Toward a Computational Model of Visual Attention

John K. Tsotsos

The Need for Attentional Processing in Vision

In principle, it seems possible to model visual perception computationally (Tsotsos, 1993a). If a vision system knows which subset of an image corresponds to an object and the object type is known, the task of matching image subset to object model is straightforward. In fact, this is a subarea of computational vision in which successful algorithms exist (e.g., Dickinson et al., 1992). The trick is to quickly determine which is the image subset of interest and the corresponding object model. However, far too much computation is required to solve this problem in its general form, guaranteeing that the optimal solution is found in all cases. There is an exponential number of possible image subsets against which to match each potential object model. This conclusion can be proved formally using the methods from theoretical computational complexity. Optimal solutions seem computationally intractable in any implementation, machine or neural (see Tsotsos, 1988, 1989, 1990, 1992, 1993a).

The prevailing argument for why the brain needs visual attention is that there is insufficient neural machinery to deal with all stimuli equally. Broadbent (1971, p. 147), for example, points out "The obvious utility of a selection system is to produce an economy in mechanism. If a complete analysis were performed even on neglected messages, there seems no reason for selection at all." Given the mismatch between brain capacity and complete analysis of all input stimuli, the task facing perceptual theorists is to discover the balance that Nature has achieved among at least three competing requirements: how much information to process and to what degree, how much brain capacity can be devoted to the task, and how quickly must an organism respond to perceptual stimuli.

The remainder of this chapter is devoted to presenting a theoretical foundation for modeling visual attention, followed by descriptions of three current computational hypotheses for modeling attention. For additional background, the reader is referred to Allport (1989) and Colby (1991).

A "First Principles" Argument

The first principles required are straightforward: images, a model base of known objects and events, and an objective function to be optimized that reflects how well an image subset matches a particular member of the model base. One common experimental paradigm, visual search, has been cast into a formal framework using these primitive elements. In Tsotsos (1989), it was proven that visual search, in the case where explicit targets are given in advance, has time complexity which is linear in the size of the image (and this linear response time vs display size is verified experimentally in a large body of work). If, on the other hand, no explicit target is provided, the task is NP-Complete; it is currently believed that such problems are computationally intractable regardless of the implementation, whether it be neural or machine. The intractability is due solely to the combinatorial nature of selecting which parts of the input image are to be processed; there are an exponential number of such image subsets. Since those proofs are based on more abstract yet equally difficult computational problems, it is instructive to consider how the computer science community deals with such combinatorial problems.

For such problems, algorithms have been developed that are not guaranteed to always find the best solution, but can find solutions quickly given some error tolerance. The goal is to find the subset of the input that maximizes an objective function. Strategies for developing partial solutions are exploited to guide search through the space of possibilities so that as few solutions are generated as possible before the best one is located. The intractability of these problems is not due to the computation of the objective function; rather the problem is so difficult only because there is an exponential number of possible solutions to explore. The best of the algorithms are ones that exploit parallel processing; even so, all of them require some serial search through a set of possible solutions (Tsotsos, 1992).

What could the objective function for vision be? Whether a given neuron computes a response that represents a specific object, a specific scene, or a portion of a code as part of a distributed representation is not relevant to this discussion. What is important is that for any particular natural scene a potentially large number of neurons will initially respond with some degree of strength simply because the receptive fields of neurons in higher level areas are so large they will contain elements that might be part of the selectivity profiles of many cells. This large initial set of responding neurons may be considered as a "first guess" as to the contents of a scene. The mapping from image subset to responding neurons is one-to-many; similarly, there are 2^R image subsets within any neuron's receptive field where R is the number of receptors in the receptive field, and thus the mapping from neuron to image subset is also one-to-many. Thus, there is no unique one-to-one mapping between image subsets and neurons. How can this ambiguity be corrected? An objective function is required that reflects this ambiguity and provides a measure for its reduction.

In the formulation of the visual search problem such an objective function was proposed (Tsotsos, 1989). There is one objective function for each known object or image event. The best match is the image subset and model that exhibits the smallest matching error and the model must explain (or cover) as much of the image subset as possible. The brute-force search strategy then is to match each objective function against all possible image subsets. Given formally, the best fit of model to data is sought such that the following is satisfied:

$$\sum_{x \in M, j_x \in I'} |x - j_x| < \theta \quad \text{and} \quad \sum_{x \in M, j_x \in I'} x \cdot j_x > \phi \qquad (20.1)$$

The first term is the error measure while the second is the cover measure. The input is the set I, I' is a subset of I, and M is a set of values corresponding to a particular object or event in the model base. θ and ϕ are two thresholds. I is not necessarily the image itself but may be a collection of all features computed from a given image. A correspondence between elements of M and elements of I' can be hypothesized where element j_x in I' is the element corresponding to x in M. Each possible combination of correspondences may be considered as a separate hypothesis.

Suppose a test image is made up of 256 pixels and a target image has 64 pixels. The correspondence required above is for each element of the target image (each pixel) to be mapped onto a unique pixel of the test image. This forms a hypothesis about where exactly in the test image the target image is believed to be represented. The spatial organization of the mapping need not preserve the structure of the target stimulus, that is, pixels chosen for the mapping may be arbitrarily distributed throughout the image. In the Marr view of the vision, this is necessarily the case since he did not believe there was a role for task (target) directed computations (Marr, 1982). So for this example, there are $\binom{256}{64}$ such possible, bottom-up mappings (256!/64! 192! is approximately 10^{56}; in general if α is the size of the test image and β is the size of the target image both in pixels, then the number of combinations can be given by a polynomial function of the image size whose highest order term is α^{β}). If spatial

structure is preserved and there is no rotation or scaling of the target in the test image, then there are only 64 possibilities such that the target image is entirely within the test image. Attentional selection may determine which mapping to attempt to verify first; if the first such mapping selected is a good one, a great deal of search can be avoided, otherwise there is the potential for a very inefficient search process. For sufficiently small images and/or massive computational power, this brute force concept will work perfectly well without attention. For the brain, this approach fails.

It is easily shown that equation (20.1), is optimized, that is,

$$\theta = 1 \qquad \text{and} \qquad \phi = \left(\sum_{x \in M} x^2 \right) - 1$$

if and only if the set I' is identical to the set M. That is, the set of features or computations that is represented by I' is of the same type and value as those represented by M. I' is a set of features at the same level of abstraction as M and spatially organized in the same manner as M. The role of attention in the image domain is to localize this set I' in a way such that any interfering or corrupting signals are minimized. In doing so, attention also seeks to increase the discriminability over other such objective functions as quickly as possible. Note that any error and cover functions may be used; they would all behave in the above fashion. The only constraint on these functions is that they lead to convex solution landscapes.

Thus, the central thesis of this chapter is that attention acts to optimize the search procedure inherent in the above "in principle" solution to vision. The main effect of attention is to reduce the number of candidates that is considered in matching, both of image subsets as well as object or event models. Attention operates continuously and automatically: without attention, vision in general is not possible. The models described in this chapter deal only with the localization of the image subset and not with model selection.

Pyramid Processes

Analysis at the complexity level (Tsotsos, 1988) confirms what several have suggested (Uhr, 1972; Burt, 1988; Anderson and Van Essen, 1987; Nakayama, 1991): the computational complexity of vision necessitates pyramidal processing. Although pyramids solve the complexity problem by reducing the size of the representations to be processed, they introduce others. Consider the simple three-level pyramid shown in figure 20.1 where each node computes some possibly non-linear weighted sum of its inputs as its output value in a feedforward manner.

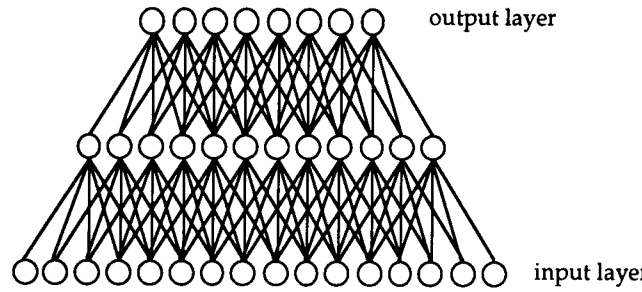

Figure 20.1
A simple pyramidal processing architecture.

Suppose that the input stimulates only one of the centrally located input layer units. That single unit will cause a response in all the units of the output layer simply due to its connectivity. This causes the input to be blurred across the output layer. Similarly, a given unit in the output layer is activated by several units of the input layer, thus responses exhibit a dependence on spatial context. If there are two separate units active in the input layer, they will both activate large parts of the output layer and will overlap for a large portion of the pyramid. This can lead to serious interpretation ambiguities. The examples described all assume that information flow is from input to output layer (data driven). However, information flow in the visual cortex appears to be bidirectional. It is easy to see that the same kinds of problems arise if information flows from top to bottom (task driven). Although pyramid structures help reduce the computational complexity of information processes via convergence of information, they corrupt the signals flowing through them unless some additional mechanisms are included. Each of the following proposals for modeling visual attention provides different solutions to these problems of information flow as well as to the problem of attentional selection.

The Major Computational Hypotheses

There are several major classes of hypotheses for the computational modeling of visual attention, described by the terms *selective routing*, *temporal tagging*, and *selective tuning*.

The Selective Routing Hypothesis

Several models fall into the *selective routing hypothesis* category. The first is that of Koch and Ullman (1985). The idea has found wide acceptance and is used as part of a number of models. The model includes the following elements: (1) an early representation, computed in parallel, permitting separate representations of several stimulus

characteristics; (2) a selective mapping from these representations into a central nontopographic representation such that this central representation at any instant contains only the properties of a single location of the visual scene; (3) a winner-take-all (WTA) network implementing the selection process based on one major rule: conspicuity of location (minor rules of proximity or similarity preference are also suggested); and (4) inhibition of this selected location causes an automatic shift to the next most conspicuous location. The other models in the selective routing category and the models in the temporal tagging category share these basic elements. The selective tuning model includes elements 1, 3, albeit with an entirely new formulation, and 4.

Feature maps code conspicuity within a particular feature dimension. The saliency map combines information from each of the feature maps into a global measure where points corresponding to one location in a feature map project to single units in the saliency map. Saliency at a given location is determined by the degree of difference between that location and its surround (as suggested by Julesz and Bergen, 1983, with their texton difference idea and further explored by Nothdurft, 1993, who showed that feature contrast is the major determinant in speed of visual search and not feature values per se). Different features may be weighted differently or their contribution may be modulated by higher-order computations. Details on the construction of this representation are not given. The WTA network implements a parallel computation based on the values in the saliency map localizing the most conspicuous location. Due to biological constraints on connectivity as well as theoretical convergence difficulties, the WTA takes a particular form; it requires a tree of intermediate nodes breaking up the computation into smaller subtasks and permitting better convergence properties. If the size of the saliency map is n units, and the branching factor of the intermediate tree is m, then the network requires $\log_m n$ comparisons to determine the globally most salient item. Then, a second pyramid marks the location of the most salient item and through another $\log_m n$ steps the most salient item reaches the output of the system. The WTA will not converge if there are two equally strong items. A shift of attention thus requires at most $2 \log_m n$ time steps. Faster convergence can be achieved if locations are physically closer to each other.

The WTA algorithm may no longer be considered biologically plausible because its time course does not agree with current observations. Kröse and Julesz (1989) show that shifts of attention do not take time proportional to the distance between items but rather are accomplished in constant time; also Remington and Pierce (1984) report no topographic relationship on time to shift attention. The intermediate tree of computations has yet to find an anatomical correlate, but perhaps most importantly, the mechanism does not immediately yield the kinds of attention-related receptive field changes observed in areas such as V4 (Moran and Desimone, 1985).

The shifter circuit model, the second in this category, presented a strategy for information flow in stereopsis, visual attention, and motion perception (Anderson and Van Essen, 1987). The model enables the realignment of successive representations in the processing stream starting in the lateral geniculate nucleus and the input layers of area V1. The realignment is based on the preservation of spatial relationships, thus the name "shifter" circuits. The shift is accomplished by a succession of stages linked by diverging excitatory inputs. Control of the direction of shift is accomplished at each stage by inhibitory neurons that selectively suppress sets of ascending inputs. For visual attention, the routing stages are grouped into small and large scale shifts. Control signals are generated externally to the main processing stream. If shifts are assumed to be contiguous it is straightforward to show that this strategy requires many thousands more connections per neuron than the accepted average figure of 1000 for each of fan-in and fan-out.

The Olshausen, Anderson, and Van Essen (1994) model is an elaboration of the shifter circuit idea; a partial implementation with simulation results is also included. The problem described above with the original shifter circuits model is remedied via a clever restructuring of the connectivity patterns between layers. By allowing the spacing between neighboring connections to increase in successively higher layers, the routing network has early layers that are well suited for small-scale shifts while the higher layers can implement larger-scale shifts. The key goal of the Olshausen et al. mechanism is to form position- and scale-invariant representations of objects in the visual field. This is accomplished via a set of control neurons, originating in the pulvinar, that dynamically modifies synaptic weights of intracortical connections so that information from a selected region of primary visual cortex is routed to higher areas. The topography of the selected portion of the visual field is preserved by the resulting transformations.

The dynamics of the control neurons are defined using simple differential equations and control neurons receive their input from a saliency map representation. They suggest that the posterior parietal areas act as the saliency map representation. Each node in the processing hierarchy performs a simple linear weighted sum operation.

Selected objects in the visual field are found by the Koch and Ullman mechanism, then routed to the top layer of the processing pyramid (inferotemporal cortex, IT). The selected object is transformed by the routing so that it spans the top-level representation. There, a Hopfield associative memory is used for recognition (Hopfield, 1982).

This model is presented in detail and the results of the computer simulations show performance as expected. Rotations are not handled and it does not seem that the shifter kinds of connectivities are sufficient to ensure rotation-invariant representations. Finally, there is no evidence yet that area IT is an image-centered representation of only a subset of the retinal image.

The Temporal Tagging Hypothesis

The *temporal tagging hypothesis* proposes that selected items are distinguished as they flow through the processing system because they are tagged by superimposing a frequency modulation of 40 Hz on the signal. Crick and Koch (1990) suggest that an attentional mechanism binds together all those neurons whose activity relates to the relevant features of a single visual object. This is done by generating coherent semisynchronous oscillations in the 40–70 Hz range. These oscillations then activate a transient short-term memory. These suggestions are not fully developed computationally in that paper. However, in a subsequent effort, Niebur, Koch, and Rosin (1993) detail a model based on those suggestions.

Niebur et al. (1993) assume that salient objects have been selected in the visual field by the Koch and Ullman mechanism. The saliency map is claimed to be found in subcortical areas (superior colliculus or the dorsomedial region of the pulvinar). That is where the attentional modulation is added and this modulation occurs only at the level of primary visual cortex V1. The modulation affects only the temporal structure of the spike trains of V1 neurons but not their mean firing rate. The existence of frequency-selective inhibitory interneurons is assumed in V4. These are required to act as bandpass filters selective to spikes arriving every 25 msec or so. Thus, they would pass temporally tagged spike trains and block other non-frequency-modulated signals. Both Crick and Koch (1990) and Niebur et al. (1993) assume that selective attention activates competition within a stack or microcolumn of neurons in V4. In the presence of multiple stimuli, neurons will compete with each other. Since the outputs of V1 neurons are tagged, their postsynaptic targets in V4 will win in the V4 level competition. They go on to say that there are no attentional effects on firing rates in V1, only in V4 or higher areas.

The model is quite detailed and provides for quantitative single-cell performance predictions; results of their simulations are in terms of firing rates. The agreement with the relevant experimental data is good. Several major issues arise from this model. First, because the model assumes the selection mechanism of Koch and Ullman, it inherits the timing problems described above. Second, if attentional modulation originates in the subcortical areas, then it is difficult to see how the effects of targets or memory items can be accounted for (Haenny, and Schiller, 1988; Chelazzi et al., 1993). In those studies, single V4 and IT neurons were found that seemed to code the target stimulus and effect the execution of the task. Within both the routing and tagging models, the path lengths required for communication with external gating control in order to affect this influence seem to be wasteful; a closer locus of attentional control seems more likely on this basis.

The Selective Tuning Hypothesis

The *selective tuning hypothesis* claims attention is used to tune the visual processing architecture in order to overcome the problems with pyramid computation and to permit task-directed processing. Selective tuning takes two forms: spatial selection is realized by inhibition of irrelevant connections and feature selection is realized by inhibition of the units that compute nonselected features. The limited space allows only a brief summary in this review article. The interested reader can refer to more detailed accounts (Tsotsos, 1990, 1993b; Culhane and Tsotsos, 1992a,b). The starting point for the model has been described. The search process that localizes the image subset I' is as follows. A winner-take-all process operates across the entire visual field at the top layer: it computes the global winner. The search process then proceeds to the lower levels. The WTA can accept guidance for areas or stimulus qualities to favor if that guidance were available but operates independently otherwise. To localize the global winner in the visual field, a hierarchy of WTA processes is activated. The global winner activates a WTA that operates only over its direct inputs. This localizes the winner within the top-level winning receptive field. In this way, all of the branches of the hierarchy that do not contribute to the winner are pruned. This pruning idea is then applied recursively to successively lower layers. The end result is that from a globally strongest response, the cause of that largest response is localized in the sensory field at the earliest levels. The paths remaining may be considered the pass zone while the pruned paths form the inhibitory zone of an atten-

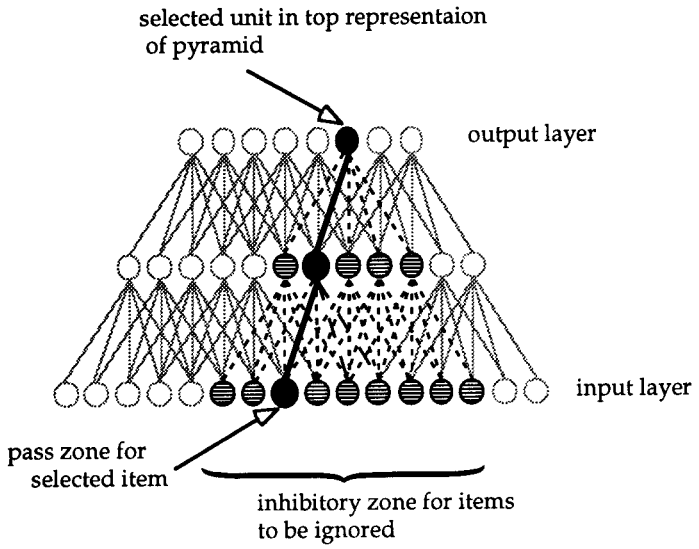

Figure 20.2
An illustration of the inhibitory beam of the selective tuning model. The black solid nodes and connections are those selected; the gray open nodes and gray connections are those that are "don't cares," and the dashed connections between the striped nodes are the connections that are inhibited.

tional beam (see figure 20.2). The WTA does not violate biological connectivity constraints. A formal relationship exists between this model and the adaptive beamforming concept of adaptive filter theory used for antenna arrays (Haykin, 1991).

Due to the localizing action of top-down pruning described above, if one were to "record" the output of a unit at the top of the processing pyramid, the time course of the response would show an initial high value, then gradually decrease over time as successively lower layers are pruned away. The decrease would not be due to any suppressive effects acting on this unit; rather, the pruning action of removing parts of its supporting subpyramid leads to a reduction in response over time (qualitatively agreeing with the time course of IT neuron responses as observed by Chelazzi et al., 1993; Gochin et al., 1991; Oram and Perrett, 1992).

The process of selection requires two traversals of the pyramid; the overall time course is consistent with that observed in Chelazzi et al. (1993) (more on this later). These traversals involve

1. Computing pyramid representation in a bottom-up fashion, modified by the biases if available

2. Detecting and localizing the most salient item in a top-down manner, pruning parts of the pyramid that do not contribute to the most salient item, and continuously propagating changes upward

The remainder of this section provides some detail on how this may be accomplished.

The model requires several different types of computing units. Interpretive units compute the visual features. Gating units compute the WTA result across the inputs of a particular interpretive unit and gate winning input through to the next higher interpretive units. Gating control units control the downward flow of selection through the pyramid and are responsible for the signals, which either activate or shut down the WTA processes. Bias units provide top-down, task-related selection via multiplicative inhibition. Figure 20.3 gives the overall architecture that ties these basic units types together. A grouping consisting of one interpretive unit, its associated gating control and bias unit, the set of WTA gating units on the inputs of the interpretive unit, and associated connections will be termed an assembly.

The notation to be used below is now introduced; the figure should be used as a supplement to this description. Physical units are distinguished from their value by the use of a "hat" ("^") where the hatted variable represents the unit, and the same variable without the hat represents the value of the unit. The first subscript gives the layer of the hierarchy in which the unit is found; the second subscript gives the assembly in which the unit is found; the third subscript represents an identifier used to distinguish units within a set. Superscripts always refer to time, in particular, time within the iterations of a given WTA process. Further,

$\hat{I}_{l,k}$ is the interpretive unit in assembly k in layer l; $I_{l,k}$ is its positive real value representing its response;

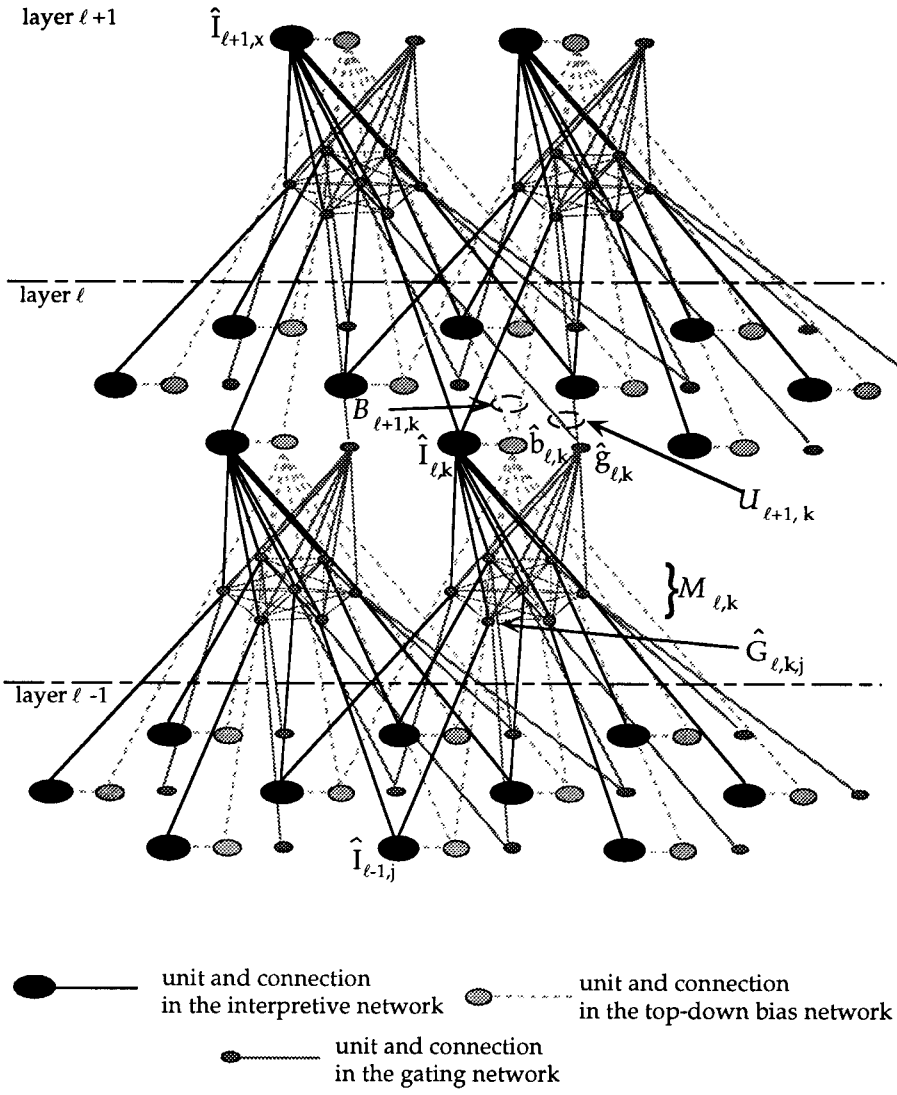

Figure 20.3
Three layers of the processing pyramid showing the details of the unit and connection types for the selective tuning model. Refer to the text for further description.

$\hat{G}_{l,k,j}$ represents the jth WTA gating unit, in assembly k in layer l linking $\hat{I}_{l,k}$ with $\hat{I}_{l-1,j}$;

$\hat{g}_{l,k}$ is the gating control unit for the WTA over the inputs to $\hat{I}_{l,k}$;

$\hat{b}_{l,k}$ is the bias unit for $\hat{I}_{l,k}$;

$M_{l,k}$ is the set of gating units for unit $\hat{I}_{l,k}$;

$U_{l+1,k}$ is the set of gating units in layer $l+1$ making efferent connections to $\hat{g}_{l,k}$;

$B_{l+1,k}$ is the set of bias units in layer $l+1$ making efferent connections to $\hat{b}_{l,k}$.

The standard iterative formulation for a WTA process is (having its roots in Feldman and Ballard, 1982):

$$C_k^t = C_k^{t-1} - \sum_{\substack{i \in V \\ i \neq k}} w_{i,k} C_i^{t-1} \tag{20.2}$$

where the values of the units in the WTA process ($C_j \in V$ for all defined j) at time t are given by C_k^t, all units are connected to all others, and the relative amount of influence of unit i on unit k is reflected by the weight $w_{i,k}$. All units decay in value with time; the process terminates when all units but one have value of 0.0. In the new formulation of the WTA for the selective tuning model, winning units (there may be more than one) maintain their actual response strength while other units decay. In this way the instantaneous representation of winners in the hierarchy always reflects the actual input. This is ac-

complished using a simple observation: if the inhibitory signal is based on the response differences, then an implicit but global ordering of response strengths is imposed on the network. The largest item will thus not be inhibited, but will participate in inhibiting all other units. The smallest unit will not inhibit any other units but will be inhibited by all. $\Delta_{i,j}$ represents this contribution based on response differences. The contribution in the WTA from unit i to unit j is set such that

$$\text{if } 0 < \theta < G^{t-1}_{l,k,i} - G^{t-1}_{l,k,j} \quad \text{then}$$

$$\Delta_{i,j} = G^{t-1}_{l,k,i} - G^{t-1}_{l,k,j}, \quad \text{else } \Delta_{i,j} = 0 \tag{20.3}$$

$G^t_{l,k,j}$ is the positive real-valued response of gating unit $\hat{G}_{l,k,j}$ at time t, such that $0 \leqslant G^t_{l,k,j}$. θ is a threshold set to

$$\theta = \frac{Z}{2^\gamma + 1} \tag{20.4}$$

assuming that at least one of the values in the competition has value greater than θ and that Z is their maximum possible value. This setting guarantees convergence within at most γ iterations (Tsotsos, 1993b). The WTA stops once the gating units in the competition are partitioned into two classes: those with value zero, and those with value greater than θ but within θ of each other (the winners). Thus, the term $w_{i,k}C^{t-1}_i$ is equation (20.2) is replaced by $\Delta_{i,j}$.

The second component of the new WTA rule is the signal for providing top-down bias. $\hat{b}_{l,k}$ is the bias unit for $\hat{I}_{l,k}$ with real-value $b_{l,k} \geqslant 0$ defined by

$$b_{l,k} = \min_{\hat{a} \in B_{l+1,k}} \{a\}. \tag{20.5}$$

$B_{l+1,k}$ is the set of bias units in layer $l+1$ making efferent connections to $\hat{b}_{l,k}$. The nature of the bias computation is to inhibit any nonselected units allowing the selected ones to pass through the pyramid without interference. The default value of bias units is 1.0; this value changes only if some other value is inserted at the top of the pyramid due to task information. Since it is assumed that the inhibitory effect is multiplicative, the simplest policy is for bias units to compute the minimum over all top-down bias signals received. Those interpretive units that compute quantities that are not selected are inhibited allowing the selected ones to pass. So, for example, if red items are being sought, the interpretive units that are selective for red stimuli would be unaffected while all other color-selective units would be biased against to some degree.

The WTA is initialized at time t_0 by setting the values of each gating unit to the output of the biased interpretive unit to which it is connected in the layer below

$$G^{t_0}_{l,k,j} = b_{l-1,j}I_{l-1,j} \tag{20.6}$$

These values are computed on the first traversal of the pyramid (the bottom-up traversal).

The next important component of the new WTA rule is the control signal, which turns the selection process on and off. $\hat{g}_{l,k}$ is the gating control unit for the WTA over the inputs to $\hat{I}_{l,k}$ and has value defined by

$$\text{if } \sum_{\hat{a} \in U_{l+1,k}} \{a\} > 0 \quad \text{then} \quad g_{l,k} = 1, \quad \text{else} \quad g_{l,k} = 0 \tag{20.7}$$

where the sum is computed after the networks involved have converged. $\hat{g}_{l,k}$ provides top-down control of the WTA processes by selecting the path of the beam's pass zone depending on the winning WTA units in the next higher layer. If the gating control unit has value one, then the WTA process is turned on; otherwise it is turned off. This is implemented by multiplicatively modifying the iterative rule so that if the WTA is off, all updated values are zero. Using this signal at the top of the pyramid, the entire process is controlled. In this way, the gating units are affected but not the interpretive units; only a pathway is closed down. The value of $\hat{g}_{l,k}$ is zero for all units during the first phase of the process (points 1 and 2 of the three-stage algorithm given earlier). During this first phase, the gating units (all the $\hat{G}_{l,k,j}$) are open and the WTAs are all disabled so that the responses computed by the interpretive units based on the stimulus in a bottom-up fashion can pass through the pyramid. Then the value of $\hat{g}_{l,k}$ becomes one for all the units at the top layer turning on the top-most WTA process. The results of this WTA process then determine the values of $\hat{g}_{l,k}$ for the successively lower layers through the application of equation (20.7) for each lower layer in order.

To enforce stability and so that no oscillations occur, the overall result is rectified (negative new unit values are set to zero) by passing the entire right side of equation (20.2) through a rectifying function $\mathbf{R}$ such that

$$\mathbf{R}[x] = x \text{ if } x > 0, \quad \text{else } \mathbf{R}[x] = 0 \tag{20.8}$$

Each of the preceding functionalities, including the control signals and the WTA action, are incorporated into a new updating rule so that after the stimulus is presented to the input layer and top-down biases are presented to the top layer, no further actions are required. The rule is given by

$$G^t_{l,k,j} = g_{l,k}R\left[G^{t-1}_{l,k,j} - \left(\sum_{\substack{i \in M_{l,k} \\ i \neq j}} \Delta_{i,j}\right)\right] \tag{20.9}$$

The most important consequence of this new rule is that convergence properties are guaranteed. It was proved in Tsotsos (1993b) that this WTA is guaranteed to converge and to not oscillate. This is possible only because the iterative update is based on differences of units and thus only the largest and second largest values need be considered; a two-unit network is thus easy to characterize. There is no logarithmic dependence on either topographic distance or numbers of competitors, thus providing a much better match to experiments (Kröse and Julesz, 1989; Remington and Pierce, 1984). The actual convergence time is dependent only on differences between strengths of signals in the same sense as that observed by Duncan and Humphreys (1989).

Because of the time course of the gating control signals, they, and in turn the units of the pyramid as well, exhibit an oscillatory pattern in time. If attention can shift every 20–50 msec or so [the time between shifts varies with experiment: Sagi and Julesz (1985) found some inspection times to be as short as 17 msec; Saarinen and Julesz (1991) found good performance at 33 msec; Bergen and Julesz (1983) noted 50 msec], then this is the cycle time of the gating control signal as well. Since gating control is set to 0.0 for part of each selection and to 1.0 for the remainder, the signal is periodic in nature with a frequency of 20–50 Hz. This may be considered as an alternative explanation for the oscillations that motivate the temporal tagging model. This gating signal may be

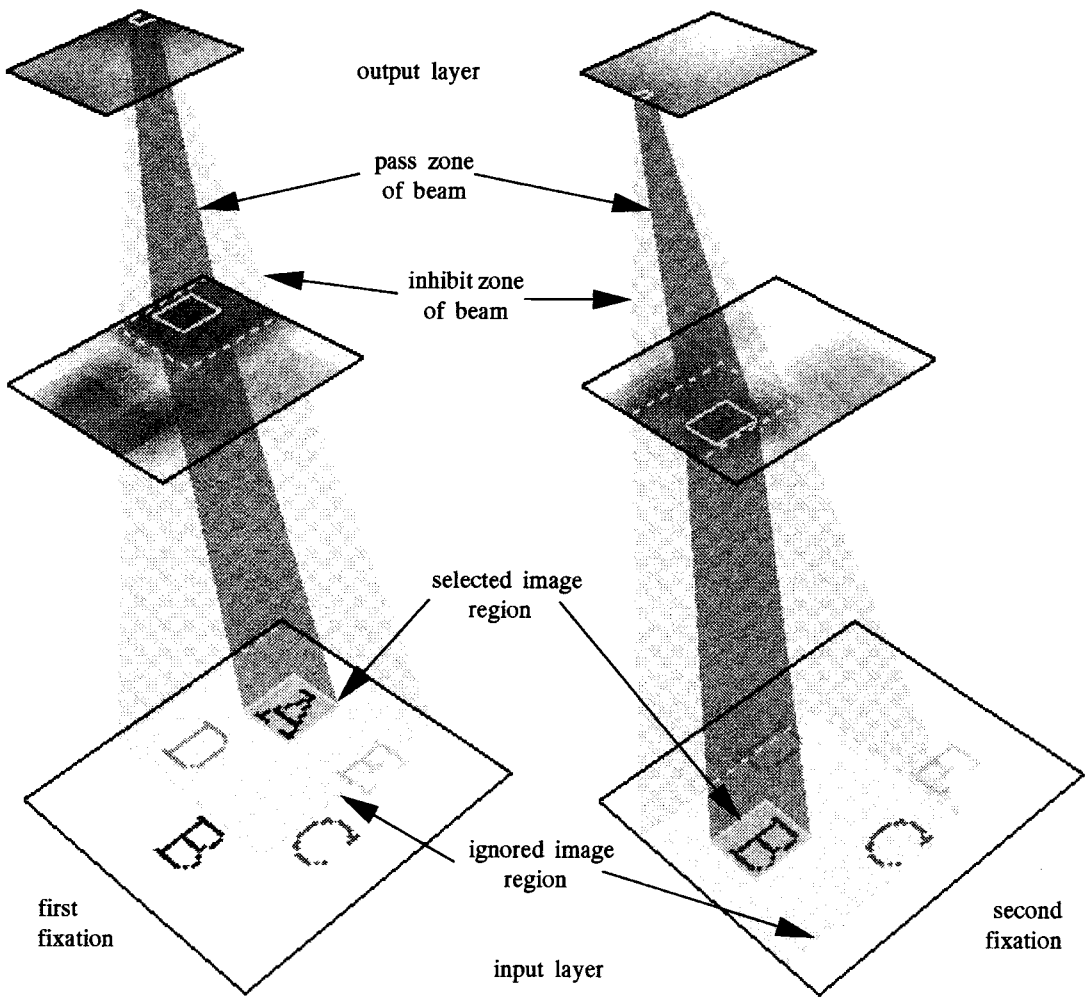

Figure 20.4
A hypothetical pyramid of three layers and a representation of shapes (letters) each with a different luminance. In other words, each letter is made up of pixels all of which are a uniform level of brightness and each letter is of a unique brightness. The selective tuning algorithm finds each in luminance order using the strategy described in the text. The inhibitory and pass zones of the inhibitory beam are clearly seen in the first two attentional fixations shown (first fixation is on the left). The computations of each of the two layers above the input layer is a simple average luminance.

considered as a sort of system clock to use a computational metaphor.

Finally, it is important to note that this algorithm, under biological connectivity constraints, very closely approximates the provably optimal parallel time complexity for finding the maximum value of a given set (Karp and Ramachandran, 1990).

The computer implementations have successfully tested many components of this mechanism (Culhane and Tsotsos, 1992a,b). An example is shown in figure 20.4 where a hypothetical test network of three layers shows the structure and successive shifts of attention for the inhibitory beam for a representation of saliency that incudes only luminance.

Conclusions

This chapter reviews the major computational hypotheses for the modeling of visual attention. They are all based on similar principles: there is insufficient brain capacity to process all visual stimuli to the same degree of detail; early representations of the scene are computed in parallel and these representations are further inspected by a serial process; selection of items to process is implemented by a winner-take-all mechanism using a representation of saliency based on the early representations; the problem of information flow through a processing pyramid must be solved. Yet, the models accomplish these tasks in very different ways. The main conclusion that can be drawn is that although there seems to be broad agreement regarding the basic foundations of modeling, insufficient biological experimentation has been done at this point that might distinguish one model from another in terms of biological realism. In addition, the functionality of all of the models is limited.

The models have much in common in terms of their performance. For example, each of the models offers a believable explanation for the observations of Moran and Desimone (1985). Each can provide accounts of a variety of human visual search experiments in that serial search processes can be simulated. However, a number of important open questions remain that may help to differentiate the models from one another.

The selective routing and temporal tagging models all assume that control of the process that distinguishes selected signals from the others has a source external to the main processing stream (V1 → V2 → V4 → IT, for example). Although the pulvinar has been implicated as playing a role in visual attention, it is by no means clear that its role is that of producing control signals (see Desimone et al., 1990). In contrast, the selective tuning model has control originating within the processing stream itself. An argument may be made supporting the latter scheme on the basis of length of connections; computationally, an argument may be made that minimization of overall connection length is important (Tsotsos, 1990).

Further distinguishing characteristics include the following:

1. The Olshausen et al. model assumes that spatial relationships must be preserved (in the topographic sense) while the temporal tagging and selective tuning models do not. These latter models permit spatial abstraction while the former does not, i.e., single units in IT seem to represent complex objects (as observed by Tanaka et al., 1991) as opposed to pixel-like retinal image copies. Spatial abstraction is a major contributor to the reduction of computational complexity (Tsotsos, 1990).

2. Only the selective tuning model explicitly includes top-down bias.

3. Each model comments on the location of saliency representations. Olshausen et al. suggest that the posterior parietal areas act as the saliency map. Niebur at al. claim it is found in superior colliculus or the dorsomedial region of the pulvinar. The selective tuning model assumes each processing layer is its own representation of what is salient.

4. Miller et al. (1993) observed suppression of response in IT neurons in a matching task that occurs within 10 msec of response onset. They conclude that the source of this suppression must be within or before IT. Chelazzi et al. (1993), in a different matching task for IT neurons, observed a first spike after 60–80 msec; 100–120 msec for full strength; 130–200 msec for full inhibitory attentional effect. Both of these works support a top-down version of attention and recognition. The routing and tagging models are bottom-up: only the attended signals ever reach the top. The tuning model relies on the initial signals to reach the top where they are used to guide further processing.

5. Although until very recently it was generally thought that attentional effects were not seen earlier than in V4 neurons (but see Haenny and Schiller, 1988), Motter (1993) has provided evidence to the contrary. This was predicted in the initial description of the selecting tuning model in Tsotsos (1990). Using an experimental paradigm that involved competing stimuli and directed attention, Motter showed that attentional effects are observed in V1, V2, as well as V4 neurons when targets were presented outside the receptive field of the neuron being

recorded. Distance was an important variable; this is the reason for the apparent difference between these results and those of Moran and Desimone (1985). The effect varies depending on the number of competing stimuli and usually manifested itself as a reduction in response if attention is directed away from the recorded neuron. There was no effect for single stimulus displays. These experiments point to a context-dependent view of attentional processing. The selective tuning model is a top-down model, and such effects arise naturally. The routing and tagging models are bottom-up models and it is not obvious how they may account for these results. The Niebur et al. model exhibits no attentional effects before area V4.

6. Schiller (this volume) presents neurophysiological evidence (which is supported by psychophysical evidence in Braun, 1994) that shows that V4 plays a significant role in the selection of less prominent stimuli from the visual scene and that this role is distinctly different than that of area MT. If V4 is lesioned, this function is destroyed for images where the target is a small item in a field of large ones in an odd-man-out task, but only little impairment is observed when the target is large in a field of small items. An MT lesion does not lead to the same effect. Such an observation is a natural one within the selective tuning model. For the large target-small distractors image the large item dominates responses at the top of the pyramid. It is the winner of the top-level WTA. If the target is small in a field of large distractors, however, the large units are the first winners; the small items would never be found unless the selective tuning is operational due to the characteristics of pyramid computation described previously. If V4 is on the path of the inhibitory beam, and it is lesioned, then the beam cannot operate correctly. In the other models, selection of the winner is made from early representations, and the difference between large and small targets would not be seen in this experiment.

The Olshausen et al. version of selective routing requires spacing of connections between layers to double with each layer; otherwise the model violates connectivity constraints. The Niebur et al. temporal tagging model requires the existence of inhibitory frequency-selective interneurons in V4. The selective tuning model requires the existence of local gating networks in each processing layer. It seems that the experimental verification of each of these points is critical for each of the models.

The models collectively form an interesting account of progress in the development of computational models of visual attention; it is clear that much research, both theoretical and experimental, remains.

Acknowledgments

Sean Culhane provided the examples of figure 20.4 with help from Eyal Shavit. I also thank Sean Culhane, Neal Davis, and Winky Wai for manuscript comments. The author is the CP-Unitel Fellow of the Canadian Institute for Advanced Research. This research was funded by the Information Technology Research Center, one of the Province of Ontario Centers of Excellence, the Institute for Robotics and Intelligent Systems, a Network of Centers of Excellence of the Government of Canada.

References

Allport, A. (1989). Visual attention. In M. Posner (Ed.), *Foundations of Cognitive Science* (pp. 631–682). Cambridge, MA: MIT Press.

Anderson, C., and Van Essen, D. (1987). Shifter circuits: A computational strategy for dynamic aspects of visual processing. *Proc. Natl. Acad. Sci. U.S.A. 84*, 6297–6301.

Bergen, J., and Julesz, B. (1983). Parallel versus serial processing in rapid pattern discrimination. *Nature (London) 303*, 696–698.

Braun, J. (1994). Visual search among items of different salience: Removing visual attention mimics a lesion in extrastriate area V4. *J. Neurosci. 14*, 554–567.

Broadbent, D. (1971). *Decision and Stress*. London: Academic Press.

Burt, P. (1988). Attention mechanisms for vision in a dynamic world. *Proc. Int. Conf. Pattern Recognition*, 977–987.

Chelazzi, L., Miller, E., Duncan, J., and Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature (London) 363*, 345–347.

Colby, C. (1991). The neuroanatomy and neurophysiology of attention *J. Child Neurol. 6*, S90–S118.

Crick, F., and Koch, C. (1990). Towards a neurobiological theory of consciousness. *Semin. Neurosci. 2*, 263–275.

Culhane, S., and Tsotsos, J. K. (1992a). A prototype for data-driven visual attention. *Proc. 11th Int. Conf.* Pattern Recognition, The Hague, August, 36–40.

Culhane, S., and Tsotsos, J. K. (1992b). An attentional prototype for early vision. *Proc. Second Eur. Conf.* Computer Vision, Santa Margherita Ligure, Italy, May, 551–560.

Desimone, R., Wessinger, M., Thomas, L., and Schneider, W. (1990). Attentional control of visual perception: Cortical and subcortical mechanisms. *Cold Spring Harbor Symp. Quant. Biol. LV*, 963–971.

Dickinson, S., Pentland, A., and Rosenfeld, A. (1992). From volumes to views: An approach to 3-D object recognition. *CVGIP: Image Understanding* 55(2), 130–154.

Duncan, J., and Humphreys, G. (1989) Visual search and stimulus similarity, *Psychol. Rev. 96*(3), 433–458.

Feldman, J., and Ballard, D. (1982). Connectionist models and their properties. *Cog. Sci. 6*, 205–254.

Gochin, P., Miller, E., Gross, C., and Gerstein, G. (1991). Functional interactions among neurons in inferior temporal cortex of the awake monkey. *Exp. Brain Res. 84*, 505–516.

Haenny, P., and Schiller, P. (1988). State dependent activity in money visual cortex I. Single cell activity in V1 and V4 on visual tasks. *Exp. Brain Res. 69*, 225–244.

Haenny, P., Maunsell, J., and Schiller, P. (1988). State dependent activity in monkey visual cortex II. Retinal and extraretinal factors in V4. *Exp. Brain Res. 69*, 245–259.

Haykin, S. (1991). *Adaptive Filter Theory*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A. 79*, 2554–2558.

Julesz, B., and Bergen, J. (1983). Textons, the fundamental elements in preattentive vision and perception of textures. *Bell Syst. Tech. J. 62*(6), Part II: 1619–1645.

Karp, R., and Ramachandran, V. (1990). Parallel algorithms for shared-memory machines. In J. van Leeuwen (Ed.), *Handbook of Theoretical Computer Science: Vol. A: Algorithms and Complexity* (pp. 871–941). Cambridge, MA: MIT Press.

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

Kröse, B., and Julesz, B. (1989). The control and speed of shifts of attention. *Vision Res. 29*(11), 1607–1619.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.

Miller, E., Li, L., and Desimone, R. (1993). Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *J. Neurosci. 13*(4), 1460–1478.

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

Motter, B. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2 and V4 in the presence of competing stimuli. *J. Neurophysiol. 70*(3), 909–919.

Nakayama, K. (1991). The iconic bottleneck and the tenuous link between early visual processing and perception. In C. Blakemore (Ed.), *Vision: Coding and Efficiency* (pp. 411–422). Cambridge: Cambridge University Press.

Niebur, E., Koch, C., and Rosin, C. (1993). An oscillation-based model for the neuronal basis of attention. *Vision Res. 33*(18), 2789–2802.

Nothdurft, H.-C. (1993). Saliency effects across dimensions in visual search. *Vision Res. 33*(5/6), 839–844.

Olshausen, B., Anderson, C., and Van Essen, D. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci. 13*(11), 4700, 4719.

Oram, M., and Perrett, D. (1992). Time course of neural responses discriminating different views of the face and head. *J. Neurophysiol. 68*(1), 70–84.

Remington, R, and Pierce, L. (1984). Moving attention: Evidence for time-invariant shifts of visual selective attention. *Percept Psychophys. 35*(4), 393–399.

Saarinen, J., and Julesz, B. (1991). The speed of attentional shifts in the visual field. *Proc. Natl. Acad. Sci. U.S.A. 88*, 1812–1814.

Sagi, D., and Julesz, B. (1985). "Where" and "What" in vision, *Science 228*, 1217–1219.

Tanaka, K., Saito, H., Fukada, Y., and Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J. Neurophysiol. 66*(1), 170–187.

Tsotsos, J. (1988). A 'complexity level' analysis of immediate vision. *Int. J. Comput. Vision 1*(4), 303–320.

Tsotsos, J. K. (1989). The complexity of perceptual search tasks. *Proc. Int. J. Conf. Artificial Intelligence*, Detroit, 1571–1577.

Tsotsos, J. K. (1990). Analyzing vision at the complexity level. *Behav. Brain Sci. 13*(3), 423–469.

Tsotsos, J. K. (1992). Is complexity theory appropriate for analyzing biological systems? *Behav. Brain Sci. 14*(4), 770–773.

Tsotsos, J. K. (1993a). The role of computational complexity in understanding perception. In S. Masin (Ed.), *Foundations of Perceptual Theory* (pp. 261–296). Amsterdam: North-Holland.

Tsotsos, J. K. (1993b). An inhibitory beam for attentional selection. In L. Harris and M. Jenkin (Ed.), *Spatial Vision in Humans and Robots* (pp. 313–331). Cambridge: Cambridge University Press.

Uhr, L. (1972). Layered 'recognition cone' networks that preprocess, classify and describe. *IEEE Transact. Comput. C-21*, 758–768.