What Roles can Attention Play in Recognition?

John K. Tsotsos, *Member, IEEE*

***Abstract*—Does attention have relevance for visual recognition and if so, under what circumstances? Is there a particular role (or roles) for attentive processes? These are not so simple to answer. Attention, if used at all in computer vision, has traditionally played one or both of the following roles: where to look next (or selection of region of interest), or top-down task influence on visual computation. In this paper, I argue that these are only two of the possible roles. Attention is also closely linked to binding and it is the triad of attention, binding and recognition that go hand in hand for non-trivial visual recognition tasks. This paper describes a set of four novel binding processes that employ a variety of attentive mechanisms to achieve recognition beyond the first feed-forward pass. The description is at a conceptual level with many pointers to papers where details may be found.**

***Index Terms*—attention, recognition, selective tuning, human vision, visual feature binding**

I. Introduction

Visual attention, recognition, and binding command a large, conflicting literature. For example, the nature of attentional influence has been debated for a long time. Among the more interesting observations are those of James' classic 1890 phrase [1] *everyone knows what attention is* juxtaposed with that of Pillsbury who in 1908 wrote [2] *attention is in disarray* and Sutherland's 1998 comment [3] *after many thousands of experiments, we know only marginally more about attention than about the interior of a black hole*. Within all of the current viewpoints, the only real constant seems to be that attention seems to be due to inherent limits in processing capacity in the brain [4]. But this does not constrain a solution. Even if we agree on a processing limit, how does it lead to the brain mechanisms that produce the observed phenomena?

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, easier problems. This paper suggests that considerations of time course can help carve seemingly monolithic problems into bite-sized pieces.

Many think 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?

Manuscript received April 14, 2008. This work was supported in part by the Canada Research Chairs Program and the Natural Engineering Research Council of Canada.

J. K. Tsotsos holds the Canada Research Chair in Computational Vision, and is with the Dept. of Computer Science & Engineering and with the Centre for Vision Research, York University, Toronto, Ontario, Canada M3J 1P3 (phone: 416-736-2100x70135; email: tsotsos@cse.yorku.ca).

Models with good performance of this first pass have appeared with hints about what happens next (e.g. [5,6,7]). 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 [4,8-12]. It is important to note that this tracing mechanism relies on a top-down maximum operation; as such it is inherently incompatible with the feed-forward max operations that most current models employ and is thus clearly distinguished from those. It should be emphasized that the experimental evidence against a feed-forward maximum operation is overwhelming[1]. 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 [13-20]. Additional studies have found the response to the preferred stimulus changes when presented along with other stimuli, a pattern inconsistent with a max operation [21,22]. A theoretical argument may also be made against a feed-forward max using the equivalence conditions between relaxation labeling processes and max selection [23], and especially considering the role of lateral processes in vision [24]. Lateral interactions necessitate a closer look at time course issues. 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 300ms after stimulus onset [25]. A single feed-forward pass ignores all but the first few V1 spikes.

II. Defining Vision Sub-tasks

Efforts to develop a computational theory of human vision must be informed by experimental observations of human (but also non-human primate) visual performance. Consequently, computational models of attention, recognition and binding

[1] I thank John Reynolds, Mazyar Fallah and Steven Zucker for discussion and literature pointers on this issue. Mazyar Fallah also provided additional general advice.

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 provide good definitions for many aspects of recognition and we can use these as a starting point [2] [26].

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. 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 (e.g., bungalows, split-level, other house types).

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 is matched by the third. *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. Additionally, responses can vary: 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.

More complex designs are also used; the point here is not to review all possibilities. Rather, the point is to present the definitions that we use in this paper. Further, if computational theories wish to have relevance to human vision, they need to consider such well-defined experimental procedures for each task when comparing their performance to experimental observations. It just does not seem right to take elements of experimental observations, model them, and then subject them to verification using different (sometimes wildly so) conditions and expect the comparison to be valid.

The need for a subject to respond 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 requested by the experimenter. In fact, 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, any of the above tasks that also include localization will be denoted by adding a superscript "*L*" to the task name.

Prior to performing any of the above tasks, subjects are provided with knowledge of the experiment, what to expect in terms of stimuli, what responses are required, and so on. In other words, subjects are 'primed' in advance for their task [27]. Thus, in any model of vision, the first set of computations to be performed is priming the hierarchy of processing areas. 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 100ms before stimulus onset to be effective [28]. This informs us that significant processing time is required for this step alone, a sufficient amount of time to complete a top-down traversal of the hierarchy before any stimulus is shown. This reflects one of the usual uses of attention referred to in the abstract.

III. Attention and Binding

Among the most misused and misunderstood concepts in perception are binding and attention. For the purposes of this paper, some concreteness is required.

The rationale for attentive processes is almost universally presented as a capacity limit with respect to processing power in the brain. Capacity limits naturally translate to considerations of computational complexity, because that is the discipline that examines the cost to achieving solutions to a problem in terms of a processing systems' capacity [4]. Problems are cast as search through a space of possibilities. In perceptual science, attention is often thought of as selection of portion of the input for preferential processing. It is really much broader and we maintain a view of attention as a *set* of mechanisms that optimize the search processes inherent in vision [29,30], a perspective that follows immediately from the capacity discussion above. These search mechanisms take many forms; perhaps the mixing of these many forms within any single experimental paradigm is a reason for why attention seems so inscrutable. Some search forms are: selection, (choosing from many); search space reduction (eliminating some of the possibilities when faced with a large search space); suppressing (improving signal-to-noise by reducing or eliminating the effect of one signal on another). There are several specific mechanisms within each category, many of which appear throughout this paper. Others include choice of world or task model, choice of viewpoint, and more [29,30].

A great deal of effort has gone into the discovery and elaboration of neural mechanisms that extract meaningful components from images 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. Current models develop representations of an image by proposing feature sets of varying complexity then resort to a classifier to make sense of these representations [5,6]. In Cognitive Science, this "Humpty-Dumpty" like task has been called the *binding problem* [31]. Binding is usually thought of as taking one sort of visual feature, such as a shape, and associating it associated

[2] I thank Allison Sekuler and Patrick Bennett for this pointer.

another feature, such as location, to provide a unified representation of an object. Such explicit association is 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" [32]. The binding literature is large; no attempt it made here to review it due to space limitations (see [33]).

Classical demonstrations of binding in vision 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 [4]. A single neuron receives converging inputs from many neurons and each provides input for many neurons. Precise location is lost in such a network of diverging feed-forward paths yet increasing convergence onto single neurons.

IV. CONNECTING RECOGNITION, ATTENTION AND BINDING

How is the right set of pathways through the visual processing network 'selected' 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. These are organized along the time dimension, i.e., the processing time each requires and the latency from stimulus onset observed for each

The first stage (leftmost element of Fig. 1), shows the priming stage. The attention processes involved include: suppression of task irrelevant features, stimuli or locations, and imposing the selectivity of a location cue or of a fixation point. The selection of task model and success criteria must also be completed. Then, the stimulus can be presented (the second element of Fig. 1).

The third element of Fig. 1 represents the one-interval *Discrimination Task* as long as no location information is required for a response (i.e., correspondence, detection, recognition, categorization, classification). Detecting whether or not a particular object is present in an image seems to take about 150ms [34]. The object recognition models cited throughout this paper as examples of modern recognition theories fall squarely within this task. This kind of 'yes-no' response can be called 'pop-out' in visual search with the added condition that the speed of response is the same regardless of number of distracters [35]. The categorization task also seems to take the same amount of time [36,37]. Interestingly, the median time required for a single feed-forward pass through the visual system is about 150ms [38]. Thus, many conclude that a single feed-forward pass suffices for this visual task. This first feed-forward pass is shown in the figure emphasizing the feed-forward divergence of neural connections and thus stimulus elements are spatially 'blurred' progressively more in higher areas of the hierarchy.

To provide more detail about a stimulus, such as for a within-category identification task, additional processing time, 65ms or so, is needed [36,37]; this is represented by the fourth from the left element of Figure 1. 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. These can be accessed by looking at which feature neurons feed into those category neurons. One way to achieve this is to traverse the hierarchy downwards, beginning with the category neuron and moving downwards through the needed feature maps[3]. 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 when full traversal is interrupted and not allowed to complete either because new stimuli enter the system before there is enough time for completion or not enough time is permitted due to other tasks. This results in the potential for localization errors and perhaps the well-known illusory conjunctions. These tasks will be termed *Identification Tasks*.

If additional localization is required for description or a motor task, (pointing, grasping, etc.), then the top-down traversal process must be allowed to sufficiently complete and additional time is required. These are the *Discrimination^{L} Tasks*, or simply, *Localization Tasks*. How much time? A lever press response seems to need 250-450ms in monkey [42]. During this task, the temporal pattern of attention modulation shows a distinct top-down pattern over a period of 35 - 350ms post-stimulus. The 'attention dwell time' needed for relevant objects to become available to influence behavior seems to be about 250ms [43]. Pointing to a target in humans seems to need anywhere from 230 to 360ms [44,45]. 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 and further experimental work is needed. Still, it seems that behavior, i.e., an action relevant to the stimulus, requires some degree of localization. The location details are available only in the early layers of the visual processing hierarchy because that is where the finer spatial resolutions of neural representation can be found. As a result, the top-down traversal initiated for the Identification Task must reach these early layers as shown in Fig. 1 (second element from the right). Examples, including difficult cases of segmentation of motion-defined form, are in [10,11,12].

The *Extended Discrimination Task* includes two-or-more interval designs [26], visual search, odd-man-out, resolving illusory conjunctions, determining transparency, recognizing objects in cluttered scenes, any task requiring sequences of saccades or pursuit eye movements, and more, e.g., [32,35,46,47]. The rightmost element of Fig. 1 depicts the start of a second feed-forward pass to illustrate this. It is likely that several iterations of the entire process, feed-forward and feedback, may be required to solve difficult tasks.

[3] This idea appeared first in Milner's 1974 paper [39], was used in Fukushima's 1986 attentive NeoCognitron [40], and appears in the 1997 Reverse Hierarchy Model of Ahissar & Hochstein [41]. Within the Selective Tuning model, it was first described in 1993 [8], with accompanying details and proofs in 1995 [9]. Only NeoCognitron and Selective Tuning provide realizations; otherwise, the two differ in all details.

The following sections relate each of these recognition stages to a specific binding process. These binding processes have been described elsewhere [48].

A. Convergence Binding

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. The attention process involved is search for maximum response. 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 [49,50]. 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.

Convergence binding requires that the image, a) contains no more than one copy of a given feature each at different locations; b) contains no more than one object/event each at different locations; and, c) does not contain objects/events that are composed of multiple features and share at least one feature type. Previous proposals for the binding problem [33] have not dealt with such constraints on problem definition.

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 Discrimination alone, the feed-forward pass can be repeated. In fact, 'waves' of stimuli continually flow through the system, but as each one passes through the full system, inspection of the results at the top of the hierarchy suffices. 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.*

B. Full Recurrence Binding

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 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 attention processes include top-down stimulus segmentation and localization and local max selection on the top-down traversal.

Full Recurrence Binding can determine the location and spatial extent of detected object/event for images that: a) contain more than one copy of a given feature each at different locations; b) contain more than one object/event each at different locations; and, c) contain objects/events that are composed of multiple features and share at least one feature.

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 projection of the receptive field 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 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, enhancing the overall signal-to-noise ratio for that neuron. This was first described in [4], the method for achieving it first described in [8] and fully detailed together with proofs of convergence and other properties in [9]. There is strong supporting evidence for an attentive suppressive surround [51]. Recent support for top-down selection has also appeared in a single-cell pre-frontal cortex study [52]. They show that in visual search tasks PFC is first to acquire the target, presumably as the strongest neural response. With human MEG evidence for recurrent processing responsible for the attentive suppressive surround [53], and the top-down timing of attentional effects throughout visual cortex [42], one can infer that a recurrent max-selection process begins in PFC.

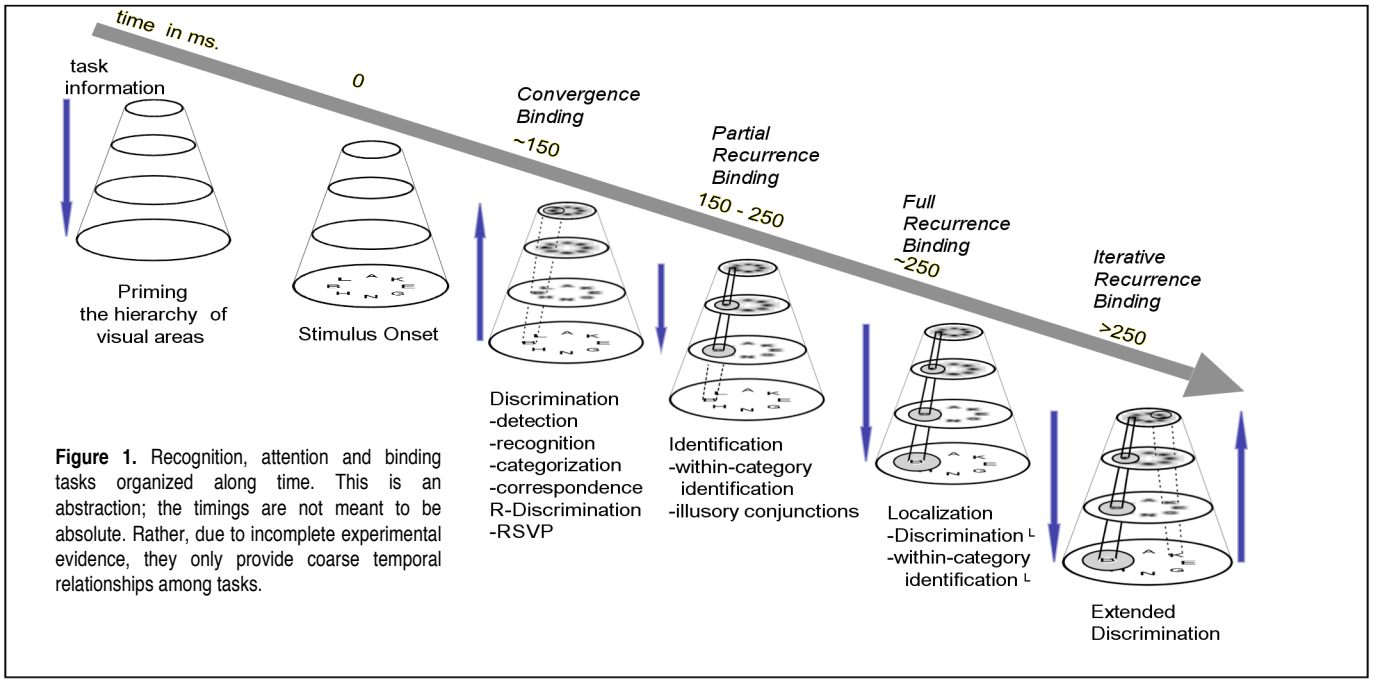

Figure 1. Recognition, attention and binding tasks organized along time. This is an abstraction; the timings are not meant to be absolute. Rather, due to incomplete experimental evidence, they only provide coarse temporal relationships among tasks.

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 but 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 [10]. There is no global representation of saliency required. Two variants on recurrence are now shown.

C. Partial Recurrence Binding

If the full recurrence binding process does not complete for any reason, this is called *Partial Recurrence Binding*. 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. Also, coarse localization tasks can be solved (such as 'in which quadrant is the stimulus?"). If this is not deployed directly due to task needs but is due to interruption, then this may result in illusory conjunctions. The attention process involved is top-down feature search guided by local max selection. A variety of different effects may be observed depending on when the top-down traversal the process is interrupted.

D. Iterative Recurrence Binding

Iterative Recurrence Binding is needed for *R-Discrimination*^{L} *Task*s and other more complex scenarios (call this class *Extended Discrimination*). 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. The attention mechanisms include sequences of convergence and recurrence binding, perhaps with task priming specific to each pass.

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 neurons in the system now see 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 [10,11].

V. Conclusion

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 [33]. 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. We view this as a first version of this decomposition and much effort remains to complete it. 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. 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: 1) top-down task directed priming before processing; 2) feed-forward traversal through the tuned visual processing hierarchy following the task-modulated neural pathways; 3) recurrent 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 feed-forward traversal.

These three basic steps are used in combination, and repeated, as needed to solve a given visual task. In simulation with artificial and real images, the model exhibits good agreement with a wide variety of experimental observations. A more detailed version appears in [54].

References

[1] James, W., (1890). **The Principles of Psychology**, H. Holt.
[2] Pillsbury W. B. (1908). **Attention**, New York: Macmillan.
[3] Sutherland, S., (1998). Feature Selection, *Nature* 392, 350.
[4] Tsotsos, J.K. (1990). A Complexity Level Analysis of Vision, *Behavioral and Brain Sciences* 13, 423-455.
[5] Serre, T., A. Oliva and T. Poggio , (2007). A Feedforward Architecture Accounts for Rapid Categorization, *Proc. National Academy of Sciences*, Vol. 104, No. 15, 6424-6429.
[6] Serre, T., L. Wolf, S. Bileschi, M. Riesenhuber and T. Poggio , (2007). Recognition with Cortex-like Mechanisms, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29, 3, 411-426.
[7] Fidler, S., Leonardis, A. (2007). Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts, IEEE CVPR 2007 Minneapolis MN, June 18-23
[8] Tsotsos, J.K. (1993). An Inhibitory Beam for Attentional Selection, in **Spatial Vision in Humans and Robots**, ed. by L. Harris and M. Jenkin, p313 - 331, Cambridge Univ. Press.
[9] Tsotsos, J.K., Culhane, S., Wai, W., Lai, Y., Davis, N., Nuflo, F. (1995). Modeling visual attention via selective tuning, *Artificial Intelligence* 78(1-2), 507 - 547.
[10] Tsotsos, J.K., Liu, Y., Martinez-Trujillo, J., Pomplun, M., Simine, E., Zhou, K. (2005). Attending to Visual Motion, *Computer Vision and Image Understanding* 100(1-2), 3 - 40.
[11] Rothenstein, A., Rodriguez-Sanchez, A., Simine, E., Tsotsos, J.K., Visual Feature Binding within the Selective Tuning Attention Framework, *Int. J. Pattern Recognition and Artificial Intelligence* (in press).
[12] Rodriguez-Sanchez, A.J., Simine, E., Tsotsos., J.K., Attention And Visual Search, *Int. J. Neural Systems*, 2007 Aug;17(4):275-88.

[13] Miller EK, Gochin PM, Gross CG. (1993). Suppression of visual responses of neurons in inferior temporal cortex of the awake macaque by addition of a second stimulus, *Brain Res.* Jul 9;616(1-2):25-9.

[14] Reynolds, J., Chelazzi, L., Desimone, R. (1999). Competitive Mechanisms Subserve Attention in Macaque Areas V2 and V4, *The Journal of Neuroscience*, 19(5):1736-1753.

[15] Missal, M., Vogels, R., Li, C-Y., Orban, G. (1999). Shape Interactions in Macaque Inferior Temporal Neurons, *The Journal of Neurophysiology* Vol. 82 No. 1, pp. 131-142.

[16] Recanzone, G., Wurtz, R., Schwarz, U. (1997). Responses of MT and MST Neurons to One and Two Moving Objects in the Receptive Field, *The Journal of Neurophysiology* Vol. 78 No. 6, pp. 2904-2915.

[17] Reynolds, J., Desimone, R. (1998). Interacting Roles of Attention and Visual Salience in V4, *J Neurophysiology* 80: 2918-2940, 1998.

[18] Chelazzi, L., Duncan, J., Miller, E., Desimone, R. (1998). Responses of Neurons in Inferior Temporal Cortex During Memory-Guided Visual Search, *The Journal of Neurophysiology* Vol. 80 No. 6, pp. 2918-2940.

[19] 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, *Journal Experimental Brain Research*, Vol 103, No 3, p.409-420.

[20] Zoccolan, D., Cox, D., DiCarlo, J., (2005). Multiple Object Response Normalization in Monkey Inferotemporal Cortex, *The Journal of Neuroscience*, 25(36):8150-8164.

[21] Sheinberg, D., Logothetis, N. (2001). Noticing Familiar Objects in Real World Scenes: The Role of Temporal Cortical Neurons in Natural Vision, *The Journal of Neuroscience*, 21(4):1340-1350.

[22] Rolls., E., Aggelopoulos, N., Zheng, F. (2003). The Receptive Fields of Inferior Temporal Cortex Neurons in Natural Scenes, *The Journal of Neuroscience*, 23(1):339-348.

[23] Zucker, S W l Leclerc, Y l Mohammed, J. (1981). Continuous relaxation and local maxima selection - Conditions for equivalence (in complex speech and vision understanding systems), *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 3, pp. 117-127.

[24] Ben-Shahar, O., Huggins, P., Izo, T., Zucker, S.W. (2003). Cortical connections and early visual function: intra- and inter-columnar processing, *J. Physiology-Paris*, Vol 97, No 2, pp. 191-208.

[25] 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*, Vol. 15, 843-856.

[26] Macmillan, N.A., Creelman, C.D., (2005**). Signal Detection Theory: A User's Guide**, Routledge.

[27] Posner, M. I. , Nissen, M., Ogden, W., (1978). Attended and unattended processing modes: The role of set for spatial locations, in Pick & Saltzmann, eds., **Modes of Perceiving and Processing Information**, 137-158, Hillsdale, NJ: Erlbaum.

[28] Müller, H., Rabbitt, P. (1989). Reflexive and Voluntary Orienting of Visual Attention: Time course of activation and resistance to interruption, *J. Exp. Psychology: Human Perception and Performance* 15, 315-330.

[29] Tsotsos, J.K., (1992). On the Relative Complexity of Passive vs. Active Visual Search, *International Journal of Computer Vision,* 7(2):127-141.

[30] Tsotsos, J.K., Motion Understanding: Task-Directed Attention and Representations that link Perception with Action, *International Journal of Computer Visio*n 45:3, 265-280, 2001.

[31] Rosenblatt, F., (1961). **Principles of Neurodynamics: Perceptions and the Theory of Brain Mechanisms.** Spartan Books.

[32] Treisman, A., Schmidt, H. (1982). Illusory conjunctions in the perception of objects, *Cognitive Psycholo*gy 14, 107-141.

[33] Roskies A.(1999). The Binding Problem - Introduction, *Neuron* 24, 7–9.

[34] Thorpe, S., Fize, D., Marlot, C. (1996). Speed of processing in the human visual system. *Natur*e 381, 520-522.

[35] Treisman, A. M., Gelade, G. (1980). A feature-integration theory of attention, *Cognitive Psychology* 12(1), 97-136.

[36] Grill-Spector, K., Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psych. Science* 16, 152-160.

[37] Evans, K., Treisman, A. (2005). Perception of Objects in Natural Scenes: Is It Really Attention Free?, *J. Experimental Psychology: Human Perception and Performance* 31-6, 1476-1492.

[39] Milner, P. (1974). A model for visual shape recognition, *Psych. Rev.* 81, 521-535.

[40] Fukushima, K.: A neural network model for selective attention in visual pattern recognition. *Biological Cybernetics* Vol 55:1 (1986) 5 - 15.

[41] Ahissar, M., Hochstein S. (1997). Task difficulty and the specificity of perceptual learning, *Nature*, 387(6631):401-6.

[42] 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.

[43] Duncan, J., Ward, J., Shapiro, K. (1994). Direct measurement of attentional dwell time in human vision, *Nature* 369, 313 - 315.

[44] Gueye, L., Legalett, E., Viallet, F., Trouche, E., Farnarier, G., (2002). Spatial Orienting of Attention: a study of reaction time during pointing movement, *Neurophysiologie Clinique* 32, 361-368.

[45] 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, *European J. Neuroscience* 17, 2481-2491.

[46] Wolfe, J. M. (1998). Visual Search. In H. Pashler (Ed.), **Attention** (pp. 13-74). Hove, UK: Psychology Press Ltd.

[47] 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. Nat.Acad.Sciences* 100(20), 11806-1181.

[48] Tsotsos, J.K., Rodriguez-Sanchez, A., Rothenstein, A., Simine, E., (2007). Different Binding Strategies for the Different Stages of Visual Recognition, **Advances in Brain, Vision, and Artificial Intelligence,** Lecture Notes in Computer Science Vol. 4729, Springer Berlin.

[49] Treisman, A. (1999). Solutions to the Binding Problem: Progress through Controversy and Convergence, *Neuron* 24:1:105-125.

[50] Reynolds, J., Desimone, R., (1999). The Role of Neural Mechanisms of Attention in Solving the Binding Problem, *Neuron* 24, 19–29.

[51] 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, *Proc. the National Academy of Sciences*, 103(4):1053-8.

[52] Buschman, T., Miller, E., (2007). Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices, *Science* 315, 1860.

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

[54] Tsotsos, J.K., Rodriguez-Sanchez, A., Rothenstein, A., Simine, E., (2008). Different Binding Strategies for the Different Stages of Visual Recognition, *Brain Research* , Available online 23 May 2008.

John K. Tsotsos received his Ph.D. in 1980 from the University of Toronto. He was on the faculty of Computer Science at the University of Toronto from 1980 to 1999. He

then moved to York University appointed as Director of York's Centre for Vision Research (2000-2006) and is currently Distinguished Research Professor of Vision Science in the Dept. of Computer Science & Engineering. He is Adjunct Professor in both Ophthalmology and Computer Science at the University of Toronto. Dr. Tsotsos has published many scientific papers, six conference papers receiving recognition. He currently holds the NSERC Tier I Canada Research Chair in Computational Vision. He has served on the editorial boards of Image & Vision Computing, Computer Vision and Image Understanding, Computational Intelligence and Artificial Intelligence and Medicine and on many conference committees. He served as General Chair for IEEE International Conference on Computer Vision 1999.