Localization of Attended Multi-feature Stimuli: Tracing Back Feed-Forward Activation Using Localized Saliency Computations

John K. Tsotsos

Dept. of Computer Science & Engineering, and, Centre for Vision Research
York University, Toronto, Ontario, Canada
tsotsos@cs.yorku.ca

Abstract. This paper demonstrates how attended stimuli may be localized even if they are complex items composed of elements from several different feature maps and from different locations within the Selective Tuning (ST) model. As such, this provides a step towards the solution of the 'binding problem' in vision. The solution relies on a region-based winner-take-all algorithm, a definition of a featural receptive field for neurons where several representations provide input from different spatial areas, and a localized, distributed saliency computation specialized for each featural receptive field depending on its inputs. A top-down attentive mechanism traces back the connections activated by feed-forward stimuli to localize and bind features into coherent wholes.

1 Introduction

Many models have been proposed to explain biological attentive behavior. Some have found utility in computer vision applications for region of interest detection. This article focuses on the Selective Tuning model (ST). Its 'first principles' foundations [1] provided the first formal justification for attention by focusing on computational complexity arguments on the nature of attention, the capacity of neural processes and on strategies for overcoming capacity limitations. The 'first principles' arise because vision is formulated as a search problem (given an image, which subset of neurons best represent image content?). This foundation suggests a specific biologically plausible architecture and its processing stages [1,2]. The architecture includes pyramid representations, hierarchical search and attentive selection.

This contribution focuses on how ST addresses the visual feature binding problem. This is a long-standing problem in cognitive science, first described by Rosenblatt [3]. In vision, as well as in other cognitive tasks, features such as an object's shape, must be correctly associated with other features to provide a unified representation of the object. This is important when more than one object is present in order to avoid incorrect combinations of features. Using the classical view of the binding problem, one can show that for a purely data-directed strategy the problem of finding the subsets of each feature map that correspond to the parts of an object has exponential complexity. It is an instance of the NP-Complete visual matching problem [4] so search is over the powerset of features and locations. In simple detection problems,

S. Kollias et al. (Eds.): ICANN 2006, Part II, LNCS 4132, pp. 471 – 480, 2006.

the complexity is manageable by a data-directed strategy because there are few features. In the general case, attentional selection is needed to limit the search.

Part of the difficulty facing research on binding is the confusion over definitions. For example, in Feature Integration Theory [5], location is a feature because it is assumed faithfully represented in a master map of locations. But, this cannot be true; location precision changes layer to layer in any pyramid representation. In any case, an object's edges do not share the same location with its interior. In the cortex, it is not accurate in a Euclidean sense almost anywhere, although the topography is qualitatively preserved [6]. The wiring pattern matters in order to get the right image bits to the right neurons. Thus binding needs to occur layer to layer because location coding changes layer to layer; it is not simply a high-level problem. In addition, features from different representations with different location coding properties converge onto single cells. The resulting abstraction of location information was shown to play an important role in the solution to complexity [1]. It also means that a binding solution requires recovery of location, as opposed to assuming it is a feature.

We define the binding task to involve the solution of three sub-problems: 1) detection (is a given object/event present?); 2) localization (location and spatial extent of detected object/event); and, 3) attachment (explicit object/event links to its constituent components). Further, binding is not a problem in simple situations and only appears when there is sufficient complexity in the image. Specifically, images must contain more than one copy of a given feature, each at different locations, contain more than one object/event each at different locations, and, contain objects/events composed of multiple features and sharing at least one feature type.

Others have proposed solutions to the feature binding problem. The Temporal Synchrony hypothesis proposes recognition of synchronized neural firing patterns [7, 8]. The Biased Competition model proposes task-biased inhibitory competition, plus the responses of higher-order neurons that encode only the attended stimuli, implicitly binds features [9]. The Saliency Map model proposes that feedback modulation of neural activity for visual attributes at the location of selected targets will suffice [10]. The difficulty with these proposals is that none present a mechanism to accomplish binding. There is even the view that recognition does not need attention for binding, and that attention is needed only for task priming and cluttered scenes [11]. This view can be rejected based on the timing observed in attentive tasks among different visual areas. As predicted by ST, higher-level areas show attentive effects before early ones. This is demonstrated in [12, 13] who show this latency pattern and show that attentive effects are mostly after 150ms from stimulus onset, the time period ignored in [11] and by those who study detection tasks exclusively.

The remainder of the paper will briefly present the ST model, and then overview the solution to binding. An example, a discussion of the limitations and behavioural predictions of the model, and a concluding discussion round out the paper.

2 The Selective Tuning Model

The details of the model have been presented previously ([1, 2, 14, 15, 16]) and thus only an overview sufficient to lead into the new work will be presented here.

2.1 The Model

The processing architecture is pyramidal, units 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 stimulus 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 first stage of 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. The result is a feed-forward, diverging cone of activity within the pyramid. 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 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 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.

2.2 ST's Winner-Take-All Process

ST's WTA is an iterative process realizable in a biologically plausible manner. 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 [2]. The time to convergence is specified by a simple relationship involving θ and the maximum possible value Z across all unit responses. The is because the partitioning procedure uses differences of values, and the smallest units will be inhibited by all other units while the largest valued units will not be inhibited by any unit. As a result, small units are reduced to zero quickly and the time to convergence is determined by the values of the largest and second largest units.

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:

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

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:

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

$G_i(0)$ is the feed-forward input to 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. Equation (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:

$$\Phi_{ij} = \mu\left(G_j(t) - G_i(t)\right)\left(1 - e^{-\frac{d_{ij}^2}{d_r^2}}\right) , \begin{cases} \text{if} \quad 0 < \theta < \mu\left(G_j(t) - G_i(t)\right)\left(1 - e^{-\frac{d_{ij}^2}{d_r^2}}\right) \\ 0 \qquad \text{otherwise} \end{cases} . \tag{3}$$

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.

3 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. This 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.

3.1 Featural Receptive Fields

For single feature maps or for the assumption of a single saliency map [5, 10] the hierarchical WTA described above will suffice. However, in our case, no such assumption is made. Saliency is not a global, homogeneous computation in this framework. A strategy for combining features from different representations and different locations is required. This requires the functionality provided by acknowledging the contributions to a neuron's response from separate locations and separate feature maps. 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

$$FRF = \bigcup_{j=1,k} f_j \quad , \tag{4}$$

where $\{f_j = \{(x_{j,a}, y_{j,a}), a=1,...,b_j\}, j=1,...,k\}$, (x,y) is a location in sub-field f_j, b_j is the number of units in sub-field f_j. The f_j's 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.

3.2 Hierarchical WTA Traces Back Feed-Forward Activations

The idea of tracing back connections in a top-down fashion was present, in part, in the Neocognitron model of Fukushima [17]; the first description of the ST hierarchical WTA method was presented in [16].

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 the top layer is populated by so-called grandmother cells whereas ST makes no such assumption. Overall, the Neocognitron model and its enhancements cannot scale and would suffer from representational and search combinatorics [1].

ST's WTA computation requires a competition among all the representations (feature maps) at the top layer of the pyramids, i.e., there can be multiple pyramids (such as ventral and dorsal stream). Task biases can weight each computation. The type of competition is determined by the relationships among the active representations. Two types are considered here. Two representations are mutually exclusive if, on a location-by-location basis, the two features 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). This also implies that the competing *FRF* sub-fields completely overlap in space. 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 colour, a line may be at some disparity, features belonging to eyes and co-exist with those from noses, etc.).

The following method is applied at the top of all pyramids at first, then recursively downwards following the *FRF* representations of the winning units. If *F* at some level of the hierarchy contains sub-fields from more than one feature map representing mutually exclusive features (call this subset *A*), then, the two WTA stages represented by Eqs. (1-3) are applied to each sub-field separately. This will yield a winning region within each sub-field, $g_f = \{(x_{i,f}, y_{i,f}) \mid i=1,2,...,n_f\}$, where n_f is the number of locations in the winning region in sub-field *f*. Call this a Type A process. Since the features are mutually exclusive, the winning feature region is the region with the largest sum of responses of its member units. This winning value V_A is given by

$$V_A = \max_{j \in F} \sum_{x,y \in g_j} r(j,x,y). \tag{5}$$

If *F* contains sub-fields representing features that can co-exist at each point (call this subset *B*), then the two stages of the WTA, represented by Eqs. (1-3), are applied to each representation separately. Here, however, the extent of the winning region is the union of all the winning regions. These winning regions are further constrained: each winning region is required to either overlap with, or to be entirely within or entirely enclose, another winning region. Call this the Type B process. The winning value is given by the sum of responses over all of the winning regions,

$$V_B = \sum_{j \in F} \sum_{x,y \in g_j} r(j,x,y) \; . \tag{6}$$

If *F* contains sub-fields representing features that are mutually exclusive (set *A*) as well as features that co-exist (set *B*), a combination of the above strategies is used. The winning value is given by the sum of Equations (5) and (6) and the extent of the winning region is the union of winning regions in sets *A* and *B*.

There is no saliency map in this model. Saliency is a dynamic and task-specific determination and one that may differ between processing layers as required. Further, this does not imply that a feature map must exist for any possible combination of features. Features are encoded separately in a set of maps and the relationships of competition or cooperation among them provide the potential for combinations. Although the above shows two forms of competition, other types can be included.

3.3 Detection, Localization and Attachment

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, the 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 *FRF*s all the way down the hierarchy. *FRF*s 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.

3.4 An Example

A very brief explanation of one example appears here. The full background for this example is available in [2, 14] and due to space limits, cannot be included here. The input is an image sequence that has three graphical objects (textured octagons) in motion; the largest item is translating and rotating, the medium sized object is rotating and the smallest object is translating. It satisfies the constraints for stimulus complexity requiring solution of the binding problem set out earlier. The visual processing hierarchy is specialized for motion alone and contains filter banks that simulate the motion selectivity of areas V1, MT, MST and 7a following experimental observations in the monkey. The V1 layer is selective for translation in 12 directions and 3 speeds (see Fig. 1a). MT, MST and 7a layers have pyramidal abstractions of this translation. MT also includes selectivity for the spatial derivative of local velocity (i.e., the representation is affine motion specific), in 12 gradient directions for each of the 12 directions and 3 speeds of translation. MST includes selectivity for generalized spiral motion (rotations, expansion, contraction and their combinations). 7a represents abstraction of generalized spiral as well as of translation. There are a total of 690 filter types in total (72 in V1, 468 in MT, 72 in MST and 78 in 7a), each operating over the visual field. Thus, there are multiple pyramids in this representation, with multiple top layers (78 top level representations). Generalized spiral neurons in MST have complex *FRF*s, building upon many features from the MT layer. However, there is no representation for the conjunction of translation with generalized spiral motion.

The feature binding process described earlier begins with a cooperative process (Type B) across the output layers: translation and generalized spiral motion can co-exist. This identifies the translation peak and the rotation peak belonging to the combined motion. The two winning regions then begin their downward search following their own *FRF*s. See Fig. 1b. The translation pyramid (on the right side of both sub-figures) needs competitive interaction (Type A) to select the best representations. The generalized spiral pyramid (on the left side of both sub-figures) also begins with competition at the top. Both pathways require competition layer by layer except for the MT layer of spatial derivatives. There, a Type B process is needed in order to find the set of common spatial gradients across a region (a rotating object has homogenous spatial derivatives to local velocity, the derivative direction being perpendicular to the direction of local motion). The winning attentional beams then split and converge on the image plane to show localization of the input stimulus (Fig. 1b). In the continuation of this example, this location would be inhibited to allow the second strongest input item to be found, and so on. The point of this example is to show how a motion not explicitly represented in the system can be found, detected and localized, involving many representations and locations bound by the attention beam.

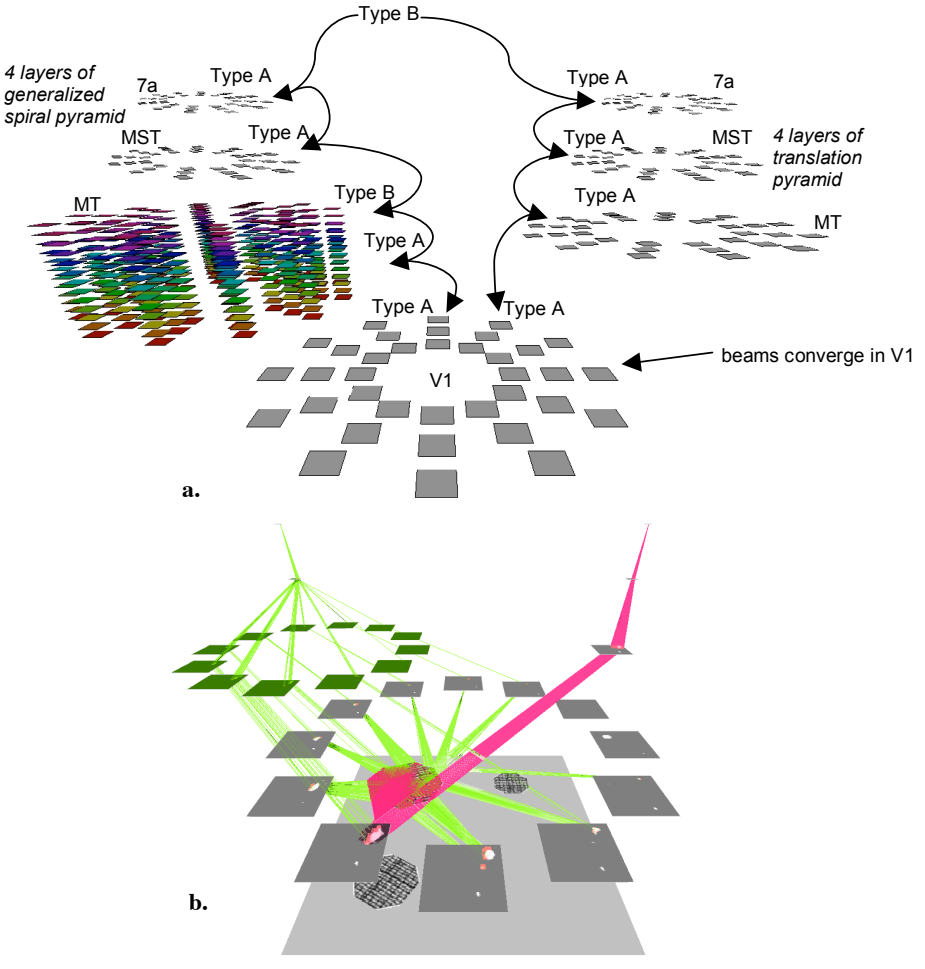

Fig. 1. a. The figure shows the full set of filter banks that are part of the motion processing system. Each small rectangle is one filter type at one pyramid layer. The rings of filters all appear at the same pyramid layer. The center rings of 36 filters show the output of V1; the translation pyramid then continues upwards on the right. The generalized spiral pyramid is on the left. V1 is their common base. The set of 432 coloured rectangles in MT depict the set of velocity gradient filters. The sequence of arrows shows the processing trail of the hierarchical WTA and the types of competition at each stage. **b.** The figure depicts the final configuration of attentive selection for the object that is translating and rotating even though no such feature conjunction has been included in the representation. The largest rectangle at the bottom represents the image plane on which are three textured octagons in motion. The other rectangles represent filter banks that contain the features of the attended stimulus. They are a subset of the full hierarchy of the left side figure with the inhibited ones removed. In some of the larger rectangles, response from the stimuli can be seen. The beams that tie together the filter bank representations are the pass zones of the attentional beam that converge on the largest of the 3 octagons. There are two roots to these beams because there is no single representation for rotating, translating objects.

4 Discussion

The solution to the feature binding problem has remained elusive for almost half a century. This paper proposes a solution and presents a very brief demonstration of the proposal. The binding problem was decomposed into three stages: detection, localization and attachment. The key element for its solution is the method for tracing connections that carry feed-forward activation downward through multiple representations so that they converge on the selected stimulus. This action links all the stimulus' component features within the pass zone of ST's attention beam.

The validation of such a model can be not only computational in the sense of performance on real images (however, see [14]). 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, among others, appeared in [1]. 1) Attention imposes a suppressive surround around attended items in space as well as in the feature dimension. 2) Selection is a top-down process where attentional guidance and control are integrated into the visual processing hierarchy. 3) The latency of attentional modulations *decreases* from lower to higher visual areas. 4) Attentional modulation appears wherever there is many-to-one, feed-forward neural convergence. 5) Topographic distance between attended items and distractors affects the amount of attentional modulation. In each of these cases, significance supporting evidence has accrued over the intervening years, recounted in [14, 15].

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.

The use of localized, distributed saliency within ST is precisely what the binding problem requires. Saliency is not a global, homogeneous process as in other models. Neurons in different representations that respond to different features and in different locations are selected together, the selection in location and in feature space, and are thus bound together via the pass zone of the attention mechanism. Even if there is no single neuron at the top of the pyramid that represents the concept, the WTA model allows for multiple threads bound through the spatial topology of the network wiring.

References

1. Tsotsos, J.K.: A Complexity Level Analysis of Vision. Behavioral and Brain Sciences Vol 13 (1990) 423-455
2. Tsotsos, J.K., Culhane, S., Wai, W., Lai, Y., Davis, N., Nuflo, F.: Modeling visual attention via selective tuning. Artificial Intelligence Vol 8:1-2 (1995) 507 - 547
3. Rosenblatt, F.: Principles of Neurodynamics: Perceptions and the Theory of Brain Mechanisms. Spartan Books (1961)
4. Tsotsos, J.K.: The Complexity of Perceptual Search Tasks. Proc. International Joint Conference on Artificial Intelligence Detroit (1989) 1571 - 1577
5. Treisman, A., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology Vol 12 (1980) 97-136
6. Felleman, D., Van Essen, D.: Distributed Hierarchical Processing in the Primate Visual Cortex. Cerebral Cortex Vol 1 (1991) 1-47
7. von der Malsburg, C.: The correlation theory of brain function, Internal Rpt. 81-2, Dept. of Neurobiology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany (1981)
8. Gray, C.M.,: The Temporal Correlation Hypothesis of Visual Feature Integration, Still Alive and Well. Neuron Vol 24:1, (1999) 31-47
9. Desimone, R., Duncan, J.: Neural mechanisms of selective visual attention. Annual Reviews of Neuroscience Vol 18 (1995) 193-222
10. Itti, L. Koch, C.: Computational modelling of visual attention. Nature Reviews Neuroscience Vol 2 (2001) 194-204
11. Riesenhuber, M. Poggio, T.: Are Cortical Models Really Bound by the "Binding Problem"? Neuron 1999, Vol 24:1 (1999) 87-93
12. Mehta, A. D. Ulbert, I. Schroeder, C. E.: Intermodal Selective Attention in Monkeys. I: Distribution and Timing of Effects across Visual Areas. Cerebral Cortex Vol 10:4, (2000) 343-358
13. Connor, D.O., Fukui, M., Pinsk, M., Kastner, S.: Attention modulates responses in the human lateral geniculate nucleus, Nature Neurosci.ence Vol 5:11, (2002) 1203–1209
14. Tsotsos, J.K., Liu, Y., Martinez-Trujillo, J., Pomplun, M., Simine, E., Zhou, K.: Attending to Motion, Computer Vision and Image Understanding Vol 100:1-2, (2005) 3 - 40
15. Tsotsos, J.K., Culhane, S., Cutzu, F.: From Theoretical Foundations to a Hierarchical Circuit for Selective Attention. Visual Attention and Cortical Circuits, (2001) 285 – 306, ed. by J. Braun, C. Koch, and J. Davis, MIT Press
16. Tsotsos, J.K.: An Inhibitory Beam for Attentional Selection. in Spatial Vision in Humans and Robots, ed. by L. Harris and M. Jenkin, (1993) 313 - 331, Cambridge University Press (papers from York University International Conference on Vision, June 1991, Toronto)
17. Fukushima, K.: A neural network model for selective attention in visual pattern recognition. Biological Cybernetics Vol 55:1 (1986) 5 - 15