available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/brainres

ELSEVIER

BRAIN RESEARCH

Research Report

The different stages of visual recognition need different attentional binding strategies

John K. Tsotsos, Antonio J. Rodríguez-Sánchez, Albert L. Rothenstein, Eugene Simine*

Department of Computer Science and Engineering, York University, 4700 Keele Street, Toronto, Ont., Canada M3J 1P3
Centre for Vision Research, York University, Toronto, Ontario, Canada

ARTICLE INFO

Article history:
Accepted 12 May 2008

Keywords:
Attention
Vision
Feature binding
Recognition
Recurrent process
Selective Tuning

ABSTRACT

Many think that visual attention needs an executive to allocate resources. Although the cortex exhibits substantial plasticity, dynamic allocation of neurons seems outside its capability. Suppose instead that the visual processing architecture is fixed, but can be 'tuned' dynamically to task requirements: the only remaining resource that can be allocated is time. How can this fixed, yet tunable, structure be used over periods of time longer than one feed-forward pass? With the goal of developing a computational theory and model of vision and attention that has both biological predictive power as well as utility for computer vision, this paper proposes that by using multiple passes of the visual processing hierarchy, both bottom-up and top-down, and using task information to tune the processing prior to each pass, we can explain the different recognition behaviors that human vision exhibits. By examining in detail the basic computational infrastructure provided by the Selective Tuning model and using its functionality, four different binding processes – Convergence Binding and Partial, Full and Iterative Recurrence Binding – are introduced and tied to specific recognition tasks and their time course. The key is a provable method to trace neural activations through multiple representations from higher order levels of the visual processing network down to the early levels.

1. Introduction[1]

We address the relationships among attention, recognition and feature binding in vision, relationships that remain poorly understood both theoretically and experimentally. The current common wisdom underlying much research into visual processing by the human brain (and the bulk of computational vision whether biologically inspired or not) includes the following assumptions: that determination of the focus of attention precedes recognition; that the location of the region of interest is used to route stimulus details to higher levels for further analysis/perception; and that it is possible to study vision in isolation from action and task. These seem rooted in the old pre-attentive/attentive distinction, a useful dichotomy in its time, but much less so now. All these concepts are common throughout most of computational vision.

The nature of attentional influence has been debated for a long time. Among the more interesting observations is that of James (1890) "everyone knows what attention is" juxtaposed with that of Pillsbury (1908) "attention is in disarray" and Sutherland's (1998) "after many thousands of experiments, we know only marginally more about attention than about the

* *Corresponding author*. Fax: +1 416 736 5872.
E-mail address: tsotsos@cse.yorku.ca (J.K. Tsotsos).
URL: http://www.cse.yorku.ca/~tsotsos (J.K. Tsotsos).

[1] An earlier, less detailed, version of this paper appeared in Tsotsos et al. (2007).

0006-8993/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.brainres.2008.05.038

interior of a black hole". Even Marr, basically discounted the importance of attention by not considering the time intervals of perception where attentive effects appear. When describing grouping processes and the full primal sketch, he says, "our approach requires that the discrimination be made quickly – to be safe, in less than 160 ms – and that a clear psychophysical boundary be present" (Marr, 1982, p. 96). Not only is the number of experimental investigations enormous, but also the number of different models, theories and perspectives is large. Attention has been viewed as early selection (Broadbent, 1958), using attenuator theory (Treisman, 1964), as a late selection process (Norman, 1968, Deutsch and Deutsch, 1963), as a result of neural synchrony (Milner, 1974), using the metaphor of a spotlight (Shulman et al., 1979), within Feature Integration Theory (Treisman and Gelade, 1980), as an object-based phenomenon (Duncan, 1984), using the zoom lens metaphor (Eriksen and St. James, 1986), as a pre-motor theory subserving eye movements (Rizzolatti et al., 1987), as Guided Search (Wolfe et al., 1989), as Biased Competition (Desimone and Duncan, 1995), as feature similarity gain (Treue and Martinez-Trujillo, 1999), and more.

Within all of these different viewpoints, the only real constant seems to be that attentional phenomena seem to be due to inherent limits in processing capacity in the brain (Tsotsos, 1990). But even this does not constrain a solution. Even if we all agree that there is a processing limit, what is its nature? How does it lead to the mechanisms in the brain that produce the phenomena observed experimentally?

We suggest that the terms attention, recognition and binding have become so loaded that they mask the true problems; each may be decomposed into smaller problems, problems whose solution depends strongly on their inter-relationships.

The paper will begin by presenting a set of vision tasks and their definitions. It must be noted that the use of many terms in vision, regardless of which discipline uses them, is not consistent but in order to follow the arguments in this paper, one must adhere to the definitions provided strictly. After the definitions, Section 1 will continue by describing needed background on our particular perspective on attention and binding. Section 2 will provide a description of our visual attention model and the reader is alerted to the fact that details have appeared previously in many papers over many years, and cannot be replicated here; literature pointers for further reading are provided. Section 3 will present the result of our connection of vision task definition and observed experimental time course of those tasks, with proposed binding processes and attentional mechanisms within our model. These connections have been realized within a model simulation, published previously with citations provided as appropriate. The paper ends with a discussion of the implications of these connections and model predictions that require new experimental work.

1.1. *Defining vision sub-tasks*

All efforts to develop a computational theory of human vision must be informed by experimental observations of human (and also non-human primate) visual performance. Consequently, the terms attention, recognition and binding should be closely tied to the experiments that attempt to discover their characteristics within human vision; yet, one currently sees the terms quite arbitrarily used, especially in the computational vision literature. Macmillan and Creelman (2005) provide good definitions for many aspects of recognition and we can use these as a starting point. It is important to note that in some instances these definitions may not match the usual use of some of the terms involved; this paper will use the definitions strictly.

One-interval experimental design involves a single stimulus presented on each trial. Between trials visual masks are used to clear any previous signal traces. *Discrimination* is the ability to tell two stimuli apart. The simplest example is a *Correspondence* experiment in which the stimulus is drawn from one of two stimulus classes and the observer has to say from which class it is drawn. This is perhaps the closest to the way much of modern computer vision currently operates; computational neuroscience models usually do not go much further. A *Detection* task is where one of the two stimulus classes is null (noise) and the subject needs to choose between noise and noise+signal and the subject responds if he sees the signal. In a *Recognition* task neither stimulus is noise. More complex versions have more responses and stimuli. If the requirement is to assign a different response to each stimulus, the task is *Identification*. If the stimuli are to be sorted into a smaller number of classes – say, M responses to sort N stimuli into categories – it is a *Classification* task. The *Categorization* task requires the subject to connect each stimulus to a prototype, or class of similar stimuli (cars with cars, houses with houses). The *Within-Category Identification* task has the requirement that a stimulus is associated with a particular sub-category from a class (bungalows, split-level, and other such house types, for example). Responses can be of a variety of kinds: verbal, eye movement to target, the press of a particular button, pointing to the target, and more. The choice of response method can change the processing needs and overall response time.

In N-interval designs, there are N stimulus displays. In the *Same–Different* task a pair of stimuli is presented on each trial and the observer must decide if its two elements are the same or different. For the *Match-to-Sample* task, three stimuli are shown in sequence and the observer must decide which of the first two the third one matches. *Odd-man-out* is a task where the subject must locate the odd stimulus from a set where all stimuli are somehow similar while one is not. More complex designs are also used and Macmillan and Creelman detail them all; the point here is not to review all possibilities. Rather, the point is to present the definitions that we use in this paper and to stress that computational theories – if they wish to have relevance to human vision – need to consider the experimental procedures for each task when comparing their performance to experimental observations.

All experiments require a response from a subject, a response that in some cases requires knowledge of location of the stimulus perceived. This leads us to define a new task that is not explicitly mentioned in Macmillan and Creelman, the *Localization* task. In this task the subject is required to extract some level of stimulus location information in order to produce the response dictated by the experimenter. That level of location information may vary in its precision. Sometimes it may be sufficient to know only in which quadrant of the visual field a stimulus is found, other times a subject may need to know location relationships among stimuli, and so on. In fact,

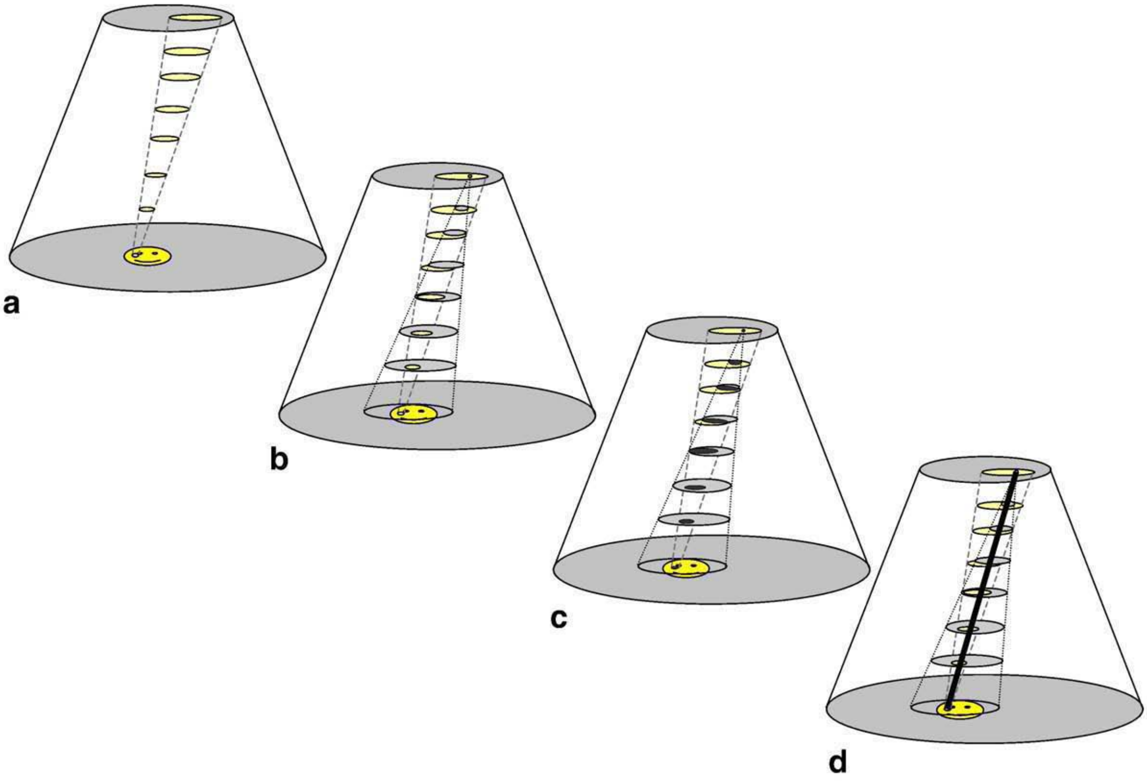

Fig. 1 – Examples of some the search problems in vision illustrated on a hypothetical processing pyramid. The figure shows a simplified hypothetical layered processing hierarchy (pyramid). A caricature of a face is the sole stimulus in the input array. The processing hierarchy is represented by the large truncated cone shape in the figure, without each layer drawn in for clarity. The activations, both feed-forward and feedback, due to the attended stimulus are drawn as smaller cones within the large one. These show only the extent of the pathways that the attended portion of the stimulus might activate. a. Feed-forward activation to be followed by maximum response selection within the region of projection of the input stimulus at the output layer. b. The receptive field of the selected neuron shown overlapping with the feed-forward projections, illustrating the extent of ambiguity that is present. c. The ambiguous regions are highlighted with dark shading. d. The target connectivity that would resolve the ambiguity.

this may be considered as an implicit sub-task for any of the standard tasks if they also require location information to formulate a response. Its importance will become apparent later in the paper. Throughout the paper, adding a superscript "L" to the task name will denote a task that also includes localization.

One final issue is worth noting here. Macmillan and Creelman (2005) point out that the basic psychophysical process is *comparison*. All psychophysical judgments of the kind described above are of one stimulus relative to another; experimental designs differ in kinds of comparisons. Thus, any theory or model of vision that does not include this basic functionality is not addressing the real issue. It is this basic comparison process that was the root of the computational complexity foundation of the Selective Tuning model (Tsotsos, 1989, 1990)[2] – its 'first principles' – and thus this forms a sound basis for a model of vision and attention.

1.2. *Attention as search optimization*

Fig. 1 shows a simplified hypothetical processing hierarchy (pyramid) with layers, for the sake of example, being retinal ganglion cells, LGN, V1, V2, V4, TEO, and TE. A caricature of a face is the sole stimulus in the input array. The point of the figure is to illustrate some of the aspects of attentive search. The processing hierarchy or pyramid is represented by the large truncated cone shape in the figure, the input layer at the bottom and the output layer at the top, without intermediate layers drawn in for clarity. The activations, both feed-forward and feedback, due to the attended stimulus are drawn as smaller cones within the large one. These show only the extent of the pathways that the attended portion of the stimulus might activate within each layer.

In the leftmost sub-figure (Fig. 1a) each point of that face would activate a feed-forward diverging network of connections and neurons. This is illustrated using a particular small region in the bottom or input layer of the figure; let's say that it is this small region that is attended. This, as well as any other portion of the stimulus, activates particular sub-regions at each layer of the visual processing hierarchy (the lightly shaded ovals). The

[2] This first principles and their relation to the comparison task are fully summarized in Tsotsos (2005).

diverging nature of activation is illustrated by the upward increasing size of activated region. That feed-forward neural connectivity has this diverging nature has been shown in many previous studies (Salin and Bullier, 1995; Angelucci et al., 2002; Gilbert and Sigman, 2007) and has been used in models for some time (Nowlan and Sejnowski, 1995; Roelfsema, 2006, to name an early one and a recent one). These works also show the reciprocal nature of feed-forward and feedback connectivity, which is required for the remainder of the figure as well as our model.

Suppose that on inspection of the topmost activated region the strongest responding neuron is selected within it, on the assumption that it might be the one that best represents the contents of the attended region in the input array (not unlike the process assumed by all computational models of visual attention). This is a search process: from a set of responses find the strongest one. That neuron has a receptive field and receives activation from a particular region in the input and that region has a well-defined set of pathways that lead to that neuron in the top layer as shown by the medium shaded ovals in Fig. 1b. This receptive field is much larger than the attended item creating a confound that must be resolved. What is that neuron responding to? The conflict does not exist only at the input or top layers, but throughout the hierarchy. The regions within each layer that exhibit this conflict in this toy example are shown by darkly shaded ovals in Fig. 1c. Some mechanism is needed to eliminate the ambiguity created by this situation, to connect the attended input directly to the neuron of interest at the top, and to ensure that it only 'sees' and responds to the attended item, as is shown in Fig. 1d.

This illustration is an abstraction, clearly simplified; imagine how the problems are compounded with several feature representations and several pathways as would be the case for any non-trivial stimulus being analyzed by the visual cortex.

Attention is most often thought of as selection of portion of the input for preferential processing, and as a result, only the portion of the above description that selects that first small region in the stimulus requires an attentional mechanism. We disagree and have maintained a view of attention as a *set* of mechanisms that optimize the search processes inherent in vision (Tsotsos, 1992, 2001). Certainly, the selection of fixation point/region in an image is a search problem, but it is hardly the only one. The selection of the strongest response at the output layer is another. The sequence in Fig. 1 is intended to illustrate another search problem, namely, that of searching for the set of pathways and neurons that best represent what is being attended. The search space is exactly the areas of the dark shaded ovals in Fig. 1c. One might consider that the search may be done in a feed-forward manner, a feedback manner, or perhaps otherwise. Search constraints can come from bottom-up (in the early stages the feed-forward activations lead to smaller regions than the receptive field) or top-down (in the later stages the receptive fields are smaller than the feed-forward activated regions). On the assumption that the decision criterion for search is maximum response within search regions at all stages, feed-forward selection of best responses cannot be guaranteed to converge on the selected neuron at the top; only local maxima will be found in each layer and there is no guarantee that using the feed-forward activation pathways only as a search guide will not lead to intermediate maxima that cause the search to veer away from the target neuron. A feedback approach will necessarily succeed as can be easily seen in the figure: from the global maximum at the top, using the receptive field boundaries within each layer as a guide, search will discover only pathways that lead to the source of that maximum response within the input layer. Assuming that the maximally responding neuron at the top is a good detector for the attended item, the stimulus that led to the global maximum is guaranteed to be found (this is formally proved in Tsotsos et al., 1995). Thus a top-down search is required to achieve the goal shown in Fig. 1d. This search process plays a major role in the remainder of the paper.

Some recent models of vision use a feed-forward maximum operation with the goal of solving the same sort of problem. Although the previous paragraph provided some rationale as to why this approach may not find a global maximum, more evidence can be presented as to why this is not likely to be biologically plausible. The experimental evidence against a feed-forward maximum operation is overwhelming. The majority of studies that have examined responses with two non-overlapping stimuli in the CRF have found that the firing rate evoked by the pair is typically lower than the response to the preferred of the two presented alone, inconsistent with a max rule (Miller et al., 1993; Reynolds et al., 1999; Missal et al., 1999; Recanzone et al., 1997; Reynolds and Desimone, 1998; Chelazzi, et al., 1998; Rolls and Tovee, 1995; Zoccolan, et al., 2005). Additional studies have found the response to the preferred stimulus changes when presented along with other stimuli, a pattern inconsistent with a feed-forward max operation (Sheinberg and Logothetis, 2001; Rolls et al., 2003). A theoretical argument may also be made against a feed-forward max using the equivalence conditions between relaxation labeling processes and max selection (Zucker et al., 1981), and especially considering the role of lateral processes in vision (Ben-Shahar et al., 2003). If lateral interactions are included time course matters. It has been observed that most V1 response increases due to lateral interactions seem to occur in the latter parts of the response profile. This hints that lateral interaction takes extra time to take effect with V1 responses continuing until about 300 ms after stimulus onset (Kapadia et al., 1995), well after the first feed-forward traversal has completed as will be described in subsequent sections.

The main point here is that attention is a set of search strategies, including search across a set of neural responses, search for the next fixation point, search for the set of neural pathways that best represent a stimulus, and search for the location and extent of a stimulus. Additional kinds of search are described in Tsotsos (1992).

1.3. *Visual feature binding*

A great deal of effort has gone into the discovery and elaboration of neural mechanisms that extract meaningful components from the images our retinae see in the belief that these components form the building blocks of perception and recognition. The problem is that corresponding mechanisms to put the pieces together again have been elusive even though the need is well accepted and many have studied the problem. This "Humpty-Dumpty" like task has been called the *binding problem* (Rosenblatt, 1961). Binding is usually thought of as taking one kind of visual feature, such as a shape, and associating it with another feature, such as location, to provide a unified representation of an object. Such explicit association ("binding") is particularly important when more than one visual object is present, in order to avoid incorrect combinations of features belonging to different objects, otherwise known as "illusory conjunctions" (Treisman and Schmidt, 1982). Binding is a broad problem: visual binding, auditory binding, binding across time, cross-modal binding, cognitive binding of a percept to a concept, cross-modal identification and memory reconstruction. The literature on binding and proposed solutions is large and no attempt is made here to review it due to space limitations (see Roskies, 1999).

Classical demonstrations of binding seem to rely on two things: the existence of representations in the brain that have no location information, and, representations of pure location for all stimuli. However, there is no evidence for a representation of location independent of any other information. Similarly, there is no evidence for a representation of feature without a receptive field. Nevertheless, location is *partially* abstracted away within a hierarchical representation as part of the solution to complexity (Tsotsos, 1990). A single neuron receives converging inputs from many receptors and each receptor provides input for many neurons. Precise location is lost in such a network of diverging feed-forward paths yet increasingly larger convergence onto single neurons (see Fig. 1b). How can location be recovered and connected to the right features and objects as binding seems to require?

The simplified example of Fig. 1 does not suffice to illustrate the magnitude of this problem. Suppose this pyramid is now replicated many times but with a common root or input, each corresponding to a particular pathway in the visual processing network, including many that have some representations in common. In other words, extend this to the actual network as shown, for example, by Felleman and Van Essen (1991). Now consider the following. Any stimulus will necessarily activate a feed-forward diverging cone of neurons through all pathways, and in each case, neural convergence causes location information to be partially lost. Furthermore there is no a priori reason to think that coordinate systems or cortical magnifications or resolutions are constant throughout the system, so there may be large differences in all of these at each level. How is the right set of pathways through this complex system identified and 'bound' together to represent an object?

Three classes of solutions to the binding problem have been proposed in the literature. Proponents of the convergence solution suggest that highly selective, specialized neurons that explicitly code each percept (introduced as cardinal cells by Barlow (1972) — also known as gnostic or grandmother neurons) form the basis of binding. The main problem with this solution is the combinatorial explosion in the number of units needed to represent all the different possible stimuli. Also, while this solution might be able to detect conjunctions of features in a biologically plausible network (i.e. a multi-layer hierarchy with pyramidal abstraction) it is unable to localize them in space on its own (Rothenstein and Tsotsos, 2008), and additional mechanisms are required to recover location information. Synchrony, the correlated firing of neurons, has also been proposed as a solution for the binding problem (Milner, 1974; von der Malsburg, 1981; Singer, 1999). Synchrony might be necessary for signaling binding, but is not sufficient by itself, as it is clear that this can at most tag bound representations, but not perform the binding process. The co-location solution proposed in the Feature Integration Theory (Treisman and Gelade, 1980) simply states that features occupying the same spatial location belong together. Due to its purely spatial nature, this solution cannot deal with transparency and other forms of spatial overlap. Also, since detailed spatial information is only available in the early areas of the visual system, simple location-based binding is agnostic of high-level image structure, which means that it cannot impose boundaries (obviously, the different edges of an object occupy different spatial locations), and arbitrary areas that belong to none, one or more objects can be selected.

It is important, given the debate over binding and vagueness of its definition, to provide something more concrete for the purposes of this paper. Here, a visual task – any of those defined earlier – will require a binding process if the input image contains more than one object in different locations (may be overlapping), the objects are composed of multiple features, and they share at least one feature type. If these conditions are not met, then no binding process is required.

1.4. *The role of time*

Regardless of visual task, it is the same pair of eyes, the same retinal cells, the same, LGN, V1, V2 and so forth, that process all incoming stimuli. Each step in the processing pathway requires processing time; no step is instantaneous or can be assumed so. In experiments such as those defined above, the timing for each of the input arrays is manipulated presumably in order to investigate different phenomena. There are many variations on these themes and this is where the ingenuity of the best experimentalists can shine. The argument made by this paper is to use time as an organizational dimension, that is, the most effective way of carving up the problem is to cut along the dimension of time. Throughout the paper, when referring to these time slices, the term 'stages' of recognition will be used. That is, a slice in time (or a contiguous set of slices) is a particular stage and the stages are identified with the above specific vision tasks each having its own temporal characteristic.

1.5. *Summary*

The first section of this paper presented the kinds of vision tasks that are of interest and introduced the visual feature binding problem. Definitions are given for the tasks and a perspective on visual attention focusing on a number of search processes required for some of these tasks are detailed. A number of questions were posed with this background. Based on these elements, we propose that the process of binding visual features to objects in each of the recognition tasks differs and that different sorts of binding actions take different amounts of processing time.

2. The model

We examined in detail the computational substrate provided by the Selective Tuning (ST) model of visual attention, to attempt to match the processing characteristics and qualitative time course of the various attentional mechanisms provided by ST to particular recognition tasks with experimental data in the literature, in order to derive the kinds of binding functionalities that each task would require, and then to abstract from these a set of procedures within the ST framework that would accomplish the visual feature binding required for each recognition task.

It is important to clarify the kind of modeling that this work represents. This is not data fitting (developing sets of equations that have good fit to existing experimental data) and it is not a learning model (whose characteristics are abstracted through statistical learning procedures from a large data base). Rather, this is a 'first principles' modeling effort. That is, beginning from issues related to the computational complexity of vision (Tsotsos, 1987, 1989, 1990, 1992, 2005; Parodi et al., 1998), moving to a definition of attention rooted in optimization of search processes in vision, and from there deriving a model that satisfies the complexity constraints and has qualitative performance in accord with known experimental observations. The details of the Selective Tuning model that are relevant for this paper are now presented; full details, have been previously presented (Tsotsos, 1990, 1993; Tsotsos et al., 1995, Tsotsos et al., 2001, 2005, 2007; Zaharescu et al., 2005; Tsotsos et al., 2005, Tsotsos et al., 2007, Rodriguez-Sanchez et al., 2007; Rothenstein et al., in press). These papers also show several examples and discussions of the performance of the model that are not repeated here.

2.1. *Selective Tuning*

Most models of vision, including ours, assume that a hierarchical sequence of computations defines the selectivity of a neuron. A feed-forward pass through the hierarchy would yield the strongest responding neurons if stimuli match existing neurons, or the strongest responding component neurons if stimuli are novel. Consider Fig. 2. The processing architecture is pyramidal, increasingly spatially coarser representations from bottom to top, units within each receiving both feed-forward and feedback connections from overlapping space-limited regions. It is assumed that response of units is a measure of goodness-of-match of stimuli within a receptive field to a neuron's selectivity. Task-specific bias, when available, allows the response to also reflect the relative importance of the contents of the corresponding receptive field in the scene. The bias is applied in a top-down manner, inhibiting neurons that code for features or elements that are not part of the task, not unlike the kind of dynamic synaptic changes that von der Malsburg (1981) described. The effect is that baseline firing rates are lowered for those task-irrelevant neurons.

The first stage of stimulus processing is a feed-forward pass. When a stimulus is applied to the input layer of the pyramid, it activates all of the units within the pyramid to which it is connected (as described in Fig. 1c). The result is a feed-forward, diverging cone of activity within the pyramid. Although it is well known that there are also lateral connections within the each level of representation these are not currently included in the model. As described earlier, the time course for the first feed-forward pass is short enough to not be affected by lateral interactions because they take longer to provide impact on responses. However, lateral interactions do

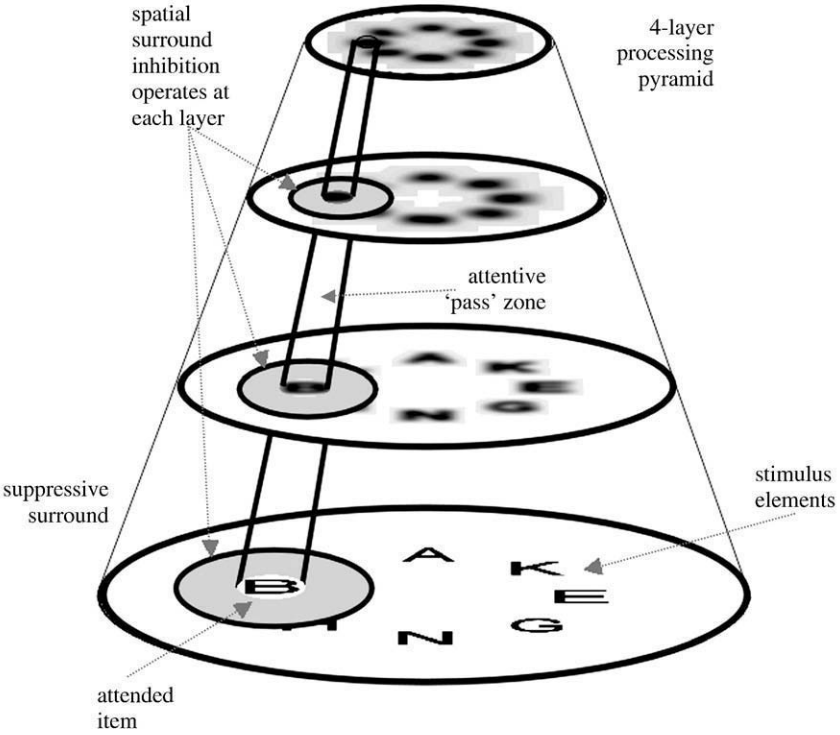

Fig. 2 – The basic Selective Tuning process of selection and surround suppression.

Please cite this article as: Tsotsos, J.K., et al., The different stages of visual recognition need different attentional binding strategies, Brain Res. (2008), doi:10.1016/j.brainres.2008.05.038

have an effect further along in processing time assuming static stimuli; these are currently under development in the model (but see Šetić and Domijan, 2007 or Ben-Shahar et al., 2003).

The second stage is a feedback pass embodying a hierarchical winner-take-all (WTA) process. The WTA can accept task guidance for areas or stimulus qualities if available but operates independently otherwise. The global winner at the top of the pyramid activates a WTA that operates only over its direct inputs. This localizes the largest response units within the top level winning receptive field. All of the connections of the visual pyramid within this receptive field that do not contribute to the winner are inhibited. This refines unit responses and improves signal-to-noise ratio. The top layer is not inhibited by this mechanism. The strategy of finding the winners within successively smaller receptive fields, layer by layer, and then pruning away irrelevant connections is applied recursively. The result is that the cause of the largest response is localized in the sensory field. The paths remaining may be considered the pass zone while the pruned paths form the inhibitory zone of an attentional beam as shown in Fig. 2.

In reality there is no single output representation for vision in the brain; there are many representations. ST requires a competition among all the representations at the top layer, biased by task. The type of competition is determined by the relationships among the active representations. Two types are considered here (and are detailed below). Two representations are mutually exclusive if, on a location-by-location basis, the two features/objects they represent cannot both be part of the same object or event (eg., an object cannot have a velocity in two directions or two speeds at the same location at the same time). Two representations may co-exist if the two features they represent can both be part of some object or event (eg., an edge may have color, a line may be at some disparity, features belonging to eyes and co-exist with those from noses, etc.). This global competition not only detects the neurons that are best tuned to the stimulus but also enables the selection of neurons that represent parts of stimuli for novel items for which no object tuning has yet been learned.

The following method is applied at the top of all pyramids at first, then recursively downwards following the representations of the winning units.

2.2. ST's winner-take-all process

WTA processes have appeared in virtually all models of visual attention since the Koch and Ullman model (1985). The one used by ST is unique; the basis for its distinguishing characteristic is that it implicitly creates a partitioning of the set of unit responses into bins of width determined by a task-specific parameter, θ. The partitioning arises because inhibition between units is not based on the value of a single unit but rather on the difference between pairs of unit values. Competition depends linearly on the difference between unit strengths. Unit A inhibits unit B if the response of A, denoted by $r(A)$, satisfies $r(A)-r(B) > \theta$. Otherwise, A will not inhibit B. The inhibition on unit B is the weighted sum of all inhibitory inputs, each of whose magnitude is determined by $r(A)-r(B)$. It has been shown that this WTA is guaranteed to converge, has well-defined properties with respect to finding largest items, and has well-defined convergence characteristics (Tsotsos et al., 1995).

The WTA process has two stages: the first is to inhibit all responses except those in the largest θ-bin; and, the second is to find the largest, strongest responding region represented by a subset of those surviving the first stage. The general form is:

$$G_i(t+1) = G_i(t) - \sum_{j=1, j\neq i}^{n} w_{ij}\Delta_{ij} \qquad (1)$$

where $G_i(t)$ is the response of neuron i at time t, w_{ij} is the connection strength between neurons i and j, (the default is that all weights are equal; task information may provide different settings), n is the number of competing neurons, and Δ_{ij} is given by:

$$\Delta_{ij} = G_j(t) - G_i(t), \quad \text{if } 0<\theta<G_j(t) - G_i(t) \text{ and otherwise } 0. \qquad (2)$$

$G_i(0)$ is the feed-forward input to neuron i. The corresponding differential equation is:

$$\frac{dG_i}{dt} = I_i(t) - \alpha_i G_i - \sum_{j=1}^{n} w_{ij}\Delta_{ij} \qquad (3)$$

where $I_i(t)$ is the external input to neuron i (at $t=0$, $I_i=G_i$), α_i is the rate constant of passive decay for neuron i.

Stage 2 applies a second form of inhibition among the winners of the stage 1 process. The larger the spatial distance between units the greater is the inhibition. A large region will inhibit a region of similar response strengths but of smaller spatial extent on a unit-by-unit basis. Eq. (1) governs this stage of competition also with two changes: the number of survivors from stage 1 is m, replacing n everywhere, and Δ_{ij} is replaced by:

$$\Phi_{ij} = \mu\left(G_j(t) - G_i(t)\right)\left(1 - e^{-\frac{d_{ij}^2}{d_r^2}}\right), \quad \text{if } 0<\theta<\mu\left(G_j(t) - G_i(t)\right)\left(1 - e^{-\frac{d_{ij}^2}{d_r^2}}\right) \text{ and otherwise } 0. \qquad (4)$$

μ controls the amount of influence of this processing stage (the effect increases as μ increases from a value of 1), d_{ij} is the retinotopic distance between the two neurons i and j, and d_r controls the spatial variation of the competition.

2.3. Summary

The Selective Tuning winner-take-call process is a provable method to trace neural connections from the strongest responding neuron in the top layer of a hierarchy to the elements in the stimulus array that caused that strongest response. Nevertheless, as presented, it does not provide for the extensions required to permit more than one feature hierarchy or pathway to contribute to that strongest responding neuron. Although it has shown good performance, this extension is the missing element that any binding process would require. These extensions are the modeling focus of this paper.

Please cite this article as: Tsotsos, J.K., et al., The different stages of visual recognition need different attentional binding strategies, Brain Res. (2008), doi:10.1016/j.brainres.2008.05.038

3. Results

The main question this work addresses is: How is the right set of pathways through this complex system identified and 'bound' together to represent an object? A novel set of four different binding processes are introduced that are claimed to suffice for solving the kinds of recognition tasks described above. Fig. 3 is the main descriptive vehicle that ties recognition, attention and binding together. The stages of the figure will be featured as main sub-sections here, within which the details of task, the kinds of attentional mechanisms involved, and the binding process are described.

3.1. *Priming*

Prior to any of the above tasks, the first set of computations to be performed is priming the hierarchy of processing areas (Posner et al., 1978). Task knowledge, such as fixation point, target/cue location, task success criteria, and so on must somehow be integrated into the overall processing; they *tune* the hierarchy. It has been shown that such task guidance must be applied 300 to 100 ms before stimulus onset to be effective (Müller and Rabbitt, 1989). This informs us that significant processing time is required for this step alone. It is a sufficient amount of time to complete a top-down traversal of the full processing hierarchy before any stimulus is shown. The first stage, the leftmost element of Fig. 3, shows this priming stage. Tuning in the ST model takes the form of multiplicative inhibition against features and locations that are not part of the target or task achieved via a full top-down pass of the visual processing hierarchy. Once complete, the stimulus can be presented (the second element of the figure from the left).

3.2. *Discrimination*

The third element of Fig. 3 represents the one-interval *Discrimination Task* as long as no location information is required for a response. This task was defined in Section 1.1 as the ability to tell two stimuli apart, and several sub-categories were defined: correspondence, detection, recognition, categorization, classification. Detecting whether or not a particular object is present in an image seems to take about 150 ms (Thorpe et al., 1996). Marr, in his definition of full primal sketch, required about this time to suffice for segregation, as mentioned in the introduction and thus his entire theory falls within this task too. This kind of 'yes–no' response can also be called 'pop-out' in visual search with the added condition that the speed of response is the same regardless of number of distractors (Treisman and Gelade, 1980). The categorization task also seems to take the same amount of time (Grill-Spector and Kanwisher, 2005; Evans and Treisman, 2005). Interestingly, the median time required for a single feed-forward pass through the visual system is about 150 ms (Bullier, 2001). Thus, we conclude that a single feed-forward pass suffices for this visual task and this is completely in harmony with many authors. This first feed-forward pass is shown in the figure emphasizing the feed-forward divergence of neural connections and thus stimulus elements are 'blurred' progressively more in higher areas of the hierarchy. These tasks do not include location or location judgments, the need to manipulate, point, or other motor commands specific to the object

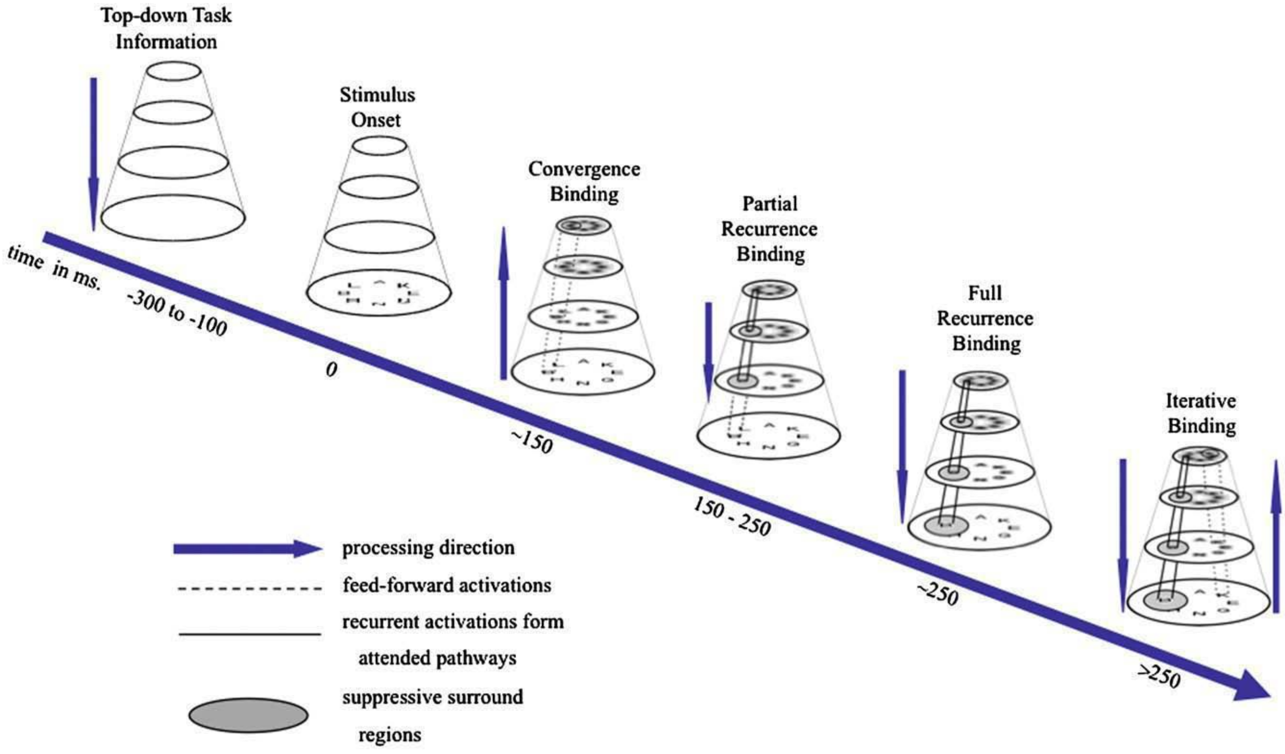

Fig. 3 – The different binding processes associated with different time periods during recognition tasks.

and usually, all objects can be easily segmented, as Marr required. That is, the background may provide clutter, but the clutter does not add ambiguity; the definition of image class that requires binding processes from Section1.3 is not satisfied.

Convergence Binding achieves the *Discrimination Task* via hierarchical neural convergence, layer by layer, in order to determine the strongest responding neural representations at the highest layers of the processing hierarchy. This feed-forward traversal follows the task-modulated neural pathways through the 'tuned' visual processing hierarchy. This is consistent with previous views on this problem (Treisman, 1999; Reynolds and Desimone, 1999). This type of binding will suffice only when stimulus elements that fall within the larger receptive fields are not too similar or otherwise interfere with the response of the neuron to its ideal tuning properties. Such interference may be thought of as 'noise' with the target stimulus being 'signal'. Convergence Binding provides neither a method for reducing this noise nor a method for recovering precise location. The accompanying attentional process is the search is over the top level representation to find the strongest responding neurons.

For a task where there is more than one stimulus in a sequence but where the information required of each stimulus can be extracted via a Discrimination alone, the feed-forward pass can be repeated. In fact, 'waves' or 'cascades' of stimuli continually flow through the system, but as each one passes through the full system, if inspection of the results at the top of the hierarchy suffices, then these are each Discrimination tasks. Each of the tasks presented in this paper may be repeated in a similar fashion; we denote this kind of process with the prefix "R-". Thus a task such as RSVP (Rapid Serial Vision Presentation) is an example of *R-Discrimination*.

3.3. *Identification*

To provide more detail about a stimulus, such as for a within-category identification task, requires additional processing time, 65 ms or so (Grill-Spector and Kanwisher, 2005; Evans and Treisman, 2005); this is represented by the fourth from the left element of Fig. 3. If the highest levels of the hierarchy can provide the basic category of the stimulus, such as 'bird', where are the details that allow one to determine the type of bird? The sort of detail required would be size, color, shape, and so forth. These are clearly lower level visual features and thus they can only be found in earlier levels of the visual hierarchy. They can be accessed by looking at which feature neurons feed into those neurons that provided the category information. One way to achieve this is to traverse the hierarchy downwards, beginning with the category neuron and moving downwards through the needed feature maps. This downward traversal is what requires the additional time observed. The extent of downward traversal is determined by the task, that is, the aspects of identification that are required. It is interesting to consider an additional impact of a partial downwards traversal. This traversal may be partial not only because of the task definition but also because a full traversal is interrupted and not allowed to complete either because new stimuli enter the system before there is enough time for completion or because not enough time is permitted due to other tasks. The result is that there is the potential for errors in localization and these may lead to the well-known illusory conjunction phenomenon (Treisman and Schmidt, 1982). This group of tasks will be termed *Identification Tasks*.

Partial Recurrence Binding can find the additional information needed to solve the *Identification Task* if it is represented in intermediate layers of the processing hierarchy. If this is not deployed directly due to task needs but is due to interruption, then this may result in illusory conjunctions. A variety of different effects may be observed depending on when during the top-down traversal the process is interrupted. There is no specific image class for which this process applies; it can be applied in all cases. Some aspects of coarse location information may also be recovered with a partial downward search (such as in which quadrant the stimulus lies). The process for recurrence is described in Section 2.2.

3.4. *Localization*

If detailed or precise localization is required for description or a motor task, (pointing, grasping, etc.), then the top-down traversal process must be allowed to complete and thus additional time is required. These are the *Discrimination^L Tasks*, or simply, *Localization Tasks*. How much time? A lever press response seems to need 250–450 ms in monkey (Mehta et al., 2000). During this task, the temporal pattern of attentional modulation shows a distinct top-down pattern over a period of 35–350 ms post-stimulus. The 'attentional dwell time' needed for relevant objects to become available to influence behavior seems to be about 250 ms (Duncan et al., 1994). Pointing to a target in humans seems to need anywhere from 230 to 360 ms (Gueye et al., 2002; Lünenburger and Hoffman, 2003). Still, none of these experiments cleanly separate visual processing time from motor processing time; as a result, these results can only provide an encouraging guide for the basic claim of our model and further experimental work is needed.

Behavior, i.e., an action relevant to the stimulus, requires localization. The precise location details are available only in the earliest layers of the visual processing hierarchy because that is where the finest spatial resolution of neural representation can be found. As a result, the top-down traversal must complete so that it reaches these earliest layers as shown in the figural element second from the right in Fig. 3 for location details. Note that intermediate points in the top-down traversal can provide intermediate levels of location details; full traversal is needed only for the most precise location needs.

Full Recurrence Binding achieves the *Localization Task*. If Convergence Binding is followed by a complete top-down traversal, attended stimuli in each feature map of the hierarchical representation can be fully localized. Recurrent traversals through the visual processing hierarchy 'trace' the pathways of neural activity that lead to the strongest responding neurons at the top of the hierarchy. The details of the algorithm for this process appeared in Section 2.

Full Recurrence Binding can determine the location and spatial extent of a detected object/event for images such as those defined for Convergence Binding, where there is no ambiguity and proper detection can occur without a special binding process. It can also do so for those images that do contain ambiguity of this kind described in Section 1.3 and for this class of images, Recurrence Binding is required for task completion. This means explicitly that segmentation is not immediate in the Marr sense, that there are multiple objects in an image that share features and thus a simple convergence via binding faces ambiguity and fails to find a clear winner.

There is one more critical component of the top-down traversal, appearing on the figures as gray regions indicating areas of neural suppression or inhibition in the area surrounding the attended stimulus. This area is defined by the feedback connections of the chosen neuron at the top. Inputs corresponding to the stimulus most closely matching the tuning characteristics of the neuron form the signal while the remainder of the input within that receptive field is the noise. Any lateral connections are also considered as noise for this purpose. Thus, if it can be determined what those signal elements are, the remainder of the receptive field is suppressed including lateral signals, enhancing the overall signal-to-noise ratio of processing for that neuron. The method for achieving this was first described in (Tsotsos, 1993) and fully detailed together with proofs of convergence and other properties in (Tsotsos et al., 1995).

However, the top-down process is complicated by the fact that each neuron within any layer may receive input from more than one feature representation. How do the different representations contribute to the selection? Different features may have different roles. For example, there are differing representations for many different values of object velocity however an object can only exhibit one velocity. These different representations can be considered as mutually exclusive, so the top-down search process must select one, the strongest. On the other hand, there are features that cooperate, such as the features that make up a face (nose, eyes, etc.). These contribute to the face neuron and the top-down search process much select appropriate elements from each. There may be other roles as well. The key here is that each neuron may have a complex set of inputs, specific to its tuning properties, and the top-down traversal must be specific to each. This is accomplished by allowing the choices to be made locally, at each level, as if there were a localized saliency representation for each neuron (Tsotsos et al., 2005). There is no global representation of saliency required. This is further explored in Section 3.6.

3.5. *Extended Discrimination*

The *Extended Discrimination Task* includes two-or-more interval designs, visual search, odd-man-out, resolving illusory conjunctions, determining transparency, any task requiring sequences of saccades or pursuit eye movements, and more (eg., Treisman and Gelade, 1980; Treisman and Schmidt, 1982; Wolfe, 1998; Schoenfeld et al., 2003). The final element of the figure, the rightmost element, depicts the start of a second feed-forward pass to illustrate this. The idea is that it is likely that several iterations of the entire process, feed-forward and feedback, may be required to solve difficult tasks.

Iterative Recurrence Binding is needed for the *R-Discrimination^{L} Task*. Iterative Recurrence Binding is defined as one of more Convergence Binding-Full Recurrence Binding cycles. The processing hierarchy may be tuned for the task before each traversal as appropriate. The iteration terminates when the task is satisfied.

There are at least two types of Iterative Recurrence Binding. The first is the more obvious one, namely, multiple attentional fixations are required for some task. The second permits different pathways to be invoked. Consider a motion stimulus; motion-defined form where a square of random elements rotates in a background of similar random elements. A rotating square is perceived even though there is no edge information present in the stimulus. After one cycle of Full Recurrence Binding, the motion can be localized and the surround suppressed. The suppression changes the intermediate representation of the stimulus so that any edge detecting neurons in the system now see edges, edges that were not apparent because they were hidden in the noise. As a result, the motion is recognized and with an additional processing cycle the edges can be detected and bound with the motion. Such examples and the model simulation results can be found in Tsotsos et al., 2005 and Rothenstein et al. (in press).

3.6. *The Selective Tuning approach to visual feature binding*

The binding strategy depends on the hierarchical WTA method to trace back the connections in the network along which feed-forward activations traveled. The WTA described above deals with a single pyramid. However, almost all neurons in visual cortex receive input from more than one representation. How is the top-down tracing guided for more than one representation? What we need is an extension, motivated in Section 2.3, that provides the solution to the localization problem and links all the component features from different representations of an object via the pass pathways of the attentional beam. The additional elements that comprise this method are now presented.

Define the Featural Receptive Field (FRF) to be the set of all the direct inputs to a neuron. This can be specified by the union of *k* arbitrarily shaped, contiguous, possibly overlapping sub-fields as

$$\mathrm{FRF} = \bigcup_{j=1,k} f_j \qquad (5)$$

where $\{f_j=\{(x_{j,a}, y_{j,a}), a=1, \ldots,b_j\}, j=1,\ldots,k\}$, (x,y) is a location in sub-field f_j, b_j is the number of units in sub-field f_j. The f_js may be from any feature map, and there may be more than one sub-field from a feature map. F is the set of all sub-field identifiers 1 though k. Response values at each (x,y) location within sub-field $i \in F$ are represented by $r(i,x,y)$.

The FRF definition applies to each level of the visual processing hierarchy, and to each neuron within each level. Suppose a hierarchical sequence of such computations defines the selectivity of a neuron. Each neuron has input from a set of neurons from different representations and each of those neurons also have a FRF and their own computations to combine its input features. With such a hierarchy of computations, a stimulus-driven feed-forward pass would yield the strongest responding neurons within one representation if the stimulus matches the selectivity of existing neurons, or the strongest responding component neurons in different representations if the stimulus does not match an existing pattern. The result is that the classical receptive field (the region of the visual field in which stimulation causes the neuron to fire) now has internal structure reflecting the locations of the stimulus features.

Features may be mutually exclusive (one depth or one velocity at each location) or features may co-exist (orientation and color both at the same location) or perhaps other kinds of inter-feature constraints may apply. Features may also be weighted differently depending on their task relevance.

If features are mutually exclusive by location, call this a Type A situation. The feature sub-fields completely overlap. The WTA, stages 1 and 2, can be extended to Type A situations. The competition proceeds as defined above for each of the sub-fields f_j separately so that within each, the largest, strongest responding contiguous region is found. The winners from each sub-region then compete for the overall largest, strongest contiguous region. The overall largest, strongest region is the region whose sum of the responses of the winning locations is greatest.

Let the winning region in feature map f be $g_f=\{(x_{i,f},y_{i,f})|i=1,2,.....n_f\}$, where n_f is the number of locations (by retinotopic location in the feature map) that comprise the winning region. The response value of the overall winner is

$$V_A = \max_{j \in F} \sum_{x,y \in g_j} r(j,x,y) \tag{6}$$

where $r(j,x,y)$ is the response value of unit at location x,y in feature map f and the extent of the winning region is the same as that of region g_f, the value of f being the one which wins the max selection.

If features can co-exist, call this a Type B situation. The feature sub-fields may be non-overlapping. In this case, the largest, strongest responding region is found in each representation separately in the same manner as for single representations defined above. The overall winner is then the union of these winners, with the response value being

$$V_B = \sum_{j \in f} \sum_{x,y \in g_j} r(j,x,y) \tag{7}$$

and the extent is given by the union of the feature maps involved. Other forms of feature combination may exist and these can be formulated analogously.

ST seeks the best matching scene interpretations (highest response) as a default (defaults can be tailored to task). This is the set of neurons chosen by the WTA competition throughout the hierarchy. If this happens to match the target of the search, then detection is complete. If not, a second candidate region is chosen and this proceeds until a decision on detection can be made. Localization is accomplished by the downward search to identify the feed-forward connections that led to the neuron's response following the network's retinotopic topology, using the FRFs all the way down the hierarchy. FRFs provide for a distributed, localized saliency computation appropriate for complex feature types and complex feature combinations. What is salient for each neuron is determined locally based on its FRF; saliency is not a global, homogeneous computation. Once localization is complete for all features, the object is attached to its components through the attention pass beams.

Simulations of this strategy show strong agreement with a variety of psychophysical and neurophysiologic experiments such as static visual searches of varying difficulties, segregation of transparent dot pattern motions, surround inhibition, and more (Rothenstein and Tsotsos, 2006; Rodriguez-Sanchez et al., 2007; Tsotsos et al., 2005; Tsotsos , 1995). In particular the surround inhibition prediction seems well supported by a variety of experimental studies (Cutzu and Tsotsos, 2003; Hopf et al., 2006; Tombu and Tsotsos, 2008). ST's top-down attentional modulation hypothesis also has good support (Mehta et al., 2000; O'Connor et al., 2002).

Table 1 – The vision tasks identified in the model with timings, attentive and binding processes summarized

Task	Approximate processing time required[a]	Attentional process	Binding process
Priming	−300 to −100 ms	Suppression of task irrelevant features, stimuli or locations; location cues, fixation points, task success criteria	N/A
Discrimination	150 ms	Search for maximum response	Convergence
Identfsification	215 ms	Top-down feature search	Partial Recurrence
Localization	200–360 ms	Top-down stimulus segmentation and localization	Full Recurrence
Extended Discrimination	>250 ms	Sequences of convergence and recurrence binding, perhaps with task priming specific to each pass	Iterative Recurrence

[a] Supported by evidence cited in text.

With respect to the basic mechanisms described above, there are two additional points to address. For Type B competition, i.e., for features that may co-exist, if the winning regions are at different locations, how can it be decided whether or not those stimuli belong together. The search is aided by the following observation: proper progress of the downwards traversal means response along the pass zone never decreases. The reason for this is that if the choice is correct, suppression of the surround (noise) will have the side-effect of increasing the neuron's response. If there is a decrease, the search fails, and a new peak must be chosen.

What if in a FRF there are peaks in both competitive and cooperative representation sets? It is in general not possible using the structure here to determine at the top level if peaks arise from the same object in the input; there is too much location abstraction to permit this. If the feature sub-fields are overlapping then the possibility that the features do arise from a single object increases, but this is not definitive. It is assumed that the winning region of the competitive set participates as a cooperation feature with the remaining cooperative features. Thus it participates in the overall winning region and the system attempts to localize it. If it is found

Please cite this article as: Tsotsos, J.K., et al., The different stages of visual recognition need different attentional binding strategies, Brain Res. (2008), doi:10.1016/j.brainres.2008.05.038

that the overall response decreases the search fails and a new feature grouping can be tried. These problems stand as model predictions; there will be situations where responses to visual tasks will be slower if stimuli are sufficiently complex or ambiguous to trigger one of these problems.

4. Discussion

A novel view of how attention, visual feature binding, and recognition are inter-related has been presented. It differs from any of those presented previously (Roskies, 1999). The greatest point of departure is that it provides a way to integrate binding with recognition tasks and with attention. The visual binding problem is decomposed into four kinds of processes each being tied to one of the classes of recognition behaviors defined by task and time course. Table 1 provides a summary of the kinds of vision tasks, their temporal requirement, the kind of attention process and binding process involved. We view this as a first version of such a decomposition and strongly believe that it requires a significant amount of effort to adequately complete and hope the community will take up the challenge to assist. In particular the Extended Discrimination Task is far too broad and requires refinement.

This view differs from conventional wisdom that considers both binding and recognition as monolithic tasks and attention as one or two simply processes only. The decomposition has the promise of dividing and conquering these problems, and the Selective Tuning strategy is proposed as the computational substrate for their solution. There are three ideas behind this solution:

- top-down task directed priming before processing;
- feed-forward traversal through the 'tuned' visual processing hierarchy following the task-modulated neural pathways;
- recurrent (or feedback) traversals through the visual processing hierarchy that 'trace' the pathways of neural activity from the strongest responding neurons at the top of the hierarchy to the input that caused the strongest response.

These three basic steps are used in combination, and repeated, as needed to solve a given visual task. In simulation with artificial as well as real images as input, the model exhibits good agreement with a wide variety of experimental observations (Tsotsos et al., 1995, 2005; Rothenstein et al., in press; Zaharescu et al., 2005; Rodriguez-Sanchez et al., 2007).

The idea of tracing back connections in a top-down fashion was present in part, in the Neocognitron model of Fukushima (1986) and suggested even earlier by Milner (1974). It also appears in the Reverse Hierarchy Model of Ahissar and Hochstein (1997). Within the Selective Tuning model, it was first described in Tsotsos (1993), with accompanying details and proofs in Tsotsos et al. (1995). Only Neocognitron and Selective Tuning provide realizations; otherwise, the two differ in all details. Fukushima's model included a maximum detector at the top layer to select the highest responding cell and all other cells were set to their rest state. Only afferent paths to this cell are facilitated by action from efferent signals from this cell. The differences between Neocognitron and ST are many. Neural inhibition is the only action of ST, with no facilitation. The Neocognitron competitive mechanism is lateral inhibition at the highest and intermediate levels that finds strongest single neurons thus assuming all scales are represented explicitly, while ST finds regions of neurons removing this unrealistic assumption. For ST, units losing the competition at the top are left alone and not affected at all. ST's inhibition is only within afferent sets to winning units. Finally, Fukushima assumes that so-called grandmother cells populate the top layer whereas ST makes no such assumption. Overall, the Neocognitron model and its enhancements cannot scale and would suffer from representational and search combinatorics (Tsotsos, 1990).

The validation of our model can be not only computational in the sense of performance on real images. Such a model can also be validated by showing that it makes counter-intuitive predictions for biological vision that gain experimental support over time. The following predictions appeared in Tsotsos 1990:

1) Attention imposes a spatial suppressive surround around attended items (Cutzu and Tsotsos, 2003; Hopf et al., 2006);
2) Attention imposes a suppressive surround around attended features, suppressing responses from nearby features in that feature dimension (Tombu and Tsotsos, 2008);
3) The surround suppression is a result of recurrent processing in the cortex (Boehler et al., submitted);
4) Selection is a top-down process where attentional guidance and control are integrated into the visual processing hierarchy;
5) The latency of attentional modulations *decreases* from lower to higher visual areas (Mehta et al., 2000);
6) Attentional modulation appears wherever there is many-to-one, feed-forward neural convergence (O'Connor et al., 2002);
7) Topographic distance between attended items and distractors affects the amount of attentional modulation;
8) There is an oscillatory nature to attention because it takes time for attention to be deployed throughout the processing network. As a result, attention appears to sample a stimulus in well-defined intervals, resulting in gaps of attention (VanRullen et al., 2007, Raymond et al., 1992).

For many of these, significant supporting evidence has accrued over the intervening years (representative citations provided).

The binding solution has some interesting characteristics that may be considered as predictions requiring investigation in humans or non-human primates:

1) Given a group of identical items in a display, say in a visual search task, subsets of identical items can be chosen as a group if they fit within receptive fields. Thus, the slope of observed response time versus set size may be lower than expected (not a strictly serial search).
2) There is no proof that selections made at the top of several pyramids will converge to the same item in the stimulus array. Errors are possible if items are very similar, if items are spatially close, or if the strongest responses do not arise from the same stimulus item.
3) Binding errors may be detected either at the top by matching the selections against a target, or if there is no target, by the end of the binding attempt when the pass

beams do not converge. The system then tries again; the prediction is that correct binding requires time that increases with stimulus density and similarity. In terms of mechanism, the ST model allows for multiple passes and these multiple passes reflect additional processing time.

4) ST's mechanism suggests that detection occurs before localization and that correct binding occurs after localization. Any interruption of any stage will result in binding errors.

Our model has a number of important characteristics: a particular time course of events during the recognition process covering the simplest to complex stimuli that can be directly compared with qualitative experimental time courses; an iterative use of the same visual processing hierarchy in order to deal with the most complex stimuli; iterative tuning of the same visual processing hierarchy specific to task requirements; suppressive surrounds due to attention that assist with difficult segmentations; a particular time course of events for recognition ranging from simple to complex recognition tasks; a top-down localization process for attended stimuli based on tracing feed-forward activations guided by localized saliency computations. Each of these may be considered a prediction for human or non-human primate vision. It would be very interesting to explore each.

Acknowledgments

Research support was gratefully received from the Natural Sciences and Engineering Research Council of Canada, the Canada Foundation for Innovation, the Ontario Innovation Trust, and the Canada Research Chairs Program. We thank Allison Sekuler and Patrick Bennett for literature pointers, Laurent Itti for discussion, John Reynolds, Mazyar Fallah and Steven Zucker for discussion and literature pointers, and Michael Arbib for comments on an early draft.

REFERENCES

Ahissar, M., Hochstein, S., 1997. Task difficulty and the specificity of perceptual learning. Nature 387, 401–406.

Angelucci, A., Levitt, J., Walton, E., Hupe, J., Bullier, J., Lund, J., 2002. Circuits for Local and Global Signal Integration in Primary Visual Cortex, J. Neurosci. 22 (19), 8633–8646.

Barlow, H.B., 1972. Single units and sensation: a neuron doctrine for perceptual psychology? Perception 1 (4), 371–394.

Ben-Shahar, O., Huggins, P., Izo, T., Zucker, S.W., 2003. Cortical connections and early visual function: intra- and inter-columnar processing. J. Physiol. (Paris) vol. 97 (No 2), 191–208.

Boehler, C.N., Tsotsos, J.K., Schoenfeld, M.A., Heinze, H.-J., Hopf, J.-M., The center-surround profile of the focus of attention arises from recurrent processing in visual cortex, (submitted).

Broadbent, D., 1958. Perception and Communication. Pergamon Press, NY.

Bullier, J., 2001. Integrated model of visual processing. Brain Res. Rev. 36, 96–107.

Chelazzi, L., Duncan, J., Miller, E., Desimone, R., 1998. Responses of Neurons in Inferior Temporal Cortex During Memory-Guided Visual Search, J. Neurophysiol. 80 (No 6), pp. 2918–2940.

Cutzu, F., Tsotsos, J.K., 2003. The selective tuning model of visual attention: testing the predictions arising from the inhibitory surround mechanism. Vis. Res. 43, 205–219.

Deutsch, J., Deutsch, D., 1963. Attention: some theoretical considerations. Psych. Rev. 70, 80–90.

Desimone, R., Duncan, J., 1995. Neural mechanisms of selective visual attention. Ann. Rev. Neurosci. 18, 193–222.

Duncan, J., 1984. Selective attention and the organization of visual information. J. Exp. Psychol. Gen. 113 (4), 501–517.

Duncan, J., Ward, J., Shapiro, K., 1994. Direct measurement of attentional dwell time in human vision. Nature 369, 313–315.

Eriksen, C., St. James, J., 1986. Visual attention within and around the field of focal attention: a zoom lens model. Percept. Psychophys. 4, 225–240.

Evans, K., Treisman, A., 2005. Perception of objects in natural scenes: is it really attention free? J. Exp. Psychol.: Hum. Percept. Perform. 31–6, 1476–1492.

Felleman, D., Van Essen, D., 1991. Distributed hierarchical processing in the primate visual cortex. Cerebral Cortex Vol 1, 1–47.

Fukushima, K., 1986. A neural network model for selective attention in visual pattern recognition. Biol. Cybern. vol 55 (1), 5–15.

Gilbert, C., Sigman, M., 2007. Brain States: Top-Down Influences in Sensory Processing, Neuron 54, 677–696.

Grill-Spector, K., Kanwisher, N., 2005. Visual recognition: as soon as you know it is there, you know what it is. Psychol. Sci. 16, 152–160.

Gueye, L., Legalett, E., Viallet, F., Trouche, E., Farnarier, G., 2002. Spatial orienting of attention: a study of reaction time during pointing movement. Neurophysiol. Clin. 32, 361–368.

Hopf, J.-M., Boehler, C.N., Luck, S.J., Tsotsos, J.K., Heinze, H.-J. Schoenfeld, M.A., 2006. Direct neurophysiological evidence for spatial suppression surrounding the focus of attention in vision. PNAS 103 (4), 1053–1058.

James, W., 1890. The Principles of Psychology, H. Holt.

Kapadia, M., Ito, M., Gilbert, G., Westheimer, G., 1995. Improvement in Visual Sensitivity by Changes in Local Context: Parallel Studies in Human Observers and in V1 of Alert Monkeys. Neuron 15, 843–856.

Koch, C., Ullman, S., 1985. Shifts in selective visual attention: towards the underlying neural circuitry. Hum. Neurobiol. 4, 219–227.

Lünenburger, L., Hoffman, K.-P., 2003. Arm movement and gap as factors influencing the reaction time of the second saccade in a double-step task. Eur. J. Neurosci. 17, 2481–2491.

Macmillan, N.A., Creelman, C.D., 2005. Detection Theory: A User's Guide,. Routledge.

Marr, D., 1982. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Henry Holt and Co., New York.

Miller, E.K., Gochin, P.M., Gross, C.G., 1993. Suppression of visual responses of neurons in inferior temporal cortex of the awake macaque by addition of a second stimulus, Brain Res. 616 (1–2), 25–29.

Milner, P.M., 1974. A model for visual shape recognition. Psychol. Rev. 81–6, 521–535.

Mehta, A., Ulbert, I., Schroeder, C., 2000. Intermodal selective attention in monkeys. I: distribution and timing of effects across visual areas. Cerebral Cortex 10 (4), 343–358.

Missal, M., Vogels, R., Li, C-Y., Orban, G., 1999. Shape Interactions in Macaque Inferior Temporal Neurons, J. Neurophysiol. 82 (No 1), pp. 131–142.

Müller, H., Rabbitt, P., 1989. Reflexive and voluntary orienting of visual attention: time course of activation and resistance to interruption. J. Exp. Psychol. Hum. Percept. Perform. 15, 315–330.

Norman, D., 1968. Toward a theory of memory and attention. Psych. Rev. 75, 522–536.

Nowlan, S., Sejnowski, T., 1995. A Selection Model for Motion Processing in Area MT of Primates, J. Neurosci. 15 (2), 1195–1214.

O'Connor, D., Fukui, M., Pinsk, M., Kastner, S., 2002. Attention modulates responses in the human lateral geniculate nucleus. Nat. Neurosci. 5 (11), 1203–1209.

Parodi, P., Lanciwicki, R., Vijh, A., Tsotsos, J.K., 1998. Empirically- derived estimates of the complexity of labeling line drawings of polyhedral scenes. Artif. Intell. 105, 47–75.

Pillsbury, W.B., 1908. Attention. Macmillan, New York.

Posner, M.I., Nissen, M., Ogden, W., 1978. Attended and unattended processing modes: the role of set for spatial locations. In: Pick Saltzmann (Ed.), Modes of Perceiving and Processing Information. Erlbaum, Hillsdale, NJ, pp. 137–158.

Raymond, J.E., et al., 1992. Temporary suppression of visual processing in an RSVP task: an attentional blink? J. Exp. Psychol. Hum. Percept. Perform. 18, 849–860.

Recanzone, G., Wurtz, R., Schwarz, U., 1997. Responses of MT and MST Neurons to One and Two Moving Objects in the Receptive Field, J. Neurophysiol. 78 (No. 6), 2904–2915.

Reynolds, J., Desimone, R., 1998. Interacting Roles of Attention and Visual Salience in V4. Neurophysiol 80, 2918–2940.

Reynolds, J., Desimone, R., 1999. The role of neural mechanisms of attention in solving the binding problem. Neuron 24, 19–29.

Reynolds, J., Chelazzi, L., Desimone, R., 1999. Competitive Mechanisms Subserve Attention in Macaque Areas V2 and V4, J. Neurosci. 19 (5), 1736–1753.

Rizzolatti, G., Riggio, L., Dascola, I., Umilta, C., 1987. Reorienting attention across the horizontal and vertical meridians — evidence in favor of a premotor theory of attention. Neuropsychologia 25, 31–40.

Rodriguez-Sanchez, A.J., Simine, E., Tsotsos, J.K., 2007. Attention and visual search. Int. J. Neural Syst. 17 (4), 275–288 Aug.

Roelfsema, P., 2006. Cortical Algorithms for Perceptual Grouping, Annu. Rev. Neurosci 29, 203–227.

Rolls, E., Aggelopoulos, N., Zheng, F., 2003. The Receptive Fields of Inferior Temporal Cortex Neurons in Natural Scenes. J. Neurosci., 23 (1), 339–348.

Rolls, E., Tovee, M., 1995. The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the receptive field. J. Exp. Brain Res. 103 (No 3), 409–420.

Roskies, A., 1999. The binding problem — introduction. Neuron 24, 7–9.

Rosenblatt, F., 1961. Principles of Neurodynamics: Perceptions and the Theory of Brain Mechanisms. Spartan Books.

Rothenstein, A., Tsotsos, J.K., 2006. Selective tuning: feature binding through selective attention. Int. Conf. Artificial Neural Networks, 10–14 Sept. 2006, Athens, Greece.

Rothenstein, A., Tsotsos, J.K., 2008. Attention links sensing with perception. Image Vis. Comput. 26, 114.

Rothenstein, A., Rodriguez-Sanchez, A., Simine, E., Tsotsos, J.K., in press. Visual feature binding within the selective tuning attention framework, Int. J. Pattern Recognition and Artificial Intelligence — Special Issue on Brain, Vision and Artificial Intelligence.

Salin, P., Bullier, J., 2008. Corticocortical Connections in the Visual System: Structure and Function, Physiol. Rev. 75 (1), 107–154.

Schoenfeld, M., Tempelmann, C., Martinez, A., Hopf, J.-M., Sattler, C., Heinze, H.-J., Hillyard, S., 2003. Dynamics of feature binding during object-selective attention. Proc. Natnl. Acad. Sci. 100 (20), 11806–11811.

Šetić, M., Domijan, D., 2007. A neural model for attentional modulation of lateral interactions in visual cortex. In: Mele, F., Ramella, G., Santillo, S., Ventriglia, F. (Eds.), Advances in Brain, Vision, and Artificial Intelligence. Springer, Berlin, pp. 42–51.

Sheinberg, D.L., Logothetis, N.K., 2001. Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. J. Neurosci. 21 (4), 1340–1350 February 15, 2001.

Shulman, G., Remington, R., McLean, J., 1979. Moving attention through visual space. J. Exp. Psychol. 92, 428–431.

Singer, W., 1999. Neuronal synchrony: a versatile code review for the definition of relations? Neuron 24, 49.

Sutherland, S., 1998. Feature selection. Nature 392, 350.

Thorpe, S., Fize, D., Marlot, C., 1996. Speed of processing in the human visual system. Nature 381, 520–522.

Tombu, M., Tsotsos, J.K., 2008. Attending to orientation results in an inhibitory surround in orientation space. Percept. Psychophys. 70 (1), 30–35.

Treisman, A., 1964. The effect of irrelevant material on the efficiency of selective listening. Am. J. Psychol. 77, 533–546.

Treisman, A., 1999. Solutions to the binding problem: progress through controversy and convergence. Neuron 24 (1), 105–125.

Treisman, A.M., Gelade, G., 1980. A feature-integration theory of attention. Cogn. Psychol. 12 (1), 97–136.

Treisman, A., Schmidt, H., 1982. Illusory conjunctions in the perception of objects. Cogn. Psychol. 14, 107–141.

Treue, S., Martinez-Trujillo, J., 1999. Feature-based attention influences motion processing gain in macaque visual cortex. Nature 399 (6736), 575–579.

Tsotsos, J.K., 1987. A'complexity level' analysis of vision Proceedings of International Conference on Computer Vision: Human and Machine Vision Workshop, London, England, June 1987.

Tsotsos, J.K., 1989. The complexity of perceptual search tasks. Proc. Int. Jt. Conf. Artif. Intell. Detroit 1571–1577.

Tsotsos, J.K., 1990. A complexity level analysis of vision. Behavi. Brain Sci. 13, 423–455.

Tsotsos, J.K., 1992. On the relative complexity of passive vs. active visual search. Int. J. Comput. Vis. 7 (2), 127–141.

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

Tsotsos, J.K., 2001. Motion understanding: task-directed attention and representations that link perception with action. Int. J. Comput. Vis. 45 (3), 265–280.

Tsotsos, J.K., 2005. Complexity and visual attention. In: Itti Rees Tsotsos (Ed.), Neurobiology of Attention. Elsevier/Academic Press.

Tsotsos, J.K., Culhane, S., Wai, W., Lai, Y., Davis, N., Nuflo, F., 1995. Modeling visual attention via selective tuning. Artif. Intell. 78 (1–2), 507–547.

Tsotsos, J.K., Culhane, S., Cutzu, F., 2001. From theoretical foundations to a hierarchical circuit for selective attention In: Braun, J., Koch, C., Davis, J. (Eds.), Visual Attention and Cortical Circuits. MIT Press, pp. 285–306.

Tsotsos, J.K., Liu, Y., Martinez-Trujillo, J., Pomplun, M., Simine, E., Zhou, K., 2005. Attending to visual motion. Comput. Vis. Image Underst. 100 (1–2), 3–40.

Tsotsos, J.K., Rodriguez-Sanchez, A., Rothenstein, A., Simine, E., Different binding strategies for the different stages of visual recognition, 2nd Int. Symposium Brain, Vision and Artificial Intelligence, Naples, Italy, Oct. 10–12, 2007.

VanRullen, R., Carlson, T., Cavanaugh, P., 2007. The blinking spotlight of attention. PNAS 104–49, 19204–19209.

von der Malsburg, Christoph, 1981. The correlation theory of brain function. Tech. Rep. 81–82 Biophysical Chemistry, MPI.

Wolfe, J., Cave, K., Franzel, S., 1989. Guided search: an alternative to the feature integration model for visual search. J. Exp. Psychol. Hum. Percept. Perform. 15, 419–433.

Wolfe, J.M., 1998. Visual search. In: Pashler, H. (Ed.), Attention. Psychology Press Ltd., Hove, UK, pp. 13–74.

Zaharescu, A., Rothenstein, A., Tsotsos, J.K., 2005. Towards a biologically plausible active visual search model, attention and performance in computational vision, Second International Workshop. Lect. Notes in Comput. Sci. Vol. 3368/2005, 133–147 Springer Berlin / Heidelberg.

Zoccolan, D., Cox, D., DiCarlo, J., 2005. Multiple Object Response Normalization in Monkey Inferotemporal Cortex. J. Neurosci. 25 (36), 8150–8164.

Zucker, S.W., Leclerc, Y., Mohammed, J., 1981. Continuous relaxation and local maxima selection — conditions for equivalence (in complex speech and vision understanding systems). IEEE Trans. Pattern Anal. Mach. Intell. vol. PAMI-3, 117–127.