ELSEVIER

Artificial Intelligence 78 (1995) 507–545

Artificial Intelligence

Modeling visual attention via selective tuning

John K. Tsotsos, Sean M. Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, Fernando Nuflo

Department of Computer Science, University of Toronto, 6 King's College Rd., Room 283D, Toronto, Ontario, Canada M5S 1A4

Received July 1994; revised January 1995

Abstract

A model for aspects of visual attention based on the concept of selective tuning is presented. It provides for a solution to the problems of selection in an image, information routing through the visual processing hierarchy and task-specific attentional bias. The central thesis is that attention acts to optimize the search procedure inherent in a solution to vision. It does so by selectively tuning the visual processing network which is accomplished by a top-down hierarchy of winner-take-all processes embedded within the visual processing pyramid. Comparisons to other major computational models of attention and to the relevant neurobiology are included in detail throughout the paper. The model has been implemented; several examples of its performance are shown. This model is a hypothesis for primate visual attention, but it also outperforms existing computational solutions for attention in machine vision and is highly appropriate to solving the problem in a robot vision system.

1. Introduction

This paper presents theoretical and computational arguments supporting a model of various aspects of visual attention based on the concept of *selective tuning*. Previously, the concept was named the *inhibitory beam model* and was first presented at the June 1991 Conference on Spatial Vision in Humans and Robots, York University [59]. The goal of the research is to develop a model of visual attention that has both biological plausibility as well as computational utility.

The central thesis of this paper is that attention acts to optimize the search procedure inherent in a solution to vision whether that solution is implemented in the brain or in a computer. This model of attention addresses the reduction of the

SSDI 0004-3702(95)00025-9

number of candidate image subsets and of feature subsets that are considered in matching; it does so by selectively tuning the visual processing network. Computational arguments linking search optimization to attention for vision and the concept of attentive selective tuning first appeared in [55]. Attention operates continuously and automatically: without attention, so-called general-purpose vision is not possible. The theory described in this paper is most closely related to the works of Koch and Ullman [29], Burt [7], Niebur et al. [39] and Olshausen et al. [42].

It is important to situate the model in its appropriate contexts. Not only is this model a hypothesis for primate visual attention, but it also outperforms existing computational solutions for attention in machine vision and is highly appropriate to solving the problem in a robot vision system. Primate vision is an existence proof for the functionality of systems computer vision researchers seek to develop. If it were possible to use the same methods as are found in primate vision embodied in a computational theory then machine vision would be successful. It is clear that the resulting system would not necessarily be the only possible computational vision system with similar functionality; however, it would be one of the solutions. Thus, the work follows in the footsteps of many other well-known works in computer vision where biological inspiration has played an important role. The inspiration is clear in our research; in fact, we go beyond simply being inspired and attempting to build in characteristics of the biology. The goal is to derive from first principles the nature of the attentional mechanism needed by any vision system, whether it be biological or machine. As a result, our past work has focused on laying the theoretical foundation [55–58]. This paper details the model of visual attention that has been developed on this foundation, presents the relationship to the neurobiology of primate visual attention, makes predictions regarding the neurobiology, and demonstrates the computational utility both in terms of theoretical results showing the method superior to past models as well as experimental results demonstrating implementations of the model. Thus, the model has two lifelines along which its success might be measured. The first is dependent on whether the biological predictions can be verified and whether new observations might be explained well by the model. The second is dependent on whether the model is useful in computational solutions of vision.

1.1. The need for attention in vision

As argued in [57], selective attention is one of the important mechanisms for dealing with the combinatorial aspects of search in vision. The visual attention mechanism seems to involve at least the following basic components: (i) the selection of a region of interest in the visual field; (ii) the selection of feature dimensions and values of interest; (iii) the control of information flow through the network of neurons that constitutes the visual system; and (iv) the shifting from one selected region to the next in time. These are discussed in turn below, and later in the paper specific solutions are proposed for some of them. Other aspects

of attention such as the transformation of task information into attentional instructions, integration of successive attentional fixations, interactions with memory and indexing into model bases are not addressed here. Some past work has dealt with some of these issues. For example, Ahuja and Abbot [1] integrate a variety of cue types in an active scheme for surface estimation. Wilkes and Tsotsos [69] discuss how to recognize object models using an active method which requires several fixations from viewpoints determined by the state of interpretation. Wilkes and Tsotsos [68] deal with the problem of indexing in an active object recognition system. However, the general problems are still open.

1.1.1. The need for region of interest selection

In [56], it was proved 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 versus display size is verified psychophysically in a large body of work). If, on the other hand, no explicit target is provided, the task is NP-complete. Thus, it may be concluded that the brain is not solving this general problem [57, 58]. 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. 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 sufficiently massive computational power, the brute-force search strategy will work perfectly well without attention. For both the primate and realizable computational visual systems and natural images and tasks, this brute-force approach fails [57].

1.1.2. The need for features of interest selection

Search within feature space seems to also have an exponential nature [57]. Although the number of feature types seems much smaller than the size of an image, the number of feature values is very large. Suppose that there are a large number of potential models and that a target containing the color red is sought in an input image and that nothing more is known about the target. As a first strategy, it would seem sensible to consider only matching to those models that contain red features. This is a large subset. Suppose now that the target may also contain the color blue (the target contains red or blue or both). Using this simple strategy, all models with blue features are added to the subset; the resulting set is larger. And so on, as feature types and values are added to the image, more and more models are added to this subset. In the extreme, this data-directed model activation strategy might include almost all models. Thus, this is not a good method. The mere presence of a feature type gives little discriminating power for a vision system unless there is an associated restriction on the set of objects or events in the task. This is exactly the situation which led to the conclusion in [53] that presence of a feature is sufficient for popout; the target is known and the feature is a sufficient discriminator. The number of feature value subsets is an

exponential function of the set size, and brute-force search in natural images for feature value subsets which may be the best candidates for matching will not suffice [57]. Recent neurobiology concludes that attentional selection acts in both the feature and spatial domains and does so independently [27].

1.1.3. The problems with information flow

The computational complexity of vision suggests pyramidal processing [3, 7, 55, 60]. Although pyramids solve part of the complexity problem by reducing the size of the representations to be processed, they introduce others: they corrupt the signals flowing through them unless some additional mechanisms are included. Assume an architecture with a hierarchical arrangement of computing units; values represented at each unit are coded by their response strength similar in spirit to other pyramid schemes (say, [7]). Connectivity from layer to layer need not be fixed and each layer (indeed, each unit) may have different connectivity patterns including overlap. There may be more than one output representation; that is, from an initial input layer several subhierarchies may be constructed, each specializing parts of the original input. This kind of configuration is consistent with that described by Van Essen et al. [63] as the starting point for their model. The hierarchy is composed of computing units (which for the remainder of the paper will be referred to as *interpretive units*) which perform processing related directly to the interpretation of their input (e.g., color, edges, motion). Each interpretive unit receives feedforward as well as feedback connections within the pyramid. Each position in a layer may be the site of several interpretive units, each specialized for some type of visual process. In other words, each spatial position within a layer may involve a column of interpretive units. Within a column, each unit is sensitive to a similar portion of the visual field (its receptive field—RF) but may process different modalities of visual information. For the remainder of the paper the examples and discussion, without any loss of generality, will focus on single pyramids composed of a single type of interpretive unit. Sizes of layers and connectivities do not affect the conclusions.

Four information flow problems due to pyramidal processing will be described; the problems arise on the assumption that no direct or indirect information flow control exists in the structure described in the previous paragraph. The first problem is depicted in Fig. 1A: the *Context* effect. A single unit at the top of the pyramid receives input from a very large subpyramid, and thus from a very large portion of the visual field. Unless an object is alone in the visual field, the response of units whose receptive fields contain the object will be affected not only by that object but also by any other image event in that receptive field, and is confounded by the object's context. Surround effects have been observed previously. For example, Van Essen et al. [64] speculate that the large feedback pathways in the cortical hierarchy may be causing this phenomenon. *Blurring* is the second problem with pyramid architectures. A single event at the input will affect an inverted subpyramid of units (Fig. 1B). Thus, although a single event may be well localized at the input layer, it is blurred as it flows upwards so that a large portion of the output layer now represents parts of it. The third problem is

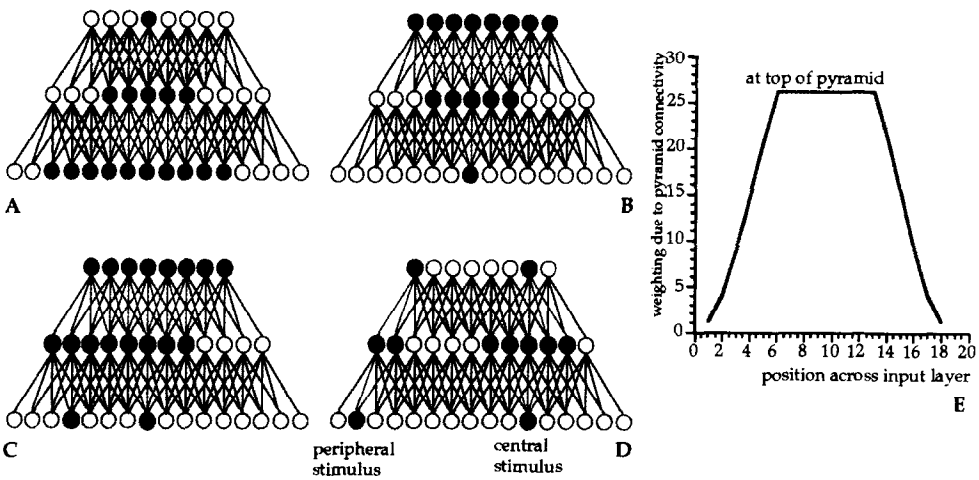

Fig. 1. A. Context effect. B. Blurring effect. C. Cross-talk effect. D. Boundary effect. E. Impulse response of the pyramid. All connections are equally weighted.

Cross-talk (Fig. 1C). Two separate visual events in the visual field will activate two inverted subpyramids which overlap. The region of overlap will contain units whose activity is a function of both events. Thus, each event interferes with the interpretation of other events in the visual field. The final problem is the *Boundary* problem: central items will appear to be stronger at the output layer than items in the visual periphery even if the peripheral items are in fact stronger. This is due solely to the numbers of connections feeding units in successively higher layers in the outer regions of the pyramid. The input layer is thus divided into a central and peripheral region. In Fig. 1D, the cause for the effect is illustrated: the parts of the pyramid affected by two stimuli in the input layer, one being located centrally and the other peripherally, are shown. Note how the number of ascending connections differs in the two cases; the peripheral units will have smaller valued responses than the central ones even if the stimuli are of equal strength. Normalization or local weighting corrections will not solve this. This is even more evident in Fig. 1E. Here, a unit strength single stimulus is swept across the input field and a plot is shown of the maximum value at the output layer produced as a function of its input position.

1.1.4. The need to shift selection in time

Once a region is selected, it then follows that the remainder of the visual field cannot be processed unless a sequence of different regions are selected which together cover the visual field. There is the possibility that many regions might be selected simultaneously and matched in parallel; the behavioral observations however do not support multiple foci of attention (see [16]). There are many choices for how to select next regions to process. An algorithm might simply tile visual space, selecting regions in some arbitrary order that will eventually cover

the entire visual field. Alternatively, an ordering might be imposed on image subregions such that after one is processed it is not processed ever again unless some new image event occurs in that region. Whatever the algorithm, solutions must be found to the following problems: (i) in what order should regions be selected? (ii) when should a previously selected region be re-selected? (iii) if the visual world is time-varying, how are changes in the image contents taken into account in determining the selected regions? Task requirements may help in this determination.

1.1.5. The need to balance task- and data-directed processes

Many of the above arguments point to the need to use task information whenever possible in order to reduce the computational cost of vision. But task information is not always available; many human activities seem mindless or casual as opposed to directed by some specific goal. Thus, one must not ignore data-directed processing and the need to balance the task- and data-directed dimensions of vision processing and of attention specifically. Neither task direction for attention nor the need for balance between the two processing modes is a new idea. For example, Roland [44] proposed the existence of a task-dependent selective attention mechanism which independently of stimulus rates and intensities enhances or inhibits the metabolism in cortical areas in a differential way. Selective attention can influence different processing levels in the visual system possibly reflecting a facilitatory effect on different visual computations or task components. For example, psychophysical sensitivity for discrimination of subtle attribute variations is observed to be enhanced with task guidance [11]. Finally, a review of the neurobiology of attention can be found in [10]. That paper concludes that attentional processes must achieve a balance between data-driven and knowledge-driven processes.

1.2. Other models

Several hypotheses for the computational modeling of biological visual attention have appeared. The connectionist hypotheses are not detailed in this brief review; most include learning models and do not address the relevant neurobiology of attention. There are many computational vision models which include aspects of attention, some of which are briefly overviewed.

1.2.1. The selective routing hypothesis

Several models fall into the *Selective Routing Hypothesis* category. The first is that of Koch and Ullman [29]. The model includes the following elements: (i) an early representation, computed in parallel, permitting separate representations of several stimulus characteristics; (ii) a selective mapping from these representations into a central non-topographic representation which at any instant contains only the properties of a single visual location; (iii) 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, (iv) inhibition of this selected location causes an automatic shift to the next most conspicuous location.

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 [28] with their texton difference idea and further explored by Nothdurft [41] 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. The WTA network implements a parallel computation based on the values on 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 this most salient item and through another $\log_m n$ steps the most salient item reaches the output of the system. A shift of attention thus requires at most $2 \log_m n$ time steps. The WTA will not converge if there are two equally strong items.

The shifter circuit model presented a strategy for information flow in stereopsis, visual attention and motion perception (Anderson and Van Essen [3]). The model enables the re-alignment of successive representations in the processing stream starting in the lateral geniculate nucleus and the input layers of primate visual 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 spatially contiguous it is straightforward to show that this strategy requires an implausibly large number of connections per neuron.

The Olshausen, Anderson and Van Essen [42] model is an elaboration of the shifter circuit idea. The problem described above with the original shifter circuits model is remedied via a clever re-structuring 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. [42] mechanism is to form position- and scale-invariant representations of objects in the visual field at the output layer of the visual processing pyramid. This is accomplished via a set of control

neurons, originating in the pulvinar, that dynamically modify 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. Each node in the processing hierarchy performs a simple linear weighted sum operation. Selected objects in the visual field are found by the Koch–Ullman mechanism using luminance saliency, then routed to the top layer of the pyramid. The selected object is transformed by the routing so that it spans the top-level representation where associative recognition takes place.

1.2.2. 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 [12] 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 semi-synchronous 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 [39] detail a model based on those suggestions.

Niebur et al. assume that salient objects have been selected in the visual field by the Koch–Ullman mechanism. Attentional modulation is added at the level of primary visual cortex V1 and 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 are assumed in V4. These are required to act as bandpass filters selective to spikes arriving every 25 ms or so. Thus, they would pass temporally tagged spike trains and block other non-frequency modulated signals. Both Crick and Koch [12] and Niebur et al. [39] 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.

Niebur and Koch [40] further modified the model to deal with the observation that oscillatory firing of neurons has been difficult to confirm experimentally. Thus, rather than suggest that oscillatory modulation is used for temporal tagging, they have proposed that firing coincidences among V2 neurons are sufficient. There is no evidence available which might favor one proposal over the other. However, Shadlen and Newsome [49] present theoretical arguments proving the existence of neurons that defect fine timing coincidences is doubtful.

1.2.3. Models of attention within computer vision

There are a great many proposals for attentive processing in computer vision, and it is not possible to review all of them here. Clark and Ferrier [9] use a salience measure where at each position in the visual field, a value is associated

that is a function of the response of some feature detector (brightness, color, etc.) and the relative importance of the particular feature to the task being solved. WTA methods then find the strongest of those responses and this becomes the focus of attention. The goal of the work was to guide the overt attention of the Harvard stereo head and not to model covert attentional fixations. Burt [7] has developed an attention mechanism based on a multi-resolution Laplacian image pyramid. A rudimentary fovea is formed within the pyramid. At the lowest frequency level, the foveal region encompasses the whole image and represents the capability of peripheral vision to resolve low resolution patterns over the full field of view. At successive levels, the region in the fovea is half the field of view of the level below it. The overall mechanism has three basic functions: foveation, tracking and prediction of next salient locations. As such, it has some of the characteristics of an overt attentional system.

Color is used as a means of locating matching candidates in work by Swain and Ballard [51], Ennesser and Medioni [20] and Grimson et al. [23], to name a few. In these cases, models of the objects sought in the images are known in advance and the color distribution of the object is used (and in conjunction with position by Ennesser and Medioni [20] and with stereo by Grimson et al. [23]) to filter images for candidates. There seems to be no relationship between this kind of attentional guidance and human behavior. The attentive part of the solution to the "Waldo" hidden pictures game described in [20, 23] appears inherently parallel, whereas humans require painstaking serial search to accomplish the same. Further, in humans it is feature contrast that affects the speed of performance, not feature values themselves, i.e., color histograms [41]. For human vision, although a particular visual task may contain many qualities which we feel are salient, it is not the case that they can all be used by the visual processing system with equal ease for search optimization. In the "Waldo" pictures for example, the complex outlines of the various figures do not appear to be useful saliency cues. Thus, there is a difference between what is perceived salient at the task level and what is usable saliency at the early vision level. Usable saliency should probably be restricted to those image qualities that can be rapidly detected, to those that "pop out" or are the output of the filters present in the early pathways. Thus these systems, although they have utility for computer vision, were not developed as models of the biological attentive system. In fact, by assuming that all salient features are usable, they go beyond what human vision seems capable of and attempt to solve too large a problem in comparison to biology.

2. The selective tuning model of visual attention

Selective tuning takes two forms integrated within a single algorithm: spatial selection is realized by inhibition of irrelevant connections in a pyramid of visual computations; and, feature selection is realized by inhibition of those units which compute non-selected features. The search process which spatially localizes the

image subset to process is as follows. A WTA process operates across the entire visual field at the top layer of the pyramid: it determines the globally most salient (or *winning*) unit in the output layer. This WTA can accept guidance for areas or stimulus qualities to favor if that guidance were available but operates independently otherwise. The search process then proceeds from the top layer to the lower layers. The globally winning unit activates another WTA that operates only over its direct feedforward inputs. This localizes the winning unit within the top-level winning receptive field. All of the feedforward branches of the pyramid that do not contribute to the winner are pruned (that is, the connections are inhibited leaving the units unaffected). 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 attentional beam (see Fig. 2). The WTA does not violate biological connectivity constraints if the top layer is constrained to contain at most a number of interpretive units equal to the lesser of the permitted neuron fan-in and fan-out. Further there is no restriction on the uniqueness or contiguity of winners; a group of equally strong yet non-contiguous units can be identified as most salient. Conflicting biases are dealt with solely within the WTA scheme much like Clark and Ferrier [9] suggest (this is discussed further below).

The process of selection requires two traversals of the pyramid. First, the representations of the interpretive units throughout the pyramid are computed in a feedforward manner. Second, the hierarchy of WTA processes is activated in a top-down manner to detect and localize the strongest item in each layer of representation, pruning parts of the pyramid that do not contribute to the most salient item and continuously propagating changes upwards.

There is similarity between the selective tuning model and the models of Koch and Ullman [29] and of Burt [7]. The selective tuning model includes some of the elements of the Koch–Ullman model as described above; differences will be

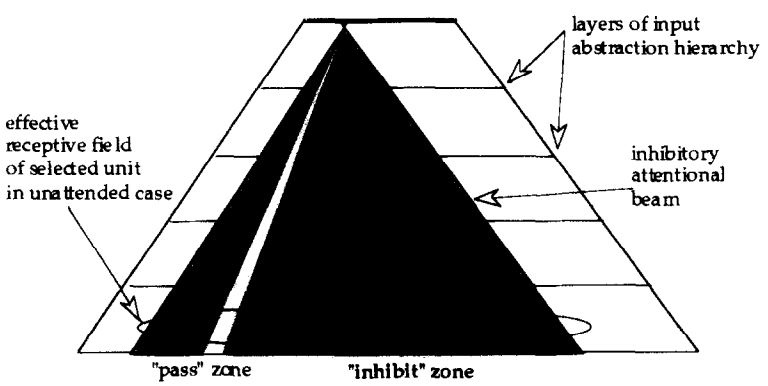

Fig. 2. The inhibitory attentional beam concept operating within a pyramid of visual computations.

highlighted below. Burt's model also includes the notion of top-down, hierarchical pruning within a pyramid structure. However, Burt does not detail exactly how the decisions are made and there is no relationship between his model and the neurobiology of early vision.

2.1. A solution for the selection of spatial region of interest

Koch and Ullman's WTA algorithm is central to all other major computational models of biological visual attention. When it was designed, it seemed consistent with the known timing of attentional shifts [50, 54]; however, this is no longer the case. Remington and Pierce [45] show that distance has no effect on attention shifts; there is no attentional gradient. They further point out a very important constraint: efficient coordination with the saccadic eye movement system in reading or visual search tasks would dictate rapid, time-invariant movements to match saccade dynamics. More recently, Kröse and Julesz [30] found no proximity effect; they show that shifts of attention do not take time proportional to the distance between items but rather are accomplished in constant time and conclude that a parallel scheme is needed to find prospective locations which are then checked by a slow serial process. Further, Koch and Ullman's mechanism does not immediately yield the kinds of attention-related receptive field changes observed in areas such as V4 [35].

2.1.1. A new winner-take-all algorithm

A new WTA updating rule is presented whose properties seem better matched to the current knowledge of the primate visual system. The model requires several different types of computing units arranged in a pyramid. *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 interpretive units in the next feedforward layer of the pyramid. *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. Fig. 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 is termed an assembly.

2.1.2. Form of the WTA computation

The notation to be used below is introduced; Fig. 3 should be used as a supplement. Physical units are distinguished from their value by the use of a hat ("$\hat{}$") 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, if present, represents an identifier

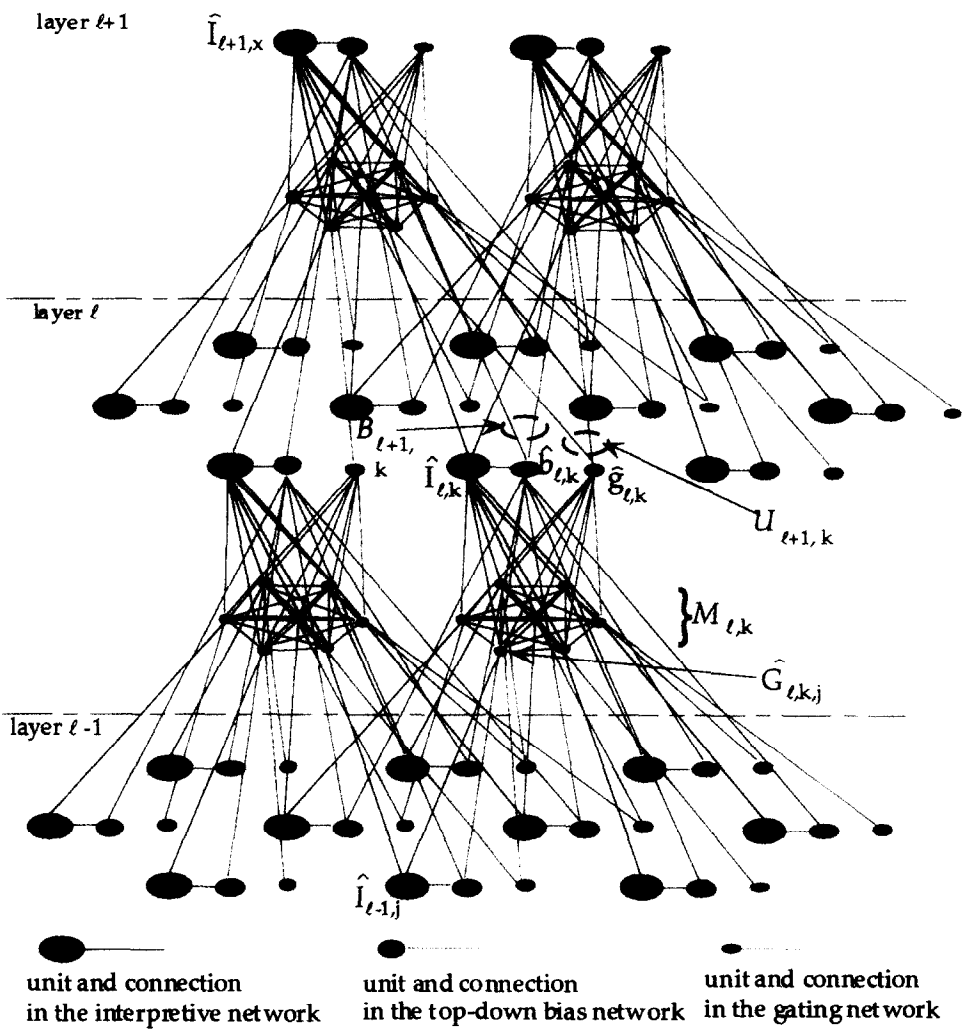

Fig. 3. The detailed wiring within four assemblies spanning three layers in the visual processing hierarchy is shown; see Section 2.1.2.

used to distinguish units within a set. Superscripts refer to time within the iterations of a given WTA process. The input layer is layer 1 and the output layer is layer L. Further:

- $\hat{I}_{l,k}$: the interpretive unit in assembly k in layer l;
- $\hat{G}_{l,k,j}$: the jth WTA gating unit, in assembly k in layer l linking $\hat{I}_{l,k}$ with $\hat{I}_{l-1,j}$;
- $\hat{g}_{l,k}$: the gating control unit for the WTA over the inputs to $\hat{I}_{l,k}$;
- $\hat{b}_{l,k}$: the bias unit for $\hat{I}_{l,k}$;
- $q_{l,j,i}$: the real-valued weight applied to $\hat{I}_{l-1,i}$ in the computation of $\hat{I}_{l,j}$;

- $n_{l,x}$: a scale normalization factor;
- $\mathcal{M}_{l,k}$: the set of gating units for unit $\hat{I}_{l,k}$;
- $\mathcal{U}_{l+1,k}$: the set of gating units in layer $l+1$ making feedback connections to $\hat{g}_{l,k}$;
- $\mathcal{B}_{l+1,k}$: the set of bias units in layer $l+1$ making feedback connections to $\hat{b}_{l,k}$.

A common iterative formulation for a WTA [21] is:

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

where V is the set of units in the competition and 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 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 accomplished using a simple observation: if the inhibitory signal is based on response differences, then an implicit but global ordering of response strengths is imposed on the entire network on the basis of pairwise local information. The largest item will thus not be inhibited at all, 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:

$$\Delta_{i,j} = \begin{cases} q_{l,k,i} G_{l,k,i}^{t-1} - q_{l,k,j} G_{l,k,j}^{t-1}, & \text{if } 0 < \theta < q_{l,k,i} G_{l,k,i}^{t-1} - q_{l,k,j} G_{l,k,j}^{t-1}, \\ 0, & \text{otherwise}. \end{cases} \tag{2}$$

$G_{l,k,j}^t$ is the value of gating unit $\hat{G}_{l,k,j}$ at time t, such that $0 \leqslant G_{l,k,j}^t$. The weighting, $q_{l,j,i} \in \mathbb{R}$, of each input to the interpretive unit $\hat{I}_{l,j}$ is included in order to reflect the importance of each input to the interpretive computation. It would not be necessary if it were the case that all inputs to a given interpretive unit were equally weighted; this is not the case. If the largest valued input to the computation of $\hat{I}_{l,j}$ is weighted negatively in that computation, then it should not be considered as a most salient input within the receptive field of $\hat{I}_{l,j}$. Using the $q_{l,j,i}$ weights in the manner above ensures that the largest value must also be positively weighted; it is the product of value and weight that is important to the computation of contributions in the WTA. θ is a threshold set to

$$\theta = \frac{Z}{2^\gamma + 1} \tag{3}$$

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 (see below). 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). Multiple winners are thus permitted under this definition. The term $w_{i,k}C_i^{t-1}$ in (1) above is replaced by $\Delta_{l,j}$.

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 $0.0 \leq b_{l,k} \leq 1.0$ defined by

$$b_{l,k} = \min_{a \in \mathcal{B}_{l+1,k}} \{a\}. \tag{4}$$

$\mathcal{B}_{l+1,k}$ is the set of bias units in layer $l+1$ making feedback connections to $\hat{b}_{l,k}$. The nature of the bias computation is to inhibit any non-selected units allowing the selected ones to pass through the pyramid without interference. For example, if red items are being sought, the interpretive units which are selective for red stimuli would be unaffected while all other color-selective units would be biased against to some degree. The default value of bias units is 1.0; this value only changes 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. 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:

$$G_{l,k,j}^{t_0} = b_{l-1,j} n_{l-1,j} I_{l-1,j}. \tag{5}$$

t_0 is the time at which a particular WTA competition begins. The normalization factor is included here in order to make results of computations at different scales directly comparable (see [13, 32]). Also note that the computations of (5) are performed 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:

$$g_{l,k} = \begin{cases} 1, & \text{if } \sum_{a \in \mathcal{U}_{l+1,k}} \{a\} > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

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. In this way, the gating units are affected but not the interpretive units; only the relevant connections are closed down allowing the unit to participate in other computations as needed. The value of $\hat{g}_{l,k}$ is zero for all units during the first phase of the process. 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 (6). As the pruning of connections proceeds downwards, new results of interpretive unit computations become available as their inputs are restricted. Time is allowed as shown in Fig. 4 for the complete upwards propagation of the new results. After this upwards propagation, a period of time is provided where the same path through the pyramid cannot be active. This inhibition of the selected region and pathway is a concept borrowed from Koch and Ullman [29]. Inhibition of return is discussed further below.

Due to 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 ms or so (the time between shifts varies with experiment: Sagi and Julesz [46] found some inspection times to be as short as 17 ms; Bergen and Julesz [6] noted 50 ms), 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 using these shift timings. This may be considered as an alternative explanation for the oscillations which motivate the temporal tagging model. This gating signal may be considered as a sort of system clock to use a computational metaphor. However, this time for attentional shifts seems to be controversial. For example, Duncan et al. [19] claim that attentional dwell time in their visual search experiments on humans is on the order of 250 ms. In the above experiments, the tasks differ and

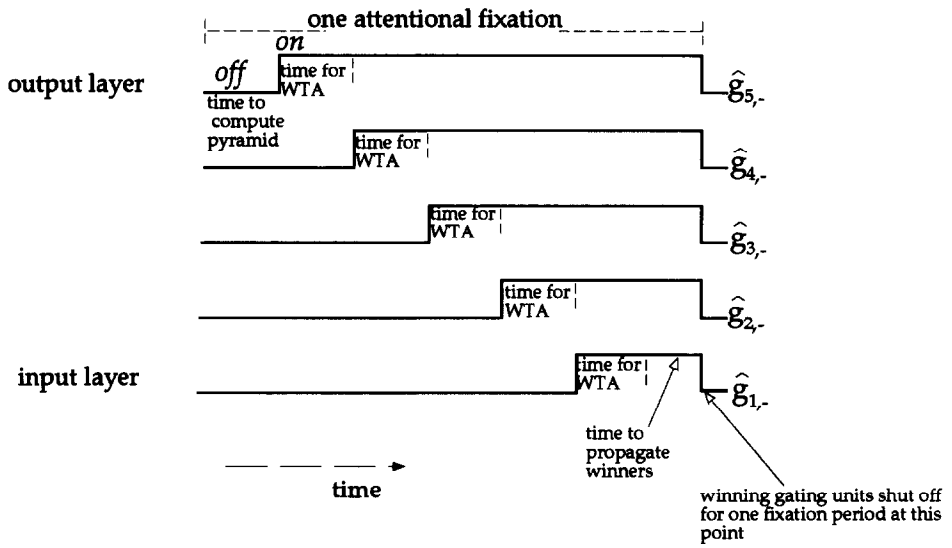

Fig. 4. The gating control signals for each of a number of layers in a pyramid.

have different time requirements. It is probably the case that attention can shift quickly if needed, but can also hold fixation for longer time intervals if the task requires it. In our model this is easily incorporated by permitting the period of the top-level gating signals to be task controlled.

In order to enforce stability and so that no oscillations occur, the overall result is rectified (negative values are set to zero) by passing the entire right side of Eq. (1) through a rectifying function $\boldsymbol{R}$ such that

$$\boldsymbol{R}[x] = \begin{cases} x, & \text{if } x > 0, \\ 0, & \text{otherwise}. \end{cases} \tag{7}$$

Each of the preceding functionalities, including the control signals and the WTA action, are incorporated into a new updating rule given by:

$$G^{t}_{l,k,j} = g_{l,k}\boldsymbol{R}\left[G^{t-1}_{l,k,j} - \left(\sum_{i \in \mathcal{M}_{l,k}; i \neq j} \Delta_{t,i} \right) \right]. \tag{8}$$

Performance issues resulting from the use of this rule are considered next.

2.1.3. Performance issues

Two theorems and their proofs are now given, the first regarding convergence of a single WTA process and the second regarding convergence of a pyramid of WTA processes. An analysis of the convergence properties of the WTA method is also given.

Theorem 1. *The WTA updating rule of* (8) *is guaranteed to converge for all inputs under the definitions presented in Section* 2.1.2.

Proof. Let $\Sigma_{i \in \mathcal{M}_{l,k}; i \neq j} \Delta_{t,i}$ be termed the contribution to a unit. Since the contribution to unit i depends on a difference function, an ordering of units is implicitly imposed depending on their response magnitude. In a given WTA competition, unit j will inhibit unit i only if the value of unit j is larger than that of unit i. Thus, the largest units (a unique maximum is not required) will have a contribution of 0 and will remain unaffected by the iterative process. All other units will have strictly positive contributions and thus will decay in magnitude or remain at zero. The rectifying function guarantees that no unit receives an updated value that is negative and thus oscillations cannot occur. The iterations are terminated when a stable state is reached (no units change in magnitude). It is thus trivially shown that the process is guaranteed to converge and locate the largest items in the input set. □

It is important that the convergence properties be investigated. Although multiple winning units are a feature of the method, in order to simplify the discussion below and without any loss of generality, we assume unique valued units. From the updating function, it is clear that the time to convergence depends only on three values: the value of the largest unit, the magnitude of the second largest unit and the parameter θ. The largest unit is not affected by the updating

process at all. The largest unit is the only unit to inhibit the second largest unit. The contribution term for all other units would be larger than for the second largest because those units would be inhibited by all larger units. This, along with the fact that they are smaller initially, means that they would reach the lower threshold faster than the second largest unit. Convergence is achieved when all units but one decay in value to θ; therefore the time to convergence is determined by the time it takes the second largest element to reach this value. This makes the convergence time independent of the number of units in the WTA process. The amount of inhibition for the first iteration of the updating rule for the second largest unit I_2, where the largest unit is I_1, is given by (8), simplifying for this situation to:

$$I_2^1 = 2I_2^0 - I_1^0 , \tag{9}$$

$I_1 = I_1^0$ is constant. Convergence is achieved when $I_2^t \leq \theta$. At the kth iteration, $I_2^k = 2^k I_2^0 - (2^k - 1)I_1$. Convergence will thus require $\log_2((I_1 - \theta)/(I_1 - I_2^0))$ iterations. There is no dependence on either topographic distance or numbers of competitors, thus providing a much better match to experiments [30, 45]. This relationship clearly shows that the more similar the values of the two items, the slower the convergence (as in [18]).

A bound on this number of iterations is desirable. Arbitrarily small differences between values are not allowed; the differences must be at least θ, so the denominator of the logarithm can be no smaller than θ. I_1 can be no larger than Z. Thus, the upper bound on the number of iterations is given by:

$$\log_2\left(\frac{Z-\theta}{\theta}\right). \tag{10}$$

For example, if $Z = 1000$ and θ is 10% of this maximum value, then the upper bound on the number of iterations is 3.2, or in practice, 4. A lower bound can be found as well; the fastest convergence that can be achieved is when the second largest element is just greater than the threshold θ and this will be denoted by θ^+; the expression becomes:

$$\log_2\left(\frac{I_1-\theta}{I_1-\theta^+}\right) = \log_2(1^+) = 0^+ , \tag{11}$$

where the superscript "+" means "just larger than". Since iterations must be performed in their entirety before decisions are made, the lower bound in practice is 1 iteration.

If convergence is required within γ iterations, then equating γ to the bound of (10) gives the appropriate value of θ that will guarantee the convergence:

$$\theta = \frac{Z}{2^\gamma + 1} . \tag{12}$$

This tacitly assumes that $I_1 > \theta$. I_1 may not be large enough in some situations; moreover a large theta may not be sensible given that it is a variance threshold for

responses caused by the same physical stimulus. A "gain" parameter will solve this problem. Redefine the contribution to include a gain parameter A, $A \geq 1$, so that:

$$G^t_{l,k,j} = g_{l,k}\boldsymbol{R}\left[G^{t-1}_{l,k,j} - A\left(\sum_{i \in \mathcal{M}_{l,k}; i \neq j} \Delta_{i,j} \right) \right]. \tag{13}$$

In this case, at the kth iteration, $I_2^k = (1 + A)^k I_2^0 - ((1 + A)^k - 1)I_1$. Convergence will require $\log_{1+A}((I_1 - \theta)/(I_1 - I_2^0))$ iterations. Using the same argument as above, if convergence is required within γ iterations, then the following expression gives the appropriate value of θ that will guarantee the convergence:

$$\theta = \frac{Z}{(1 + A)^\gamma + 1}. \tag{14}$$

In general, if the allowable variance is known and the maximum number of permissible iterations is given, then the gain may be set as:

$$A = \left(\frac{Z - \theta}{\theta} \right)^{1/\gamma} - 1. \tag{15}$$

It now remains to determine what the WTA network is guaranteed to find in a pyramid of such processes.

Theorem 2. *The WTA algorithm is guaranteed to find a path through a pyramid of L layers such that it includes the largest-valued interpretive node in the output layer (m_L) and interpretive nodes m_k, $1 \leq k < L$, such that m_k is the largest-valued node within the support set of m_{k-1} and where m_1 must be within the central region of the input layer.*

Proof. Each interpretive node of the pyramid provides a measure of fit to some visual event or feature. The set of units providing feedforward connections is the support set. It is known that in the input layer there is an annulus on the boundary in which the computations have undefined value since there is insufficient support data. This effect is compounded in successive layers of the pyramid. The annulus of the input layer for which the following theorem does not hold has width defined by

$$d_1 - 1 + \sum_{i=2}^{L-1} \frac{d_i - 1}{2}.$$

where d_i is an odd integer and is the diameter of the RF at layer i.

The proof is by induction on the number of layers of the pyramid. Theorem 1 proved that for a single layer the WTA is guaranteed to converge and to find the maximum-valued elements in the WTA set. This is true regardless of the type of computation which derived those values. The same single layer guarantees hold throughout the pyramid.

There is an important reason for the restriction on the node which the method

is guaranteed to find. Recall the discussion of the context and boundary effects of Section 1.1. The absolute maximum value in the input layer is confounded by these two characteristics of pyramid processing and would not be preserved under any conditions. Thus, the WTA can only find the maximal values within a layer and not the maximum values which are inputs to those layers. This seems to be exactly what is required, however, if the unit operations are measures of fit for particular features or events in the input.

Assume that for a pyramid of n layers the theorem is true. By the induction principle, if it can be proved that the theorem holds for $n+1$ layers (where the $(n+1)$th layer is added to the pyramid on the input layer side), the proof is complete for a pyramid with an arbitrary number of layers. Suppose that the beam path includes node m_n in the nth layer. The WTA process rooted at node m_n is guaranteed by Theorem 1 to find the largest-valued node in the added layer in the support set of node m_n and include it in the beam path. □

The new WTA greatly improves the signal-to-noise ratio of the visual process. If the connections from the unselected items are inhibited leaving the selected unit connections intact it is clear that signal-to-noise improvement is dramatic. This bears resemblance to the results reported by Bashinski and Bacharach [5] who found that events can be reported at a lower threshold with attention.

Finally, a rough comparison of this WTA algorithm with the Koch–Ullman algorithm and a provably optimal one shows that this WTA not only can outperform the Koch–Ullman method, but also achieves an efficiency that approaches that of the provably optimal scheme for finding the maximum value of a given set [62] in a biologically plausible manner. This optimal time complexity for finding a maximum in a set of n elements using a set of p parallel processors is

$$\log_2 \log_2 n - \log_2 \log_2\left(\frac{2p}{n}\right) \quad \text{for } \frac{n}{2} \leqslant p \leqslant \frac{n(n-1)}{2} \tag{16}$$

within some integer constant. The basic operation is a comparison between two elements resulting in a decision of which is the greater. The algorithm works equally well whether there is a unique or multiple maxima in the input set. Valiant [62] points out that optimal complexity is only possible assuming a rather large amount of overhead computation per time step. This overhead grows as $\log_2 p$, faster than the expression above and thus dominates asymptotically. It is reasonable to assume that the overhead can be captured in constant time if p is fixed. Also note that the number of two-element comparisons required to completely determine the ordering of the set is $\frac{1}{2}n(n-1)$; if there were this many processors available, the result can be found in one time step.

It must be stressed that comparing the Koch–Ullman WTA, the WTA of this paper and the optimal max-finding procedure is not possible in a direct fashion. Each has a different definition of the amount of computation that must be performed within a time interval and there are different multiplicative and additive constants involved. The latter algorithms require significant pre-processing as well; the hierarchy within which the WTA operates must first be defined.

Finally, the time complexities are stated in different ways: Valiant's is an asymptotic worst-case complexity while the other two are upper bounds. Nevertheless, a few interesting conclusions can be drawn from the comparison.

The Koch–Ullman algorithm requires $2\log_m n$ operations to determine the globally most salient item. The selective tuning WTA has an upper bound on the number of time steps of $\log_2((Z-\theta)/\theta)$ from (10). θ can be defined in terms of Z, i.e., if θ is some fraction ξ of the value Z, this expression reduces to $\log_2((1-\xi)/\xi)$ and is independent of the number of elements as well as the maximum value of those elements. Within a competition, it is assumed that some processing is associated with each connection in the WTA network equivalent to the computation of Eq. (2). Thus n^2-n comparisons are performed per iteration; this will not violate connectivity constraints as described earlier. In Valiant's terms, $p=n^2-n$. Finally, this expression gives the number of iterations per layer of the pyramid; if the pyramid has L layers, then the overall number of iterations is given by

$$L\log_2\left(\frac{1-\xi}{\xi}\right). \tag{17}$$

Koch and Ullman suggest a biologically plausible network of six layers and an optic nerve size number of elements in a saliency map ($n=1{,}000{,}000$). This yields 12 time steps for convergence. The Valiant algorithm for this size of input set requires fewer than 4.3 time steps (as long as the number of processors is in the range $\frac{1}{2}n \leq p \leq \frac{1}{2}n(n-1)$). In the selective tuning WTA, for a six-layer network, the upper bound on convergence time is lower than that of Koch and Ullman for all values of $\xi > 0.2$, a reasonable decision threshold. The selective tuning WTA algorithm is faster than the Valiant scheme for values of $\xi > 0.38$, again not an unreasonable decision threshold (although the noise tolerance properties will suffer). Note that each method has different processor demands and this accounts for part of the conclusions.

2.2. *Implementations and performance examples*

Good experimental results using real and simulated images have been achieved using an implementation of the model. Although the model is inherently a parallel one requiring large numbers of processors and connections, the implementation is essentially serial. In all cases, at least two feature qualities (scale plus one or more feature types) are included in the saliency representation. A simple method for arbitration among feature types is used, namely, another WTA across feature dimensions. The implementation is integrated to control the attentional performance of the TRISH stereo head [33].

Two types of images were used, real 8-bit gray scale digitized images and binary simulated images. In each case the structure of the pyramid was different, primarily because the nature of the visual information computed differed. These details will appear with the examples; it should be clear that the choices of

pyramid sizes and other structural parameters are quite arbitrary and do not in any way impact the conclusions.

2.2.1. Scale normalization

If units represent more than one stimulus quality, say position, size, luminance, wavelength, or edge contrast, then the competition must consider the interactions among dimensions as well as the absolute magnitudes of response. For example, suppose the definition of saliency is of the following form: the most salient visual event is the one which is the brightest over the largest region of visual field. Ambiguities will arise in the competition because units may have the same response yet differ in size (as well as location).

A simple method is used to resolve this specific ambiguity and no behavioral significance is claimed for this solution [13]. If a unit with a small RF has a response of ψ, and a larger competing unit has a RF with response $(\psi - \varepsilon)$, then for a sufficiently small ε, we would like the larger RF to win over the smaller one. To illustrate, consider a RF of size 3×3 that has a response of 255, and a competing RF of size 30×30 that has a response of 254. Since the latter unit represents a visual event that is 100 times the size of the event that the former unit represents, and is over 99% its strength, it seems reasonable to favor the latter over the former.

This bias on RF size is accomplished by multiplying the responses of all units by a normalizing factor that is a function of the size of the corresponding RF. A normalization whose rate of change is greatest for small RFs, without weighting very large RFs excessively is desired. In the experiments conducted, the following empirically satisfying function is used:

$$F(x) = \frac{\alpha + 1}{\alpha + \beta^{-\sqrt{x}}}, \tag{18}$$

where x represents the number of basic elements (pixels) in the receptive field. The number 1 in the numerator is a result of normalizing $F(x)$ for $x = 0$. The $\sqrt{x}$ is used to account for the area of the RF. This function may be used for linear features by replacing $\sqrt{x}$ with x. $n_{l,x}$ is the variable used in the formulation of the WTA process to denote scale normalization. Note in (5) that both task bias and scale normalization contribute to the determination of the strength of a particular interpretive unit result. The interaction is strictly multiplicative. It was found empirically that values of $\alpha = 10$ and $\beta = 1.03$ in (18) generally give good results in most instances. A detailed analytical account of this issue appears in [32]; our empirical function asymptotically approaches the optimal presented by Lindeberg [32].

2.2.2. Luminance

In this example, salient items are those which are the brightest and largest regions in the image [13]. The lowest level of the processing hierarchy is the digitized image, and each successive level is a simple local average of the previous

level. The image has size 256×256 pixels. The pyramid has five layers, including the input layer. Rectangular receptive fields of a range of sizes were represented with all combinations of sizes ranging from 6 to 50 pixels on a side. The layers of the pyramid were of sizes (beginning with the input layer and proceeding upwards): 256×256; 208×208; 144×144; 80×80; and 32×32. There is no particular significance to these numbers; changes in sizes and indeed in numbers of layers do not affect overall behavior of the algorithm. Thus at each location of any layer, there are a large number of competing units receiving input from different size receptive fields. Each unit computes the average of its input. The result for the first attentional fixation in an intensity image of a toy boat is shown in Fig. 5A while the scan path of fixations for the first few covert fixations is shown in Fig. 5B.

2.2.3. Oriented edges

In the next example, a 128×128 pixel 8-bit gray scale image of a hand was used as the input and the definition of salient items was changed. The most salient item is the longest and highest contrast straight line [14]. Difference operators were used to extract the edges from the input image; this choice was based on simplicity and ease of computation. Four orientations are computed by applying the appropriate difference template (0, 45, 90 and 135 degrees). Each orientation is preserved by creating separate pyramids for each orientation, while still maintaining a single overall beam. Each pyramid has five layers. Alternatively, this may be considered as a single pyramid where at each location there is a column of units, each unit representing a different orientation. The range of receptive fields was 3×3 to 35×35. The sizes of the layers were 128×128, 108×108, 80×80, 48×48 and 28×28. Within each orientation hierarchy, a WTA process chooses the winning RF like in the intensity simulations. Then the winners from each of these separate WTA competitions are input into an additional WTA process that determines a single overall winner from among the individual hierarchy winners. This overall winning RF determines which regions of the next level of all orientation hierarchies are to compete. Fig. 5C illustrates the scan path showing the movement of the pass zone on the input layer for successive fixations.

2.2.4. Instantaneous optical flow patterns

In this experiment, simulated instantaneous full velocity optic flow patterns were used. The goal was not to fully interpret the motion (i.e., extract motion parameters); it was thought that this could be a fast method of localizing and labeling salient motion patterns. Once localized, they can be examined in more detail for motion parameters [31].

The images were small, 64×64 pixels, and the pyramid used four layers. Templates were constructed to match against each of 16 types of motion pattern. It should be clear that there is no claim that this set of motions is complete; it is, however, a representative set of simple and complex motions. Goodness-of-fit measures for how well a template explains a given subset of optic flow in the

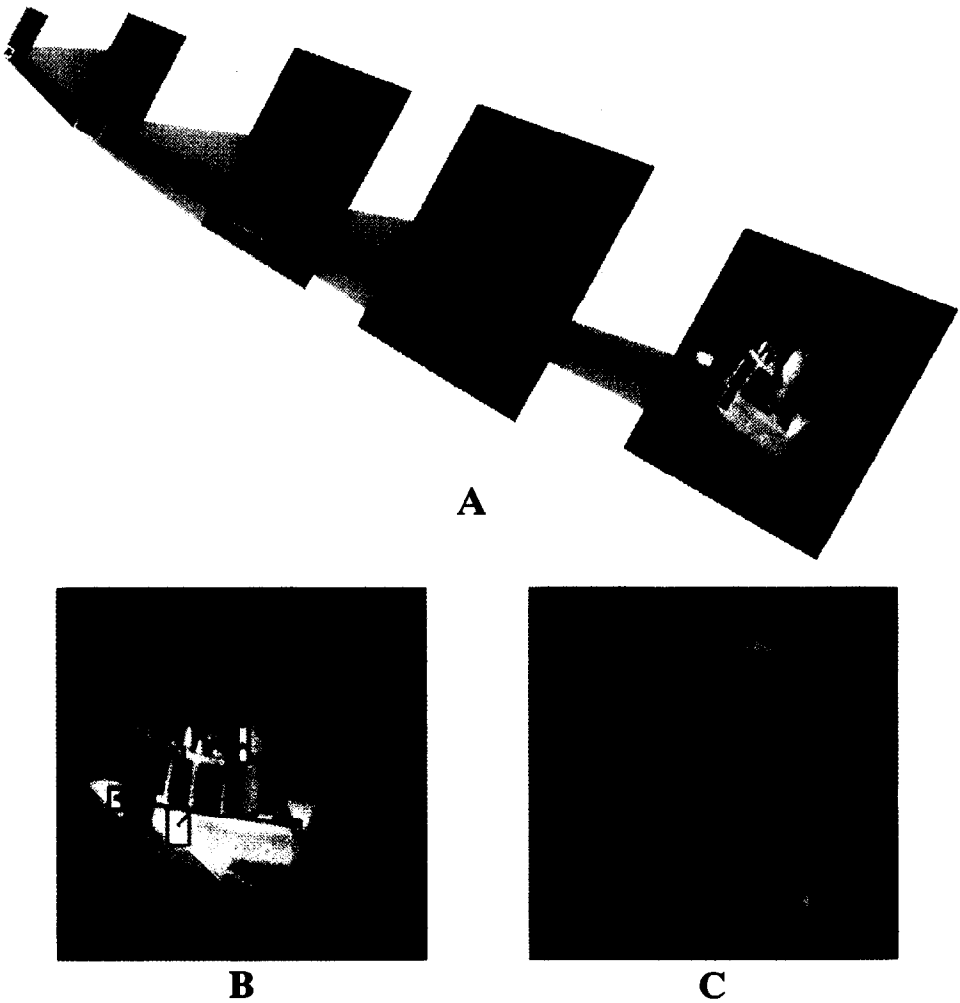

Fig. 5. The first selected region (luminance and scale) of the image is outlined in yellow in each layer of the pyramid. Dark red highlights the beam's pass zone, while the red shaded portions show the inhibitory zone (not cumulative). B. Luminance salience—the first five fixations. C. Edge salience—the first six fixations

image were computed using straightforward correlation and then the remainder of the pyramid was constructed using local averages. The patterns fall into two categories: motion of the environment (full field motions) and motion of objects in the visual field. The types of flow patterns (color-coded in Fig. 6) in the first category are: translate (Env.Trans.), clockwise rotate (Env.C.Rotate), counter-clockwise rate (Env.AC.Rotate), clockwise rotate off-center (Rotate.EM.Cue), recede off-center (Recede.EM.Cue), approach (Env.Approach) and recede

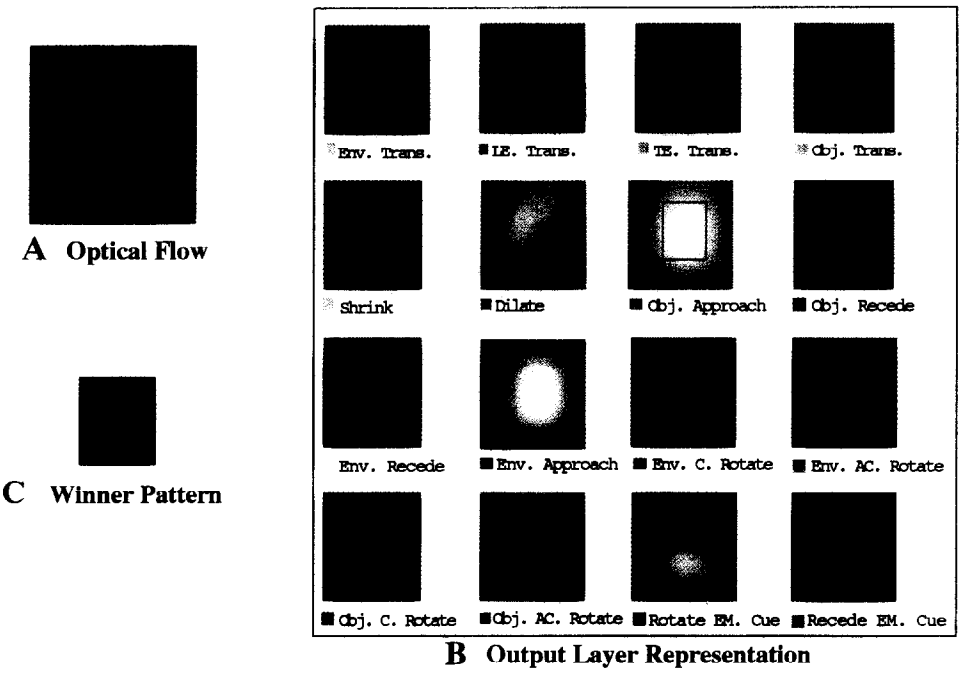

Fig. 6. Optic flow example: a single object approaching an observer. The overall map is all orange (object approach); there are no competing stimuli and the approach is blurred over the entire output. It is strongest only in the region corresponding to the object.

(Env.Recede). In each case, the environment is exhibiting this flow which is induced by self-motion. Clockwise rotate off-center is a rotation of the environment about a pre-determined point that is not the center of the image, and recede off-center is self-motion away from objects in the environment with focus of expansion off the image center by a pre-determined amount. The types in the second category are: leading edge translate (LE.Trans.), trailing edge translate (TE.Trans.), object translate (Obj.Trans.), shrink (Shrink), dilate (Dilate), approach (Obj.Approach), recede (Obj.Recede), clockwise rotate (Obj.C.Ro-

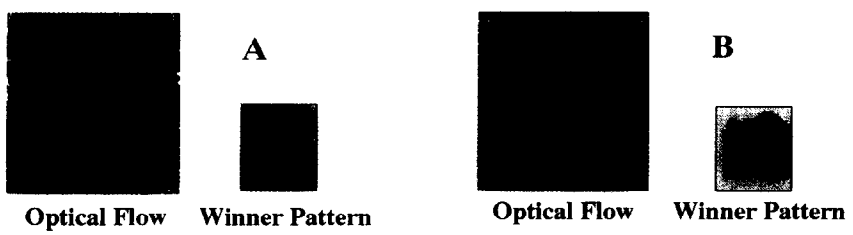

Fig. 7. Two more optic flow examples: A. A noisy field. B. An object approaching in a translating environment.

tate), and counter-clockwise rotate (Obj.AC.Rotate). In each case, subregions of the image (hypothesized objects) are exhibiting this flow pattern. Test images were created by selecting some combination of motion, selecting specific motion parameters (size of object, velocity, etc.) and generating the appropriate flow field. Successful tests were run on noise-free images as well as on noisy images and using images with multiple, differently moving objects. Similar to the edge example earlier, separate pyramids were maintained, one for each motion pattern, and an overall winner was selected.

For display purposes, different colors are assigned to each of the flow patterns. The integrated result of applying the WTA selection scheme over all the motion patterns is represented by a color map with the same size as the top layer. For each position in the color map, the color is that of the winning pattern which has the strongest response over all the patterns within RFs centered at that position. The color composition of the color map reflects the cause of the input optical flow. The largest and strongest response, determined by running yet another WTA on this map similar to the luminance example above, is the overall winner.

The model can correctly label and locate the most salient pattern when the input is the optical flow of one of the stored single flow patterns. Fig. 6A shows the input pattern for a single object in motion, approaching the camera. The representations at the output layer of each of the 16 pyramids are shown in Fig. 6B, where black represents no response and white represents maximal response. Small colored squares beneath each of the gray scale representations give the color code for that type of motion. The best response is shown by the overall winner pattern color (Fig. 6C) and the red square in the object approach representation corresponding to the object in motion.

When 35% noise is added in the same input optical flow (35% of the image is randomly set to some error value), the model can still correctly classify the motion pattern (Fig. 7A). Another example is that of an object in motion with the background exhibiting some other motion (Fig. 7B). The winner map clearly shows an object approaching with a background in translation. There are several small error patterns at the boundaries, a region where all motion algorithms seem to have difficulty. If multiple objects are moving in different (single) motion patterns, patterns are localized and labeled in order of salience sequentially.

It is not claimed that this is all there is to motion processing; rather, this simple scheme appears to be sufficient to detect, localize and label regions where salient motion is occurring so that further analysis may consider only that sub-image for detailed inspection. The templates are consistent with stimuli found effective for neurons in motion areas (such as MST) in monkey (e.g., [17]).

Note that the extension of this particular method and examples to the case of texture fields is straightforward.

2.2.5. Task direction

The following set of examples show the effect of simple task tuning on the attentional process. Although most nontrivial tasks would be more involved than

shown in these examples, in combination, the different types of task guidance demonstrated would have significant utility.

Feature direction

Three separate forms of task guidance are shown: for the boat image, bias against a subregion of the image ($b_{l,k} = 0.0$ for units whose receptive fields are within that region); for the hand image, bias against horizontal lines ($b_{l,k} = 0.0$ for the horizontal line operators); and finally, for the hand image bias against 45-degree lines by 30% ($b_{l,k} = 0.7$ for the 45-degree line operators). The resulting scan paths are shown in Figs. 8A, 8B and 8C respectively. The algorithm

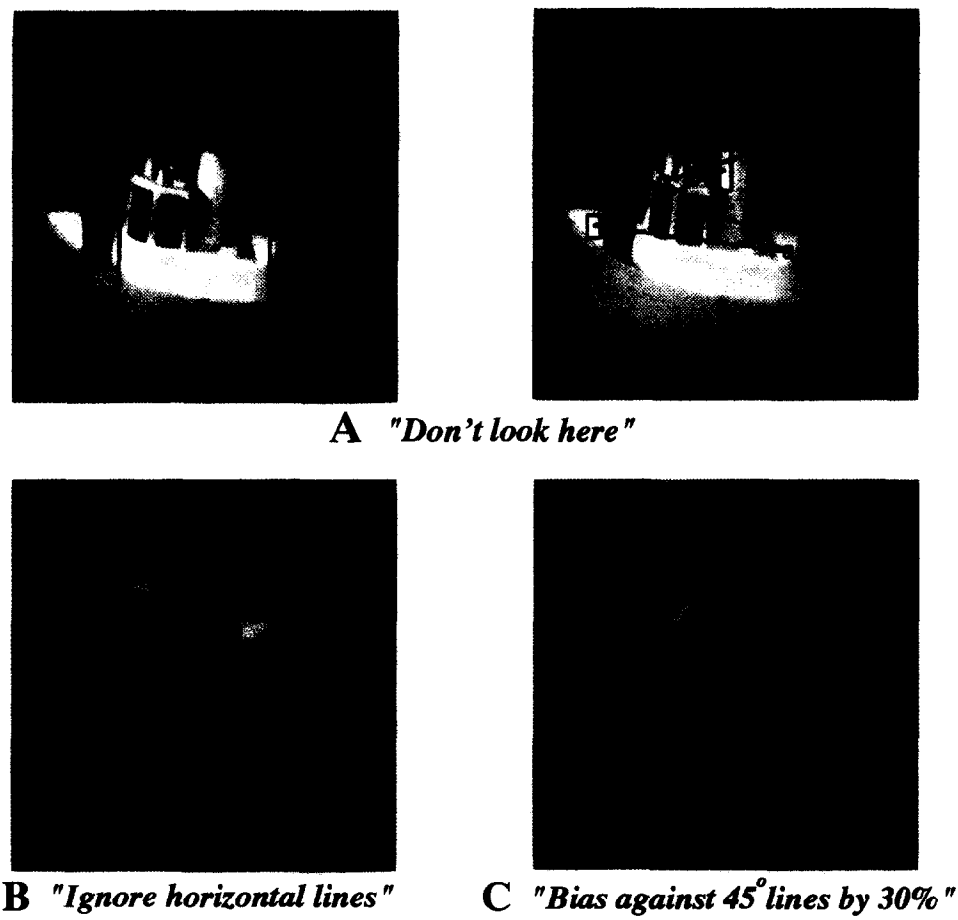

Fig. 8. A. The system is instructed to ignore the region enclosed in blue. The resulting scan path is shown; compare with that of Fig. 5B. B. The second example of task guidance is to ignore edges of a given orientation (horizontal) for the hand image. The scan path is shown; compare with that of Fig. 5C. C. 45-Degree lines are biased against. The scan path should be compared with Fig. 5C. One of the 45-degree lines (the first fixation) is so strong that a 30% bias has little effect.

performs as expected. In order to clearly see the difference in scan paths, the corresponding no-task case must be compared to the task-directed versions. Fig. 8A must be compared to Fig. 5B. The blue box is the region to be ignored and the algorithm skips over the first salient region found in Fig. 5B. Fig. 8B must be compared to Fig. 5C. Horizontal lines are completely ignored. Finally, Fig. 8C is also compared to Fig. 5C as well as to Fig. 8B to see that not all of the 45-degree lines are ignored. The first line found is sufficiently strong to overcome the bias.

It is important not to draw too strong a conclusion from these experiments. The images are of a limited form (no color, no depth, no motion), and the interpretive process as well as task guidance is simple. Nevertheless, the system behavior has all of the right qualities and the results are encouraging for more sophisticated task guidance.

Location cues

It is well known that a location cue can be very effective in speeding up recognition if a target is actually present at that location; and if the target is not present there, the wrong cue can slow down the recognition, both comparisons to the uncued situations (see [2], for review). Thus, it is important that the model include a method for pointing the inhibitory beam at a particular location. If the beam is already set up before the stimulus appears, the time savings are clear: the attended node at the output layer has as input only the stimulus at the cued location with any other stimuli in its receptive field inhibited. Thus, the time to compute the WTAs for each layer is avoided. If the target is not there, then all of this computation must proceed as with no cue, but the time to check the cued location has already been incurred. Thus, if the cue is wrong, the time for recognition will be longer than if no cue is given.

How does the attentional system know where the cue is? When a cue is presented visually, it must be attended even if not fixated. Knowing that it is a location cue instructs the subject to simply not allow attention to shift even after the cue a disappears. The large assumption is that a subject has such voluntary control over that aspect of the attentional mechanism. Other algorithms [42, 47] deal with location as well. Yet, in each case, their system is given positional coordinates in a retinotopic reference frame in order to identify the region where attention is to be centered. It seems highly unlikely that any such coordinates are being passed around in the visual cortex. Rather, it is much more plausible that attention is directed to locations which are referenced by the external world itself (as Ballard [4] suggests for animate vision in general). In the case of all experimental paradigms which employ a location cue, subjects are given the cues visually as a brief flash or a sustained visual marker, and thus the location is given with respect to the external world.

Abrupt onset and offset events

A well-known attention capture mechanism is a flashed cue [70]. Psychophysically, the observations include: (i) In visual search tasks with a target among distractors: when the target is an abrupt onset, no response time (RT) versus

display size effect is observed; otherwise, serial search is observed. (ii) Attentional capture can be overridden by top-down control if enough time is provided for set-up and the cue is good (such as a location cue). In these cases, an abrupt onset does not interrupt performance. (iii) If there are multiple onsets, all are tagged as high priority for attention but only for first 100 ms or so; then they exhibit the same priority as non-onset items. (iv) Onset in any representation has the same effect (luminance, depth, motion and texture have been tested).

A simple algorithm for detecting and localizing abrupt image events at different spatial and temporal scales is apparent [66]: (1) convolve images in a sequence with on-center and off-center difference of Gaussians (DOGs) at several spatial scales; (2) compute temporal differences over several scales; (3) if there is sufficient change, signal an event (on or off); (4) normalize responses for scale; (5) choose strongest response via WTA. Sufficient change means: For onset events, a region of a given scale exhibits a sufficiently large increase in strength over some period of time, and the change occurs within the center region of the DOG operator. For offset events, a region of a given scale exhibits a decrease in strength over some period of time, and the change occurs within the center region of the DOG operator. There is no need for constraint on the new response for the off events: if an object disappears, the new contrast at that position is zero or very small. For a given location and scale,

$$\text{if } \frac{N(t_2)-N(t_1)}{t_2-t_1} > \theta_1 \text{ and } |N(t_2)| > \theta_2 \text{ and } \frac{N_c(t_2)-N_c(t_1)}{t_2-t_1} > \theta_3 ,$$

$$\text{then signal "on"} ; \tag{19}$$

$$\text{if } \frac{F(t_2)-F(t_1)}{t_2-t_1} > \theta_4 \text{ and } \frac{F_c(t_2)-F_c(t_1)}{t_2-t_1} > \theta_5 ,$$

$$\text{then signal "off"} ; \tag{20}$$

where $N(t)$ is the response of an "on" unit at a given point at time t, $F(t)$ is the response of an "off" unit at a given point at time t, $N_c(t)$ and $F_c(t)$ are the responses of the center portions of the receptive fields only, and thresholds are set at percentages of the maximum responses of the relevant DOG operator for the class of images investigated (see [66]).

The largest responses within a given scale are found first via WTA and then across scale in order to find the globally most salient onset and offset. This has been implemented and tested on real gray scale image sequences of blocks on a black background using a luminance representation only, but at multiple spatial and temporal scales. Spatial and temporal scale preferences are easily incorporated.

Fig. 9 depicts an example in which changes are caused by motions of objects. The on events for this example are the moving of the two rectangular blocks and the cylindrical block to their new positions. The off events are the disappearance of the triangular block and three other blocks moving away from their original positions. Note that not all pixels at the new position of the cylindrical block are classified as having an on event. This is because the new position of the cylindrical

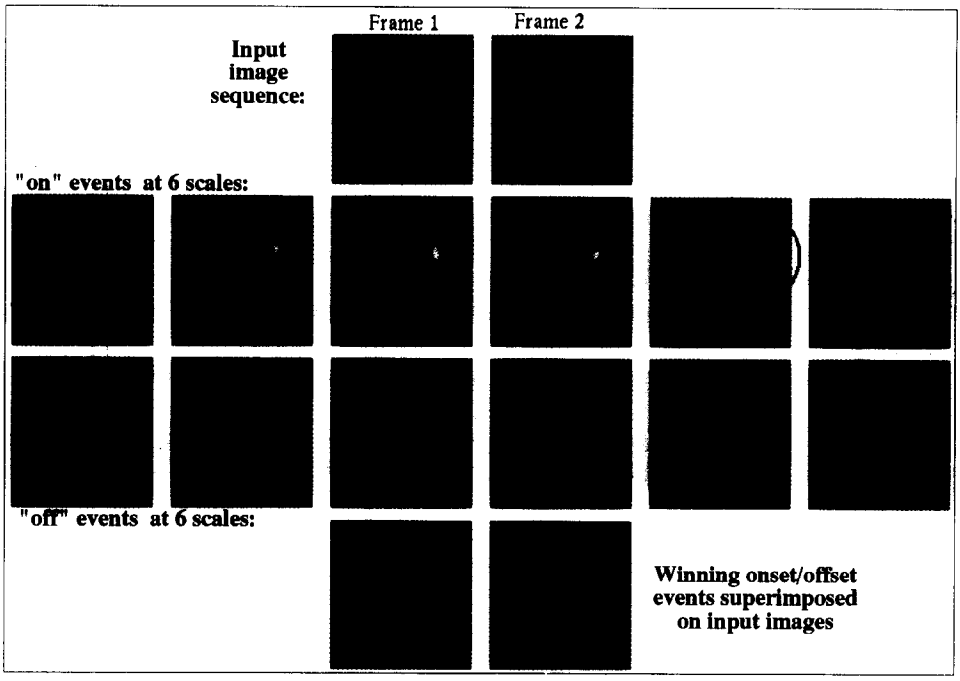

Fig. 9. An example of several blocks moving among other stationary blocks. The most prominent on event is at the new position of the rectangular block at the upper right (denoted by the red circle); the winning off event is at the old position of the cylindrical block (denoted by the yellow block). Winning center and surround regions of the DOGs are shown within each scale.

block partly overlaps with the former position of the rectangular block. Since the intensity levels of the two objects are roughly the same, there is no change in response at the overlap part. The same effect is observed when detecting off events. We can also observe from this example that different events are detected from operators of different sizes.

The algorithm works equally well for images where only luminance changes are present, such as the spotlight from a flashlight which moves around a fixed scene. Simulations using a temporal window other than 2 were also tried with the expected results. For example, suppose the luminance of blocks is decreased gradually and events are detected over a longer time course instead of a shorter one. As the amount of light is decreased continuously for the whole scene, we can expect that there will be off events only. Additional examples can be found in [65].

2.3. The boundary problem and foveating saccades

The boundary problem requires a unique solution and analysis [15]. Although previous solutions have concentrated on extending the edges of the input images in a variety of ways [61], a different and more biologically plausible solution exists

if it is recognized that the eyes which capture the image can move. Suppose that instead of artificially altering the input layer size and contents, an independent mechanism is provided which could detect peripheral salient items separately and saccade to fixate them. To detect the most salient item in the input layer peripheral annulus (described and quantified in Section 2.1.3), an independent WTA whose inputs are only the biased-against units of the input layer is executed. Then a simple algorithm emerges: (1) Compute both the overall most salient item (call this the central winning item) and the independent annulus salient item. (2) If they differ, compare the values of the two after compensating for the processing which occurs through the pyramid; otherwise attend the common item. (3) If the annulus item is more salient than the central item, move the eye to fixate the annulus winner; otherwise the central item is the one attended.

This is not a biologically implausible solution; foveating saccades have been previously described which appear to have similar function [26, 67]. These saccades are elicited in response to a visual stimulus in the periphery (differences most apparent with 10 degrees eccentricity or more) where the exact location of the stimulus is unpredictable. The saccade results in the approximate foveation of the stimulus. Hallett [26] and Whittaker and Cummings [67] hypothesized that a separate mechanism must be present to drive these special saccades.

The realization of the algorithm is mostly straightforward. There are two issues that are worth some discussion. The first is how to determine the amount of compensation in step (2) above. The boundary problem arises because peripheral stimuli are weighted less through the pyramid than central ones. The impulse response for a hypothetical pyramid shown in Fig. 1E demonstrates this. The figure also gives the solution to the compensation issue. Once a peripheral stimulus has been localized by the independent WTA, its position relative to the impulse response can be easily found. The value of that curve at that position gives the relative weighting through the pyramid for that position. The compensating factor then is the maximum weighting in the pyramid (that for a central item) divided by the weighting at the selected position. This method will apply for pyramids where weights are applied in a linear fashion and for peripheral stimuli that are small in spatial extent. For nonlinear pyramids, a more complex scheme is needed which may involve a family of impulse response curves indexed by the value of the peripheral stimulus. For stimuli with large spatial extent, the weighting will be different depending on position within the stimulus. In this case, different compensations, computed in the manner described for each position, will be applied across the stimulus.

The second issue deals with inhibition of return. In the original Koch–Ullman formulation an inhibitory step was included once an item was attended so that attention may shift to the next most salient item. All models seem to have reached the same conclusion on this point. If however the covert system is linked with the overt system a new dimension is added to the inhibition. Not only must locations within an image be inhibited but also locations outside the image. When the eyes move to attend to a peripheral item, previously attended items may not

be present in the new image. Subsequent movement of the eyes may bring those previously attended locations back into view: should they be attended again? In fact, there are many cases where, if no action is taken, the eyes can oscillate perpetually between two or three locations. There is little guidance from behavioral experiments on this point and it is probably the case that task requirements and some kind of internal spatial working memory of what has been seen must play a role.

Fig. 10A shows a typical blocks world real image with the camera view outlined in green; the foveating saccade mechanism is turned off. The first fixation (based on luminance and scale saliency only as described in Section 2.2.2) is correctly on the central block, and successive fixations move around the scene without any camera motion. They do so incorrectly; the second most salient object in this case happens to be the last one found in this sequence. When the independent foveating saccade mechanism is turned on, the second most salient item is detected by the independent WTA in the periphery and it competes with the winner found by the central process. The peripheral item wins, a new camera fixation location is chosen (the centroid of the winning item), the camera moves acquiring the new image (the green rectangle moves in the second fixation of Fig. 10B), and attention is now fixated on the winning item which is centered in the

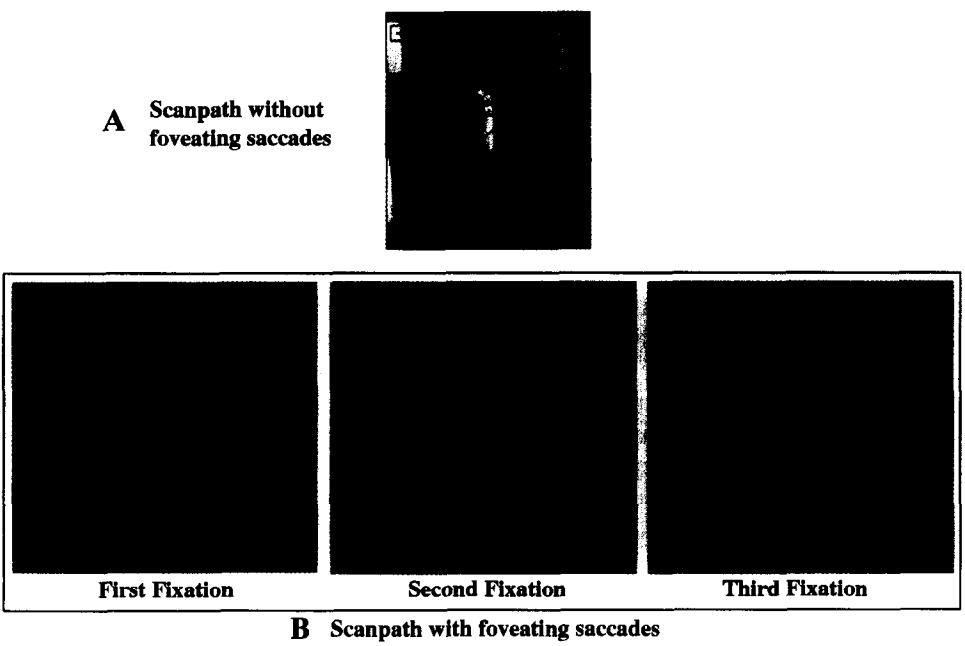

Fig. 10. A. A sequence of fixations without the foveating saccades process. The blue square outlines the cumulative unbiased region. B. The sequence of fixations in the same scene with the foveating saccade scheme. The green box outlines the extent of the visual field and the blue box the extent of the central unbiased region.

new image. The third fixation is now on an item not seen in the first image; note that the first item fixated (in the first fixation) is inhibited and not re-attended even though it is the strongest in the visual field. If oscillations between attended locations are to be prevented, an inhibition based on the object rather than on image coordinates is required (object-based inhibition of return as Gibson and Egeth [22] describe). The current implementation includes such a crude spatial map with temporally-decaying inhibition of attended locations. This demonstrates an important heretofore unexplored computational connection between covert and overt attentional fixations. Foveating saccades not only lead to the solution to one of the information flow problems, but also can play an important role in the exploration of the visual world.

3. Relationship to neurobiology of attention and other models

This section provides detailed discussions driven by the results of several key experimental works in visual attention. Any model of attention with claims on biological plausibility must be able to explain these findings. In all cases, it is assumed that the neurons examined experimentally correspond to the interpretive units of the selective tuning model.

Moran and Desimone [35] discovered that single neurons in trained monkeys as early as in area V4 (but not in V1) can be tuned so that separate stimuli within the same receptive field can be individually attended in a dynamic and task-specific manner. They claim that unwanted information is filtered from the receptive fields of neurons in extrastriate cortex as a result of selective attention on either stimulus location and/or stimulus quality almost as if the receptive field has contracted around the attended stimulus. The attenuation was quite pronounced in V4, somewhat smaller in IT, and not found in V1.

The experiments and observations described in that paper arising from the setup of their figures 1A and 1B are now addressed. In the figures below, five separate set-ups are shown A four-level hierarchy is used; for convenience only, call the layers V1, V2, V4 and IT. Each unit is connected to 7 other units in the pyramid—the exact choice does not matter. Again for simplicity, assume that effective (green) stimuli are effective regardless where they appear in a unit's receptive field. The configurations of input stimuli correspond to experiments in the Moran and Desimone paper.

In each figure, the tick marks below the input layer denote the extent of the receptive field for the neuron being recorded, marked by the arrow. In Fig. 11A, no attentional cue is provided and there is only a single stimulus. This is the situation when one is searching for cells that respond strongly to stimuli and is mapping out their receptive fields before conducting the attentional experiment. Note that even if the input layer contains a single stimulus, the units activated within the pass zone of the beam in the higher layers have a width larger than one unit. This is due to the property of the WTA that finds groups of units whose response is within the error tolerance and labels the group as winner. This

redundancy is not a feature of any other model and may play important roles in ensuring that processing is redundant and thus noise and fault-tolerant. It is not known which units Moran and Desimone actually recorded in relationship to the hierarchical structure. One particular V4 unit was chosen in these examples because it permits the stimuli to be arranged at roughly equal distances both inside and outside the RF as required later on.

Both an effective and an ineffective stimulus are within the chosen RF as shown in Fig. 11B and it is clear that the chosen neuron responds well. This is the performance when there is no task as well as when the effective item is attended. The connections from the ineffective stimulus to the neuron being recorded are inhibited. In Fig. 11C the ineffective stimulus is attended within the chosen neuron and as in the experiments, the chosen neuron does not fire well. This was the major surprise in the experiments of Moran and Desimone. Even though the effective stimulus was still within the neuron's receptive field, it did not cause the neuron to fire. In the model, the connections from the effective stimulus to the neuron recorded are inhibited; however, other V4 neurons do receive input and if the recording probe were moved would find good responses outside the beam structure.

The ineffective stimulus is moved outside the RF for Fig. 11D so that the distance between effective and ineffective stimuli remains the same. When the effective stimulus is attended, the selected neuron fires well. Finally, the ineffective stimulus placed outside the RF is now attended as shown in Fig. 11E. The selected neuron still fires well.

This explanation seems to fit Moran and Desimone's observations nicely and does not deviate from them. However, in order to experimentally verify this explanation properly, an entire pathway must be tested; that is, the entire route from V1 through to IT must be recorded simultaneously, including its breadth across each area—under the kinds of conditions Moran and Desimone use. Further, the distance between effective and ineffective stimuli must be varied since it is clear that a large enough distance might overcome any interactions between subpyramids activated by individual stimuli. Moran and Desimone do not sufficiently detail the experiment in terms of the relationships among the levels, the spread of the areas activated, and the units recorded in order to rule out this explanation.

The conclusion that can be drawn from this illustration using the Moran and Desimone experimental setup is that distance between attended stimulus and receptive field being studied matters for this model. If the attended stimulus is near but not in the receptive field studied, the inhibitory effect of attention on the recorded neuron should be large. If it is far, the effect should disappear, and in between the inhibitory effect of attention will gradually decrease with increasing distance. This should be clear from the figures above. This corresponds very closely to the kind of activity observed in visual-movement neurons in the frontal eye fields when tested with visual search tasks that include distractors [48]. Schall and Hanes [48] found that neural activity peaked when the target was in the response field and was suppressed when the target was beside but not distant from

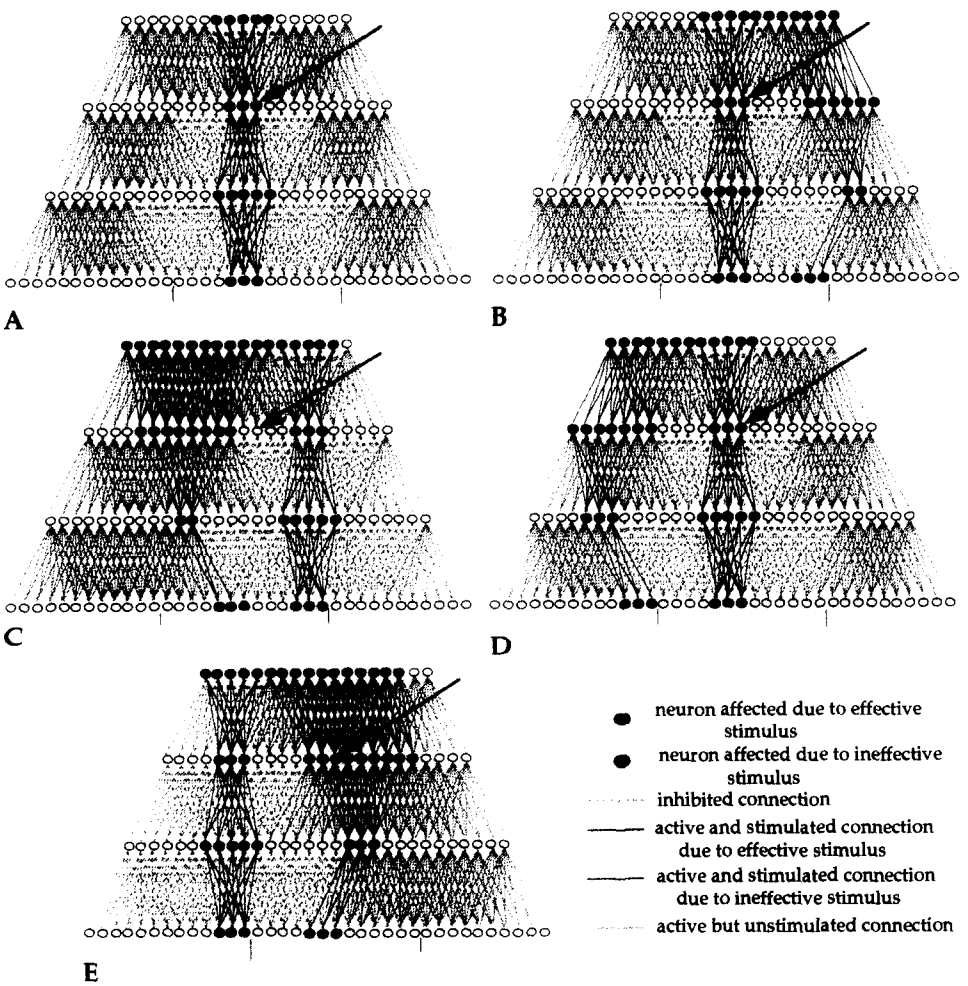

Fig. 11. These figures show the information routing within a hypothetical visual pyramid for situations corresponding to the experiments in Moran and Desimone [35].

this field. The magnitude of these effects are also affected by the position of the neuron within the hierarchy. Schall and Hanes hypothesized that this might be due to a lateral inhibition mechanism; the selective tuning model is an alternative explanation. There is insufficient information to accomplish such a task-specific inhibition if only lateral connections are considered.

Motter [37] concluded that the topographic representation of the neural activity in area V4 highlights potential candidates for matching to targets while minimizing the impact of any background items. In other words, the computations which create this representation seem to maximize signal-to-noise ratios for the features which are relevant to the task. Neural activity was attenuated when the stimulus did not match a cue, independent of spatial location, but was about twice as large

as the attenuated value if the stimulus and cue did match. He used color and luminance as features. Interestingly, he found that neural activity was not affected due to the cueing conditions prior to presentation of stimulus arrays. This is consistent with a model which de-emphasizes connections which are not of interest. In [38] he goes one step further and concludes that the attentional control system seems to be able to "shut down" the synaptic impact of all but one of many color inputs. This too is consistent with the selective tuning model and was suggested by Tsotsos [57] as an important search optimization. Finally, Motter suggests that a sequential combination of the two processes of a full field pre-attentive focal attentive selection based on features which identifies candidate targets, followed by a spatially restrictive focal attentive process which localizes targets, would be an interesting explanation of both his and Moran and Desimone's results; this is exactly the concept initially sketched out by Tsotsos [57] and embodied in the selective tuning model presented here.

The routing, temporal tagging and selective tuning models have much in common in terms of their performance. For example, each of the models offers a believable explanation for the Moran–Desimone [35] observations. Each can provide accounts of a variety of human visual search experiments in that several search processes can be simulated. However, a number of important open questions remain which may help to differentiate the models from one another.

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. [52]) as opposed to pixel-like retinal image copies. Spatial abstraction is a major contributor to the reduction of computational complexity [57]. Note that the image preservation of Olshausen et al. also makes no real improvement in signal-to-noise ratio of the computation; their model in fact preserves the noise.

Miller et al. [34] observed suppression of response in IT neurons in a matching task which occurs within 10 ms of response onset. They conclude that the source of this suppression must be within or before IT. Chelazzi et al. [8] in a different matching task for IT neurons observed a first spike after 60–80 ms, 100–120 ms for full strength and 130–200 ms 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.

Although until very recently, it was generally thought that attentional effects were not seen earlier than in V4 neurons (but see [24]), Motter has provided evidence to the contrary [36]. This was predicted in the initial description of the selective tuning model in [57]. It should not be surprising that attention and task requirements might affect very early levels of visual processing. In audition, it has been found that attentive processes can modify the responses of even the earliest of sensory cells [25, 43]. 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 [35]. 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.

4. Conclusions

A model based on the concept of selective tuning has been presented as an explanation for aspects of visual attention. The overriding goal is to provide a computational explanation to primate visual attentional performance; yet, the computational utility of the resulting method for robot vision is evident. It provides for a solution to the problems of selection in an image, information routing through the visual processing hierarchy and task-specific attentional bias. There are several key characteristics which distinguish the model from its major competitors:

- The timing and convergence characteristics of the new WTA provide a much better match to the behavioral observations than previous WTAs, the WTA is near-optimal in its convergence properties.
- The model makes strong predictions regarding the micro-circuitry of visual cortex and visual search performance.
- The model includes a first link to the eye movement system via the foveating saccade mechanism. A link to eye movements is a characteristic which must be present in any attention model but does not appear in the other major models.
- The model has been implemented and is used for solving real attentional problems in controlling a robotic vision system.

Overall, the match to the neurobiology of attention is very good; although the other models also demonstrate good matches in different ways. Further experimentation that is guided by this and other computational models is required to help differentiate the models with respect to biological plausibility.

Acknowledgments

The first author is 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 and Natural Sciences and Engineering Research Council of Canada.

References

[1] N. Ahuja and L. Abbot, Active stereo: integrating disparity, vergence, focus, aperture and calibration for surface estimation, *IEEE Trans. Pattern Anal. Mach. Intell.* **15** (10) (1993) 1007–1029.

[2] A. Allport, Visual attention, in: M. Posner, ed., *Foundations of Cognitive Science* (MIT Press/Bradford Books, Cambridge, MA, 1989) 631–682.

[3] C. Anderson and D. Van Essen, Shifter circuits: a computational strategy for dynamic aspects of visual processing, *Proc. Nat. Acad. Sci. USA* **84** (1987) 6297–6301.

[4] D. Ballard, Aninate vision, *Artif. Intell.* **48** (1991) 57–86.

[5] H. Bashinski and V. Bacharach, Enhancement of perceptual sensitivity as the result of selectively attending to spatial locations, *Perception Psychophys.* **28** (1980) 241–248.

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

[7] P. Burt, Attention mechanisms for vision in a dynamic world, in: *Proceedings Ninth International Conference on Pattern Recognition*, Beijing, China (1988) 977–987.

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

[9] J.J. Clark and N. Ferrier, Modal control of an attentive vision system, in: *Proceedings ICCV*, Tarpon Springs, FL (1988) 514–523.

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

[11] M. Corbetta, F. Miezin, G. Schulman and S. Petersen, Selective attention modulates extrastriate visual regions in humans during visual feature discrimination and recognition, in: *Exploring Brain Functional Anatomy with Positron Tomography Wiley*, Ciba Foundation Symposium 163 (1991).

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

[13] S. Culhane and J.K. Tsotsos, An attentional prototype for early vision, in: *Proceedings 2nd European Conference on Computer Vision*, Santa Margherita Ligure, Italy (1992) 551–560.

[14] S. Culhane and J.K. Tsotsos, A prototype for data-driven visual attention, in: *Proceedings 11th International Conference on Pattern Recognition*, The Hague (1992) 36–40.

[15] N. Davis, A model for a foveating saccade mechanism, M.Sc. Thesis, Department of Computer Science, University of Toronto, Ont. (to appear).

[16] R. Desimone and J. Duncan, Neural mechanisms of selective attention, in: *Annual Review of Neuroscience* (Annual Reviews, Palo Alto, CA, to appear).

[17] C. Duffy and R. Wurtz, Sensitivity of MST neurons to optic flow stimuli II: mechanisms of response selectivity revealed by small-field stimuli, *J. Neurophysiol.* **65** (6) (1991) 1346–1359.

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

[19] J. Duncan, R. Ward and K. Shapiro, Direct measurement of attentional dwell time in human vision, *Nature* **369** (6478) (1994) 313–315.

[20] F. Ennesser and G. Medioni, Finding Waldo, or focus of attention using local color information, in: *Proceedings Conference on Computer Vision and Pattern Recognition* (1993) 711–712.

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

[22] B. Gibson and H. Egeth, Inhibition of return to object-based and environment-based locations, *Perception Psychophys.* **55** (3) (1994) 323–339.

[23] W.E.L. Grimson, A. Lakshmi Ratan, P. O'Donnell and G. Klanderman, An active visual attention system to play 'Where's Waldo', in: *Proceedings Conference on Computer Vision and Pattern Recognition*, Seattle, WA (1994) 85–90.
[24] P. Haenny and P. Schiller, State dependent activity in money visual cortex I: single cell activity in V1 and V4 on visual tasks, *Experimental Brain Res.* **69** (1988) 225–244.
[25] E. Hafter, Focusing attention on specific auditory filters, in: *Proceedings Annual Meeting of the International Society for Psychophysics* (1991) 39–43.
[26] P. Hallett, Primary and secondary saccades to goals defined by instructions, *Vision Res.* **18** (1978) 1279–1296.
[27] G. Humphreys, C. Romani, A. Olson, M. Riddoch and J. Duncan, Non-spatial extinction following lesions of the parietal lobe in humans, *Nature* **372** (6504) (1994) 357–359.
[28] B. Julesz and J. Bergen, Textons, the fundamental elements in preattentive vision and perception of textures: Part II, *Bell Syst. Tech. J.* **62** (6) (1993) 1619–1645.
[29] C. Koch and S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, *Human Neurobiol.* **4** (1985) 219–227.
[30] B. Kröse and B. Julesz, The control and speed of shifts of attention, *Vision Res.* **29** (11) (1989) 1607–1619.
[31] Y. Lai, Experiments with motion grouping, M.Sc. Thesis, Department of Computer Science, University of Toronto, Ont. (1992).
[32] T. Lindeberg, Discrete scale-space theory and the scale-space primal sketch, Ph.D. Thesis, CVAP, Royal Institute of Technology, Stockholm (1991).
[33] E. Milios, M. Jenkin and J.K. Tsotsos, Design and performance of TRISH, a binocular robot head with torsional eye movements, *Int. J. Pattern Recogn, Artif. Intell.* **7** (1) (1993) 51–68.
[34] E. Miller, L. Li and R. Desimone, Activity of neurons in anterior inferior temporal cortex during a short-term memory task, *J. Neurosci.* **13** (4) (1993) 1460–1478.
[35] J. Moran and R. Desimone, Selective attention gates visual processing in the extrastriate cortex, *Science* **229** (1985) 782–784.
[36] B. Motter, Focal attention produces spatially selective processing in visual cortical areas V1, V2 and V4 in the presence of competing stimuli, *J. Neurophysiol.* **70** (3) (1993) 909–919.
[37] B. Motter, Neural correlates of attentive selection of color or luminance in extrastriate area V4, *J. Neurosci.* **14** (4) (1994) 2178–2189.
[38] B. Motter, Neural correlates of feature selective memory and pop-out in extrastriate area V4, *J. Neurosci.* **14** (4) (1994) 2190–2199.
[39] E. Niebur, C. Koch and C. Rosin, An oscillation-based model for the neuronal basis of attention, *Vision Res.* **33** (18) (1993) 2789–2802.
[40] E. Niebur and C. Koch, A model for the neuronal implementation of selective attention based on temporal correlation among neurons, *J. Comput. Neurosci.* **1** (1994) 151–158.
[41] H.-C. Nothdurft, Saliency effects across dimensions in visual search, *Vision Res.* **33** (5/6) (1993) 839–844.
[42] B. Olshausen, C. Anderson and D. Van Essen, A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information, *J. Neurosci.* **13** (11) (1993) 4700–4719.
[43] S. Parker and B. Schneider, The stimulus range effect: Evidence for top-down control of sensory intensity, *Perception Psychophys.* **65** (1) (1994) 1–11.
[44] P. Roland, Cortical regulation of selective attention in man. A regional bicerebral blood flow study, *J. Neurophysiol.* **48** (5) (1982) 1059–1078.
[45] R. Remington and L. Pierce, Moving attention: Evidence for time-invariant shifts of visual selective attention, *Perception Psychophys.* **35** (4) (1984) 393–399.
[46] D. Sagi and B. Julesz, "Where" and "What" in vision, *Science* **228** (1985) 1217–1219.
[47] P. Sandon, Simulating visual attention, *J. Cogn. Neurosci.* **2** (3) (1989) 213–231.
[48] J. Schall and D. Hanes, Neural basis of saccade target selection in frontal eye field during visual search, *Nature* **366** (1993) 467–469.
[49] M. Shadlen and W. Newsome, Noise, neural codes and cortical organization, in: Hudspeth and Stryker, eds., *Current Opinion in Neurobiology* **4** (4) (1994).

[50] G. Shulman, R. Remington and J. McLean, Moving attention through visual space, *J. Exp. Psychol. Human Perception* **5** (1979) 522–526.
[51] M. Swain and D. Ballard, Color indexing, *Int. J. Comput. Vision* **7** (1) (1991) 11–32.
[52] K. Tanaka, H. Saito, Y. Fukada and M. Moriya, Coding visual images of objects in the inferotemporal cortex of the macaque monkey, *J. Neurophysiol.* **66** (1) (1991) 170–187.
[53] A. Treisman, Features and objects: the fourteenth Bartlett memorial lecture, *Q. J. Experimental Psychol.* **40A** (2) (1988) 201–237.
[54] Y. Tsal, Movement of attention across the visual field, *J. Exp. Psychol. Human Perception* **9** (1983) 523–530.
[55] J.K. Tsotsos, A complexity level analysis of vision, in: *Proceedings International Conference on Computer Vision: Human and Machine Vision Workship*, London, England (1987).
[56] J.K. Tsotsos, The complexity of perceptual search tasks, in: *Proceedings IJCAI-89*, Detroit, MI (1989) 1571–1577.
[57] J.K. Tsotsos, Analyzing vision at the complexity level, *Behav. Brain Sci.* **13** (3) (1990) 423–469.
[58] J.K. Tsotsos, The role of computational complexity in understanding perception, in: S. Masin, ed., *Foundations of Perceptual Theory* (North-Holland, Amsterdam, 1993) 261–296.
[59] J.K. Tsotsos, An inhibitory beam for attentional selection, in: Harris and Jenkin, eds., *Spatial Vision in Humans and Robots* (Cambridge University Press, Cambridge, 1993) 313–331.
[60] L. Uhr, Layered recognition cone networks that preprocess, classify and describe, *IEEE Trans. Comput.* **21** (1972) 758–768.
[61] G. van der Wal and P. Burt, A VLSI pyramid chip for multiresolution image analysis, *Int. J. Comput. Vision* **8** (3) (1992) 177–190.
[62] L. Valiant, Parallelism in comparison problems, *SIAM J. Comput.* **4** (3) (1975) 348–355.
[63] D. Van Essen, C. Anderson and D. Felleman, Information processing in the primate visual system and integrated systems perspective, *Science* **255** (5043) (1992) 419–422.
[64] D. Van Essen, D. Felleman, E. DeYoe and J. Knierin, Probing the primate visual cortex: pathways and perspectives, in: A. Valberg and B. Lee, eds., *Advances in Understanding Visual Processes* (Plenum, New York, 1991).
[65] W. Wai, A model for the detection of image change, M.Sc. Thesis, Department of Computer Science, University of Toronto, Ont. (1994).
[66] W. Wai and J.K. Tsotsos, Directing attention to onset and offset of image events for eye-head movement control, in: *Proceedings IAPR Conference on Pattern Recognition*, Jerusalem (1994).
[67] S. Whittaker and R. Cummings, Foveating saccades, *Vision Res.* **30** (9) (1990) 1363–1366.
[68] D. Wilkes and J. Tsotsos, Efficient serial associative memory, in: *Proceedings Conference on Computer Vision and Pattern Recognition* (1993) 701–702.
[69] D. Wilkes and J. Tsotsos, Integration of camera motion behaviours for active object recognition, in: *Proceedings IAPR Workshop: Visual Behaviors*, Seattle, WA (1994).
[70] S. Yantis and J. Jonides, Abrupt visual onsets and selective attention: evidence from visual search, *J. Exp. Psychol. Human Perception Performance* **10** (1984) 601–621.