it substitutes a considerable machinery of interpretive units, gating units, bias units, and gating control units to do attentional selection. The Tsotsos *et al.* [46] algorithm represents a significant advance over the Koch and Ullman [45] algorithm. As in that algorithm, however, attention operates serially on only one target at a time. Although the authors note that the WTA can, in principle, be biased in favor of some features over others, they do not say how such priming, or more generally how category learning and recognition, can be accomplished in a self-organizing way.

Attentional selection within the Tsotsos model uses an array of gating and bias units that selectively inhibit unwanted *connections* throughout the pyramid. A large auxiliary network of highly specific connections is needed to inhibite the unwanted connections. An ART network, in contrast, achieves attentional selection by inhibiting the activities of unmatched *nodes*, not the much larger number of connections that feed these nodes. No auxiliary gating system is required within the ART model. Instead, top-down nonspecific inhibitory signals combine with excitatory top-down prototype signals at target nodes. These simple operations help to explain large behavioral and neural databases [58, 59, 65] and have been used in large-scale pattern recognition applications in technology, ranging from the design of the Boeing 777 and the control of nuclear reactors to medical database analysis, Landsat satellite image analysis, and the analysis of multispectral infrared, LADAR, and SAR imagery; see Carpenter and Grossberg [75] for some references.

The Tsotsos *et al.* [46] model incorporates a discussion of how saccades can move an eye or a camera to foveate a particular object in a scene. A full analysis of this issue would take us too far afield, but some comparative remarks may be helpful. The authors summarize psychophysical data, such as that of Remington and Pierce [76] and Kröse and Julesz [77] showing that attentional shifts can occur over variable distances in equal time, and that such rapid time-invariant attention shifts would be needed to control saccadic eye movements during reading. The authors note that this property is inconsistent with the Koch and Ullman [45] WTA algorithm and develop a new one to replace it.

We have elsewhere proposed an alternative solution. Grossberg and Rudd [78, 79] and Francis and Grossberg [80] have modeled the cortical dynamics of apparent motion by proposing how a wave of apparent motion can interpolate spatially separated and temporally staggered flashes of light. This wave can travel at a variable speed to join an earlier flash with one that occurs a fixed time later, even if the second flash is at variable distances from the first. This classical equal-time property of apparent motion [81] is simulated along with beta motion, gamma motion, delta motion, split motion, Ternus motion, and Korte's laws, among other data.

Such a traveling wave has been proved to occur in any system wherein the effects of each input are Gaussianly filtered across space, and the activity due to one input is waning while that due to the next input is waxing, within prescribed spatiotemporal bounds. Such a wave is therefore called a G-wave. The peak of the wave is chosen by a WTA operation.

Grossberg [82] suggested that G-waves carry spatial attention shifts via the magnocellular visual cortical processing stream that feeds the parietal cortex of the brain's Where system; see also [83]. By the equal-time property, a spatial attention shift due to a G-wave can occur over variable distances in equal time. G-waves are proposed to solve the ecologically ubiquitous problem of continuously tracking a prey or predator as it moves at variable speed between dense occluding cover. The intermittently occluded target produces a series of temporally discrete "flashes" that the G-wave continuously interpolates. Grossberg [19] outlined how this attentional tracking mechanism can be joined to SOS-type search mechanisms so that static 3D boundary and surface properties can compete with target motion properties to control attention shifts.

These motion mechanisms automatically realize an "attention capture" mechanism by enhancing transient responses to flashed events. The interaction of these transient enhancement effects with SOS mechanisms helps to explain how competition can occur between top-down priming and bottom-up energetic demands for attention. Tsotsos *et al.* [46] construct a special algorithm to enhance abrupt image events. Although this algorithm includes some of the properties of the Grossberg–Rudd motion model, it does not include the key operations that are needed to explain psychophysical data about motion percepts.

Finally, Tsotsos *et al.* [46] discuss a possible algorithm for causing saccadic eye movements to points of interest, but they do not analyse how the brain achieves its self-organizing control and calibration of eye movements. Such a theory was developed by Grossberg and Kuperstein [50].

Another connection of the present model is with work on steerable filters [84]. This work proves some nice theorems about the circumstances under which one can synthesize filters of arbitrary orientation from linear combinations of basis filters so as to adaptively steer a filter to any orientation. The goal of steerable filters is somewhat different from our use of Gaussian interpolation of oriented filters. Our goal herein is partly based on computational efficiency and partly on biological plausibility. Freeman and Adelson [84] demonstrate steerability of the directional derivatives $-2x\exp[-(x^2+y^2)]$ and $-2y\exp[-(x^2+y^2)]$ of the Gaussian $\exp[-(x^2+y^2)]$ by using sins, cosines, and more complex trigonometric functions to interpolate across orientations. This approach does not yet seem to have biological support, and was not needed to achieve the computational compression that we found using Gaussian interpolation.

In summary, alternative models either have different goals or different computational properties than the What-and-Where filter that is proposed herein. These differences can ultimately be traced to our group's focus on *self-organizing* modules for invariant pattern recognition, 3D boundary and surface representation, motion analysis, and visual search. The What-and-Where filter adds a component to this emerging architecture that enables fast, parallel search to occur for desired targets in a scene. None of the algorithms reviewed above yet seem able to do this.

15. CONCLUSION

The parallel and hybrid serial–parallel What-and-Where filters use a combination of cliff-like oriented filters, Gaussian interpolation, and suitably organized competitive interactions across orientation and size to produce an output image that is invariant under translation, rotation, and scaling of the input. By breaking this preprocessing stage into What-and-Where channels, the amount of information that is lost about the figure's form is minimized. The What channel provides invariant form information for purposes of pattern learning and recognition. The Where channel retains the location, orientation, and size of the image for use in applications such as the allocation of spatial attention, image understanding, and the planning of motor trajectories to contact the figure in space. This analysis has disclosed some of the computational issues, uncertainties, and trade-offs, such as the role of cliff-receptive fields to achieve good positional localization, competition across scales to deal with the double peak problem, self-similar interpolation across orientation but not scale, and compensations for normalized or unnormalized filters, that are needed for accurate and efficient computation. In particular, for image processing applications carried out in software, partial serialization of the Where channel, combined with Gaussian interpolation across orientation, achieves accurate invariance using a relatively small number of filters.

The present algorithm has a number of limitations that need to be overcome by future research. When a target has an almost symmetric shape, it may generate an ambiguous rotational estimate, even though its features are not symmetrically distributed over its surface. This degenerate case may most simply be handled by generating multiple rotated images of objects that fail to activate a winning orientation. If the target is partially occluded by a nearer object, the present 2D algorithm may generate a biased representation of its position and orientation. To overcome this problem, a prior stage of 3D figure–ground separation would be needed to separate the occluding and occluded objects onto different depth planes, and complete the boundary and surface representations of the occluded objects on its own depth plane. Such 3D algorithms are presently under development [19–21]. They highlight other problems that an image preprocessor must handle in order to process realistic 3D scenes.

More generally, primate brains use a What-and-Where strategy to divide the cortical processing load between object recognition and spatial localization tasks (see Section 1). Carpenter and Grossberg [59], Grossberg [19], and Grossberg, Mingolla, and Ross [35] have modeled how the brain's What-and-Where strategy may be embedded in a larger image processing architecture wherein 3D boundary and surface representations of a scene interact reciprocally with attentive learned object categories (What stream) and spatial maps (Where stream) to focus attention upon desired objects in a 3D scene and to search for such targets amid various types of distractors. These studies indicate how a multiplexed spatial map, such as the Where filter described herein, may organize the interactions between spatial and object representations that are used to interpret and understand the visual world. Future research will work to further develop these models into an autonomous architecture for image understanding and to explain progressively larger databases about primate 3D vision, visual search, and object recognition. The present research contributes to this task by disclosing some of the computational problems that need to be solved by a Where system that is based upon oriented filters and by defining several efficient algorithms that solve them.

The present work does not, however, show how the brain uses Where information to generate a representation of objects in the What stream that is invariant under changes in position, size, and orientation. This process is handled here, for purposes of short-term application, by simply shifting the object into a canonical representation using the Where information. The present work also does not integrate the Where process into the larger image understanding architecture that is summarized in Figs. 2 and 4. Further study is needed of how a self-organizing algorithm like Fusion ARTMAP can autonomously learn which combinations of What and Where information predict a particular interpretation of a scene. On the other hand, as noted above, models of each stage in this architecture are now available, so the process of system synthesis can begin.

ACKNOWLEDGMENTS

The authors thank Cynthia E. Bradford for her valuable assistance in the preparation of the manuscript. Dr. Courosh Mehanian participated in the specification of the oriented filters defined by Eqs. (1)–(4).

REFERENCES

1. D. Casasent and D. Psaltis, Position, rotation and scale invariant optical correlation, *Appl. Opt.* **15**, 1976, 1795–1799.
2. P. Cavanagh, Image transforms in the visual system, in *Figural Synthesis* (P. C. Dodwell and T. Caelli, Eds.), pp. 185–218, Erlbaum Associates, Hillsdale, NJ, 1984.
3. A. Mishkin, L. G. Ungerleider, and K. A. Macko, Object vision and spatial vision: Two cortical pathways, *Trends in Neurosci*. **6**, 1983, 414–417.
4. L. G. Ungerleider and M. Mishkin, Two cortical visual systems: Separation of appearance and location of objects, in *Analysis of Visual Behavior* (D. L. Ingle, M. A. Goodale, and R. J. W. Mansfield, Eds.), pp. 549–586, MIT Press, Cambridge, MA, 1982.
5. M. A. Goodale and D. Milner, Separate visual pathways for perception and action, *Trends in Neurosciences* **15**, 1992, 20–25.

6. S. Grossberg, Neural networks for visual perception in variable illumination, *Optics News* **8**, 1988, 5–10.
7. S. Grossberg and D. Todorović, Neural dynamics of 1-D and 2-D brightness perception: A unified model of classical and recent phenomena, *Perception and Psychophysics* **43**, 1988, 241–277.
8. A. G. Andreou and K. A. Boahken, Modeling inner and outer plexiform retinal processing using nonlinear coupled resistive networks, in *Human Vision, Visual Processing, and Digital Display, II.* Vol. 1453, pp. 270–281, Society of Photo-Optical Instrumentation Engineers, Bellingham, WA, 1991.
9. S. Grossberg, Cortical dynamics of 3-D form, color, and brightness perception, I. Monocular theory. *Perception Psychophys.* **41**, 1987, 87–116.
10. A. Gove, S. Grossberg, and E. Mingolla, Brightness perception, illusory contours, and corticogeniculate feedback, *Visual Neurosci*. **12**, 1995, 1027–1052.
11. S. Grossberg and E. Mingolla, Neural dynamics of perceptual grouping: Textures, boundaries, and emergent segmentations, *Perception and Psychophysics* **38**, 1985, 141–171.
12. S. Grossberg and E. Mingolla, Neural dynamics of surface perception: Boundary webs, illuminants, and shape-from-shading, *Computer Vision, Graphics, and Image Processing* **37**, 1987, 116–165.
13. S. Grossberg, E. Mingolla, and D. Todorović, A neural network architecture for preattentive vision, *IEEE Transactions on Biomedical Engineering* **36**, 1989, 65–84.
14. S. Grossberg, E. Mingolla, and J. Williamson, Synthetic aperture radar processing by a multiple scale neural system for boundary and surface representation, *Neural Networks* **8**, 1995, 1005–1028.
15. G. A. Carpenter, S. Grossberg, and C. Mehanian, Invariant recognition of cluttered scenes by a self-organizing ART architecture: CORT-X boundary segmentation, *Neural Networks* **2**, 1989, 169–181.
16. S. Grossberg and L. Wyse, A neural network architecture for figure-ground separation of connected scenic figures, *Neural Networks* **4**, 1991, 723–742.
17. S. Grossberg and L. Wyse, Figure-ground separation of connected scenic figures: Boundaries, filling-in, and opponent processing, in *Neural Networks for Vision and Image Processing* (G. A. Carpenter and S. Grossberg, Eds.), pp. 161–194, MIT Press, Cambridge, MA, 1992.
18. S. Grossberg, A solution of the figure-ground problem for biological vision, *Neural Networks* **6**, 1993, 463–483.
19. S. Grossberg, 3-D vision and figure-ground separation by visual cortex, *Perception and Psychophysics* **55**, 1994, 48–120.
20. S. Grossberg, Cortical dynamics of 3-D figure-ground perception of 2-D pictures, 1997.
21. S. Grossberg and N. P. McLoughlin, Cortical dynamics of 3-D surface perception: Binocular and half-occluded scenic images, *Technical Report CAS/CNS-TR-95-022*, Boston University, 1996. *Neural Networks*, in press.
22. 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**, 1987, 54–115.
23. G. A. Carpenter and S. Grossberg, ART 2: Stable self-organization of pattern recognition codes for analog input patterns, *Applied Optics* **26**, 1987, 4919–4930.
24. G. A. Carpenter, S. Grossberg, and D. B. Rosen, ART2-A: An adaptive resonance algorithm for rapid category learning and recognition, *Neural Networks* **4**, 1991, 493–504.
25. G. A. Carpenter, S. Grossberg, and D. B. Rosen, Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system, *Neural Networks* **4**, 1991, 759–771.
26. Y. Asfour, G. A. Carpenter, S. Grossberg, and G. W. Lesher, Fusion ARTMAP: A neural network architecture for multi-channel data fusion and classification, *Technical Report CAS/CNS-93-006*, Boston University, 1993. [In *Proceedings of the World Congress on Neural Networks, Portland, II*, pp. 210–215, Erlbaum Associates, Hillsdale, NJ]
27. G. A. Carpenter and S. Grossberg, Fuzzy ARTMAP: Supervised learning, recognition, and prediction by a self-organizing neural network, *IEEE Communications Magazine*, September, 1992, 38–49.
28. G. A. Carpenter, S. Grossberg, and J. H. Reynolds, ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network, *Neural Networks* **4**, 1991, 565–588.
29. G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen, Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks* **3**, 1992, 698–713.
30. J. R. Williamson, Gaussian ARTMAP: A neural network for fast incremental learning of noisy multidimensional maps, *Neural Networks*, 1996.
31. G. A. Carpenter and W. D. Ross, ART-EMAP: A neural network architecture for learning and prediction by evidence accumulation, in *Proceedings of the World Congress on Neural Networks, Portland, III*, pp. 649–656, Erlbaum Associates, Hillsdale, NJ, 1993.
32. G. A. Carpenter and W. D. Ross, ART-EMAP: A neural network architecture for object recognition by evidence accumulation, *IEEE Transactions on Neural Networks* **6**, 1995, 805–818.
33. G. Bradski and S. Grossberg, A neural architecture for 3-D object recognition from multiple 2-D views, in *Proceedings of the World Congress on Neural Networks, San Diego, IV,* pp. 211–219, Erlbaum Associates, Hillsdale, NJ, 1994.
34. G. Bradski and S. Grossberg, Fast learning VIEWNET architectures for recognizing 3-D objects from multiple 2-D views, *Neural Networks* **8**, 1995, 1053–1080.
35. S. Grossberg, E. Mingolla, and W. Ross, A neural theory of attentive visual search: Interactions of boundary, surface, spatial, and object representations. *Psychological Review* **101**, 1994, 470–489.
36. Y. Yeshurun and E. L. Schwartz, Shape description with a space-variant sensor: Algorithms for scan-path, fusion and convergence over multiple scans, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 1989, 1217–1222.
37. P. M. Daniel and D. Whitteridge, The representation of the visual field on the cerebral cortex in monkeys, *Journal of Physiology* **159**, 1961, 203–221.
38. B. Fischer, Overlap of receptive field centers and representation of the visual field in the cat's optic tract, *Vision Research* **13**, 1973, 2113–2120.
39. E. L. Schwartz, Spatial mapping and spatial vision in primate striate and inferotemporal cortex, in *Sensory Experience, Adaptation, and Perception* (L. Spillmann and B. R. Wooten, Eds.). pp. 73–104, Erlbaum Associates, Hillsdale, NJ, 1984.
40. R. B. H. Tootell, M. S. Silverman, E. Switkes, and R. L. DeValois, Deoxyglucose analysis of retinotopic organization in primate striate cortex. *Science* **218**, 1982, 902–904.
41. D. C. van Essen, W. T. Newsome, and J. H. R. Maunsell, The visual representation in striate cortex of macaque monkey: Asymmetries, anisotropies, and individual variability, *Vision Research* **24**, 1984, 429–448.
42. S. Grossberg, Contour enhancement, short-term memory and constancies in reverberating neural networks, *Stud. Appl. Math.* **52**, 1973, 217–257. [Reprinted in S. Grossberg, *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control*. Boston: Reidel Press, 1982]
43. J. Feldman and D. Ballard, Connectionist models and their properties. *Cognitive Science* **6**, 1982, 205–254.
44. K. P. Hadeler, On the theory of lateral inhibition, *Kybernetik* **14**, 1974, 161–165.
45. C. Koch and S. Ullman, Shifts in selective visual attention: Towards the under-laying neural circuitry, *Human Neurobiology* **4**, 1985, 219–227.

46. J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, Modeling visual attention via selective tuning, *Artificial Intelligence*, 1995, in press.
47. D. H. Hubel and T. N. Wiesel, Functional architecture of macaque monkey visual cortex, *Proceedings of the Royal Society of London (B)* **198**, 1977, 1–59.
48. M. Seibert and A. M. Waxman, Spreading activation layers, visual saccades, and invariant representations for neural pattern recognition systems, *Neural Networks* **2**, 1989, 9–27.
49. S. Grossberg, A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans, in *Progress in Theoretical Biology*, R. Rosen and F. Snell (Eds.), Vol. 5, Academic Press, New York, 1978.
50. S. Grossberg and M. Kuperstein, *Neural Dynamics of Adaptive Sensory Motor Control: Expanded Edition*, Elsevier Science, Tarrytown, NY, 1986/1989.
51. M. A. Cohen and S. Grossberg, Absolute stability of global pattern formation and parallel memory storage by competitive neural networks, *IEEE Transactions on Systems Man, and Cybernetics* **SMC-13**, 1983, 815–826.
52. M. Cohen and S. Grossberg, Neural dynamics of speech and language coding: Developmental programs, perceptual grouping, and competition for short term memory, *Human Neurobiology* **5**, 1986, 1–22.
53. R. Coultrip, R. Granger, and G. Lynch, A cortical model of winner-take-all competition via lateral inhibition, *Neural Networks* **5**, 1992, 47–54.
54. S. A. Ellias and S. Grossberg, Pattern formation, contrast control, and oscillations in the short term memory of shunting on-center off-surround networks, *Biological Cybernetics* **20**, 1975, 69–98.
55. B. Ermentrout, Complex dynamics in winner-take-all neural nets with slow inhibition, *Neural Networks* **5**, 1992, 415–431.
56. S. Grossberg, Competition, decision, and consensus, *Journal of Mathematical Analysis and Applications* **66**, 1978, 470–493.
57. S. Grossberg and D. Levine, Some developmental and attentional biases in the contrast enhancement and short term memory of recurrent neural networks, *Journal of Theoretical Biology* **53**, 1975, 341–380.
58. G. A. Carpenter and S. Grossberg (Eds.), *Pattern Recognition by Self-Organizing Neural Networks*, MIT Press, Cambridge, MA, 1991.
59. G. A. Carpenter and S. Grossberg, Normal and amnesic learning, recognition, and memory by a neural model of cortico-hippocampal interactions, *Trends in Neurosciences* **16**, 1993, 131–137.
60. R. Desimone, Neural circuits for visual attention in the primate brain, in *Neural Networks for Vision and Image Processing* (G. A. Carpenter and S. Grossberg, Eds.), pp. 343–364, MIT Press, Cambridge, MA, 1992.
61. E. Miller, L. Li, and R. Desimone, Activity of neurons in anterior inferior temporal cortex during a short-term memory task, *Journal of Neuroscience* **13**, 1993, 1460–1478.
62. B. Motter, Neural correlates of attentive selection of color or luminance in extrastriate area V4, *Journal of Neuroscience* **14**, 1994, 2178–2189.
63. J. Reynolds, L. Chelazzi, S. Luck, and R. Desimone, Sensory interactions of selective spatial attentive in macaque area V2, *Society for Neuroscience Abstracts* **20**, 1994, 1054.
64. J. Reynolds, J. Nicholas, L. Chelazzi, and R. Desimone, Spatial attention protects macaque V2 and V4 cells from the influence of non-attended stimuli, *Society for Neuroscience Abstracts* **21**, 1995, 356.
65. S. Grossberg, The attentive brain, *American Scientist* **83**, 1995, 438–449.
66. M. Bravo and R. Blake, Preattentive vision and perceptual groups, *Perception* **19**, 1990, 515–522.
67. A. Cohen and R. B. Ivry, Density effects in conjunctive search: Evidence for a coarse location mechanism of feature integration, *Journal of Experimental Psychology: Human Perception and Performance* **17**, 1991, 891–901.
68. J. T. Enns and R. A. Rensink, Influence of scene-based properties on visual search, *Science* **247**, 1990, 721–723.
69. Z. J. He and K. Nakayama, Surface features in visual search, *Nature* **359**, 1992, 231–233.
70. J. T. Mordkoff, S. Yantis, and H. Egeth, Detecting conjunctions of color and form in parallel, *Perception and Psychophysics* **5**, 1990, 157–168.
71. J. Wolfe, K. R. Cave, and S. L. Franzel, Guided search: An alternative to the feature integration model of visual search, *Journal of Experimental Psychology: Human Perception and Performance* **15**, 1989, 419–433.
72. J. M. Wolfe and S. R. Friedman-Hill, Part-whole relationships in visual search, *Investigative Ophthalmology and Visual Science* **33**, 1992, 1355.
73. B. A. Olshausen, C. H. Anderson, and D. C. van Essen, A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information, *Journal of Neuroscience* **13**, 1993, 4700–4719.
74. J. J. Hopfield, Neurons with graded response have collective computational properties like those of two-state neurons, *Proceedings of the National Academy of Sciences* **81**, 1984, 3058–3092.
75. G. A. Carpenter and S. Grossberg, Integrating symbolic and neural processing in a self-organizing architecture for pattern recognition and prediction, in *Artificial Intelligence and Neural Networks: Steps towards Principled Predictions* (V. Honavar and L. Uhr, Eds.), pp. 387–421, Academic Press, San Diego, 1994.
76. R. Remington and L. Pierce, Moving attention: Evidence for time-invariant shifts of visual selective attention, *Perception and Psychophysics* **35**, 1984, 393–399.
77. B. Kröse and B. Julesz, The control and speed of shifts of attention, *Vision Research* **29**, 1989, 1607–1619.
78. S. Grossberg and M. Rudd, A neural architecture for visual motion perception: Group and element apparent motion, *Neural Networks* **2**, 1989, 421–450.
79. S. Grossberg and M. E. Rudd, Cortical dynamics of visual motion perception: Short-range and long-range apparent motion, *Psychological Review* **99**, 1992, 78–121.
80. G. Francis and S. Grossberg, Cortical dynamics of form and motion integration: Persistence, apparent motion, and illusory contours, *Vision Research* **36**, 1996, 149–173.
81. P. A. Kolers, *Aspects of Motion Perception*, Pergamon Press, Oxford, 1972.
82. S. Grossberg, Why do parallel cortical systems exist for the perception of static form and moving form? *Perception and Psychophysics* **49**, 1991, 117–141.
83. S. Grossberg and E. Mingolla, Visual motion perception, in *Encyclopedia of Human Behavior*, Vol. 4, (V. S. Ramachandran, Ed.), pp. 469–486. Academic Press, New York, 1994.
84. W. T. Freeman and E. H. Adelson, Steerable filters, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**, 1991, 891–906.