3

Cognitive Vision Needs Attention to Link Sensing with Recognition

John K. Tsotsos

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

Abstract. "Cognitive computer vision is concerned with integration and control of vision systems using explicit but not necessarily symbolic models of context, situation and goaldirected behaviour" (Vernon 2003 [473]). This paper discusses one small but critical slice of a cognitive computer vision system, that of visual attention. The presentation begins with a brief discussion on a definition for attention followed by an enumeration of the different ways in which attention should play a role in computer vision and cognitive vision systems in particular. The Selective Tuning Model is then overviewed with an emphasis on its components that are most relevant for cognitive vision, namely the winner-take-all processing, the use of distributed saliency and feature binding as a link to recognition.

3.1 Towards a Definition of Attention

What is 'attention'? Is there a computational justification for attentive selection? The obvious answer that has been given many times that the brain is not large enough to process all the incoming stimuli, is hardly satisfactory (Tsotsos 1987 [456]). This answer is not quantitative and provides no constraints on what processing system might be sufficient. Methods from computational complexity theory have formally proved that purely data-directed visual search in its most general form is an intractable problem in any realization (Tsotsos 1989[457]). There, it is claimed that visual search is ubiquitous in vision, and thus purely data-directed visual processing is also intractable in general. Those analyses provided important constraints on visual processing mechanisms and led to a specific (not necessarily unique or optimal) solution for visual perception. One of those constraints concerned the importance of attentive processing at all stages of analysis: the combinatorics of search are too large at each stage of analysis otherwise. Attentive selection based on task knowledge turns out to be a powerful heuristic to limit search and make the overall problem tractable (Tsotsos 1990 [458]). This conclusion leads to the following view of attention: *Attention is a set of strategies that attempts to reduce the computational cost of the search processes inherent in visual perception.* It thus plays a role in all aspects of vision.

H.I. Christensen and H.-H. Nagel (Eds.): Cognitive Vision Systems, LNCS 3948, pp. 25–35, 2006.

Many (the active/animate vision researchers) seem to claim that attention and eye movements are one and the same; certainly none of the biological scientists working on this problem would agree. That one can attend to particular locations in the visual field without eye movements has been known since Helmholtz (1924 [186]), but eye movements require visual attention to precede them to their goal (Hoffman 1998 [192] surveys relevant experimental work). Active vision, as it has been proposed and used in computer vision, necessarily includes attention as a sub-problem.

3.2 Attention in Computer Vision

What is it about attention that makes it one of the easiest topics to neglect in computer vision? The task of tracking, or active control of fixation, requires as a first step the detection of the target or focus of attention. How would one go about solving this? Knowing that with no task knowledge and in a purely-data-directed manner, this sub-task of target detection is NP-Complete means that one is attempting to solve a problem that includes known intractable sub-problems. Is the problem thought to be irrelevant or is it somehow assumed away?

Those who build complete vision application systems invoke attentional mechanisms because they must confront and defeat the computational load in order to achieve the goal of real-time processing (there are many examples, two of them being Baluja & Pomerleau 1997 [27] and Dickmanns 1992 [100]). But the mainstream of computer vision does not give attentive processes, especially task-directed attention, much consideration.

A spectrum of problems requiring attention has appeared (Tsotsos 1992 [459]): selection of objects, events, tasks relevant for domain, selection of world model, selection of visual field, selection of detailed sub-regions for analysis, selection of spatial and feature dimensions of interest, selection of operating parameters for low level operations. Take a look at this list and note how most research makes assumptions that reduce or eliminate the need for attention:

- Fixed camera systems negate the need for selection of visual field.
- Pre-segmentation eliminates the need to select a region of interest.
- 'Clean' backgrounds ameliorate the segmentation problem.
- Assumptions about relevant features and the ranges of their values reduce their search ranges.
- Knowledge of task domain negates the need to search a stored set of all domains.
- Knowledge of which objects appear in scenes negates the need to search a stored set of all objects.
- Knowledge of which events are of interest negates the need to search a stored set of all events.

The point is that the extent of the search space is seriously reduced before the visual processing takes place, and most often even before the algorithms for solution are designed! However, it is clear that in everyday vision, and certainly in order to understand vision, these assumptions cannot be made. More importantly, the need for attention is broader than simply vision as the above list shows. It touches on the relevant aspects of visual reasoning, recognition, and visual context. As such, cognitive vision systems

should not include these sorts of assumptions and must provide mechanisms that can deal with the realities inherent in real vision.

3.3 The Selective Tuning Model (STM) of Visual Attention

The modeling effort described herein features a theoretical foundation of provable properties based in the theory of computational complexity (Tsotsos 1987, 1989, 1990, 1992 [456, 457, 458, 459]). The 'first principles' arise because vision is formulated as a search problem (given a specific input, what is the subset of neurons that best represent the content of the image?) and complexity theory is concerned with the cost of achieving solutions to such problems. This foundation suggests a specific biologically plausible architecture as well as its processing stages as will be briefly described in this article. A more detailed account can be found in (Tsotsos 1990 [458], Tsotsos et al. 1995 [460]).

3.3.1 The Model

The visual processing architecture is pyramidal in structure with units within this network receiving both feed-forward and feedback connections. When a stimulus is presented to the input layer of the pyramid, it activates in a feed-forward manner all of the units within the pyramid with receptive fields (RFs) mapping to the stimulus location; the result is a diverging cone of activity within the processing pyramid. It is assumed that response strength of units in the network is a measure of goodness-of-match of the stimulus within the receptive field to the model that determines the selectivity of that unit.

Selection relies on a hierarchy of Winner-Take-All (WTA) processes. WTA is a parallel algorithm for finding the maximum value in a set. First, a WTA process operates across the entire visual field at the top layer where it computes the global winner, i.e., the units with largest response (see Section 3.3.3 for details). The fact that the first competition is a global one is critical to the method because otherwise no proof could be provided of its convergence properties. The WTA can accept guidance to favor areas or stimulus qualities if that guidance is available, but operates independently otherwise. The search process then proceeds to the lower levels by activating a hierarchy of WTA processes. The global winner activates a WTA that operates only over its direct inputs to select the strongest responding region within its receptive field. Next, all of the connections in the visual pyramid that do not contribute to the winner are pruned (inhibited). The top layer is not inhibited by this mechanism. However, as a result, the input to the higher-level unit changes and thus its output changes. This refinement of unit responses is an important consequence because one of the important goals of attention is to reduce or eliminate signal interference (Tsotsos 1990 [458]). By the end of this refinement process, the output of the attended units at the top layer will be the same as if the attended stimulus appeared on a blank field. This strategy of finding the winners within successively smaller receptive fields, layer by layer, in the pyramid and then pruning away irrelevant connections through inhibition is applied recursively through the pyramid. The end result is that from a globally strongest response, the cause of that largest response is localized in the sensory field at the earliest levels. The paths remaining may be considered the

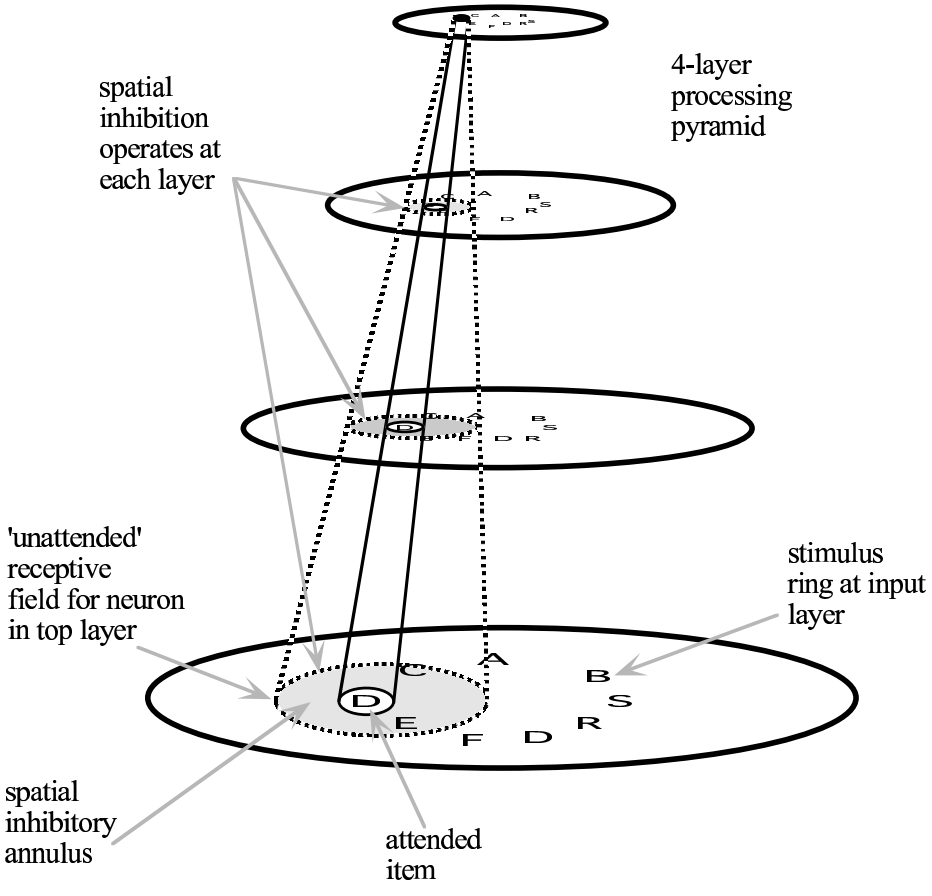

Fig. 3.1. Attentional beam

pass zone of the attended stimulus while the pruned paths form the inhibitory zone of an attentional beam. The WTA does not violate biological connectivity or relative timing constraints. Figure 3.1 gives a pictorial representation of this attentional beam.

An executive controller is responsible for implementing the following sequence of operations for visual search tasks:

1. Acquire target as appropriate for the task, store in working memory.
2. Apply top-down biases, inhibiting units that compute task-irrelevant quantities.
3. 'See' the stimulus, activating feature pyramids in a feed-forward manner.
4. Activate top-down WTA process at top layers of feature pyramids.
5. Implement a layer-by-layer top-down search through the hierarchical WTA based on the winners in the top layer.
6. After completion, permit time for refined stimulus computation to complete a second feed-forward pass. Note that this feed-forward refinement does not begin with the completion of the lowermost WTA process; rather, it occurs simultaneously with

completing WTA processes (step 5) as they proceed downwards in the hierarchy. On completion of the lowermost WTA, some additional time is required for the completion of the feed-forward refinement.
7. Extract output of top layers and place in working memory for task verification.
8. Inhibit pass zone connections to permit next most salient item to be processed.
9. Cycle through steps 4 - 8 as many times as required to satisfy the task.

This multi-pass process may seem to not reflect the reality of biological processes that seem very fast. However, it is not claimed that all of these steps are needed for all tasks. Several different levels of tasks may be distinguished, defined as:

Detection - is a particular item present in the stimulus, yes or no?
Localization - detection plus accurate location;
Recognition - localization plus accurate description of stimulus;
Understanding - recognition plus role of stimulus in the context of the scene.

The executive controller is responsible for the choice of task based on instruction. If detection is the task, then the winner after step 4, if it matches the target, will suffice and the remaining steps are not needed. Thus simple detection in this framework requires only a single feed-forward pass. If a localization task is required, then all steps up to 7 are required because, as argued in below, the top-down WTA is needed to isolate the stimulus and remove the signal interference from nearby stimuli. This clearly takes more time to accomplish. If recognition is the task, then all steps, and perhaps several iterations of the procedure, are needed in order to provide a complete description. The understanding task seems to fit the concept of cognitive vision best; however, the model described here does not include all of the required functionalities for this.

3.3.2 Top-Down Selection

STM features a top-down selection mechanism based on a coarse-to-fine WTA hierarchy. Why is a purely feed-forward strategy not sufficient? There seems to be no disagreement on the need for top-down mechanisms if task/domain knowledge is considered, although few non-trivial schemes seem to exist. Biological evidence, as well as complexity arguments, suggests that the visual architecture consists of a multi-layer hierarchy with pyramidal abstraction. One task of selective attention is to find the value, location and extent of the most salient image subset within this architecture. A purely feed-forward scheme operating on such a pyramid with:

i) fixed size receptive fields with no overlap, is able to find the largest single input with local WTA computations for each receptive field but location is lost and extent cannot be considered.
ii) fixed size overlapping receptive fields, suffers from the spreading winners problem, and although the largest input value can be found, the signal is blurred across the output layer, location is lost and extent is ambiguous.
iii) all possible RF sizes in each layer, becomes intractable due to combinatorics.

While case i) might be useful for certain computer vision detection tasks, it cannot be considered as a reasonable proposal for biological vision because it fails to localize

targets. Case iii) is not plausible as it is intractable. Case ii) reflects a biologically realistic architecture, yet fails at the task of localizing a target. Given this reality, a purely feed-forward scheme is insufficient to describe biological vision. Only a top-down strategy can successfully determine the location and extent of a selected stimulus in such a constrained architecture as used in STM.

3.3.3 WTA and Saliency

The Winner-Take-All scheme within STM is defined as an iterative process that can be realized in a biologically plausible manner insofar as time to convergence and connectivity requirements are concerned. The basis for its distinguishing characteristic comes from the fact that it implicitly creates a partitioning of the set of unit responses into bins of width determined by a task-specific parameter, θ. The partioning arises because inhibition between units is not based on the value of a single unit but rather on the absolute value of the difference between pairs of unit values. Further, this WTA process is not restricted to converging to single points as all other formulations. The winning bin of the partition, whose determination is now described, is claimed to represent the strongest responding contiguous region in the image (this is formally proved in Tsotsos et al. 1995 [460]).

First, the WTA implementation uses an iterative algorithm with unit response values updated by each iteration until convergence is achieved. Competition in an iteration depends linearly on the difference between unit strengths in the following way. Unit A will inhibit unit B in the competition if the response of A, denoted by $\rho(A)$, satisfies $\rho(A) - \rho(B) > \theta$. Otherwise, A will be inhibited by B. The overall impact of the competition on unit B is the weighted sum of all inhibitory effects, each of whose magnitude is determined by $|\rho(A)-\rho(B)|$. It has been shown that this WTA is guaranteed to converge, has well-defined properties with respect to finding strongest items, and has well-defined convergence characteristics (Tsotsos et al. 1995 [460]). The time to convergence, in contrast to any other iterative or relaxation-based method is specified by a simple relationship involving θ and the maximum possible value, Z, across all unit responses. The reason for this is that because the partitioning procedure uses differences of values. All larger units will inhibit the units with the smallest responses, while no units will inhibit the largest valued units. As a result the small response units are reduced to zero very quickly while the time for the second largest units to be eliminated depends only on the values of those units and the largest units. As a result, a two-unit network is easy to characterize. The time to convergence is given by

$$\log_2 \left(\frac{A - \theta}{A - B} \right)$$

where A is the largest value and B the second largest value. This is also quite consistent with behavioral evidence; the closer in response strength two units are the longer it takes to distinguish them.

Second, the competition depends linearly on the topographical distance between units, i.e., the features they represent. The larger the distance between units is, the greater the inhibition. This strategy will find the largest, most spatially contiguous subset within

the winning bin. A spatially large and contiguous region will inhibit a contiguous region of similar response strengths but of smaller spatial extent because more units from the large region apply inhibition to the smaller region than inhibit the larger region from the smaller one. At the top layer, this is a global competition; at lower layers, it only takes place within receptive fields. In this way, the process does not require implausible connectivity lengths. For efficiency reasons, this is currently only implemented for the units in the winning bin. With respect to the weighted sums computed, in practice the weights depend strongly on the types of computations the units represent. There may also be a task-specific component included in the weights. Finally, a rectifier is needed for the whole operation to ensure that no unit values go below zero. The iterative update continues until there is only one bin of positive response values remaining and all other bins contain units whose values have fallen below θ. Note that even the winning bin of positive values must be of a value greater than some threshold in order to eliminate false detections due to noise.

The key question is how is the root of the WTA process hierarchy determined? The following is a conceptual description of this, and not the iterative implementation of the WTA that depends on the process described in the previous paragraphs. The "max" function used below is implemented using the iterative process just described. Let F be the set of feature maps at the output layers overall, and $F^i, i = 1$ to n, be particular feature maps. Values at each x, y location within map i are represented by $M^i_{x,y}$. The root of the WTA computation is set by a competition at the top layers of the pyramid depending on network configuration (task biases can weight each computation).

To allow full generality, define a receptive field as the set of n contiguous locations $R = \{r_i = (x_i, y_i), i = 1...n\}$. The neuron receives input from these locations from an arbitrary set of other neurons, not necessarily from the same representation. Define the receptive field of a neuron as a set of arbitrarily shaped, contiguous, sub-fields

$$F = \{f_j = \{(x_{j,a}, y_{j,a}), a = 1...b_j\}, j = 1...k\},$$

such that

$$\bigcup_{j=1,k} f_j = R\,.$$

Each subfield is a retinotopic representation of a particular feature. The WTA competitions are defined on the subfields f_i . For spatially overlapping parts of these subfields, the features represented can be either mutually exclusive or can co-exist. The winning value is W, and this is determined by:

1. If $k = 1$, that is, there is only a single sub-field f,

$$W = \max_{x,y}\ M^f_{x,y}\ . \tag{3.1}$$

2. If F contains more than one sub-field, representing mutually exclusive features (subfields are fully overlapping in location), then

$$W = \max_{x,y}\left(\max_{i \in F}\ M^i_{x,y}\right)\ . \tag{3.2}$$

3. If F contains more than one sub-field, all fully overlapping in location, representing features that can co-exist at each point, then there is more than one WTA process, all rooted at the same location but operating through different feature pyramids

$$W = \max_{x,y} \left(\sum_{i \in F} M^i_{x,y} \right) . \tag{3.3}$$

4. If F contains sub-fields representing features that are mutually exclusive (the set A, as in case 2 above) as well as complementary (the set B, as in case 3 above), the winning locations are determined by the sum of the strongest response among set B (following method 3) plus the strongest response within set A (using method 2). Thus, a combination of the above strategies is used. There is more than one WTA process, all rooted at the same location but operating through (3.4)

$$W = \max_{x,y} \left[\sum_{b \in B} M^b_{x,y} \; + \; \max_{a \in A} \left(M^a_{x,y} \right) \right] . \tag{3.4}$$

For sub-fields or portions thereof that are not spatially overlapping with any other subfield, then the WTA process operates within that region following Rule 1.

As a result, there is no single saliency map in this model as there is in all other models. Indeed, there is no single WTA process necessarily, but several simultaneous WTA threads. Saliency is a dynamic, local, distributed and task-specific determination and one that may differ even between processing layers as required. Although it is known that feature combinations of high complexity do exist in the higher levels of cortex, the above does not assume that all possible combinations must exist. Features are encoded separately in a pre-defined set of maps and the relationships of competition or cooperation among them provide the potential for combinations. The above four types of competitions then select which combinations are to be further explored. This flexibility allows for a solution (at least in part) to binding issues.

The WTA process is implemented utilizing a top-down hierarchy of units. There are two main unit types: gating control units and gating units. Gating control units are associated with each competition in each layer and at the top, are activated by the executive in order to begin the WTA process. An additional network of top-down bias units can also provide task-specific bias if it is available. They communicate downwards to gating units that form the competitive gating network for each WTA within a receptive field. Whether the competition uses Eqs. 1, 2, 3, or 4 depends on the nature of the inputs to the receptive field. Once a particular competition converges, the gating control unit associated with that unit sends downward signals for the next lower down competition to begin. The process continues until all layers have converged.

The model has been implemented and tested in several labs applying it to computer vision and robotics tasks. The current model structure is shown in Figure 3.2. The executive controller and working memory, the motion pathway (V1, MT, MST, 7a), the peripheral target area PO, the gaze WTA and gaze controller have all been implemented and examples of performance can be found in (Culhane & Tsotsos 1992 [90], Wai & Tsotsos 1994 [482], Tsotsos et al. 1995 [460], Tsotsos et al. 2002 [461]).

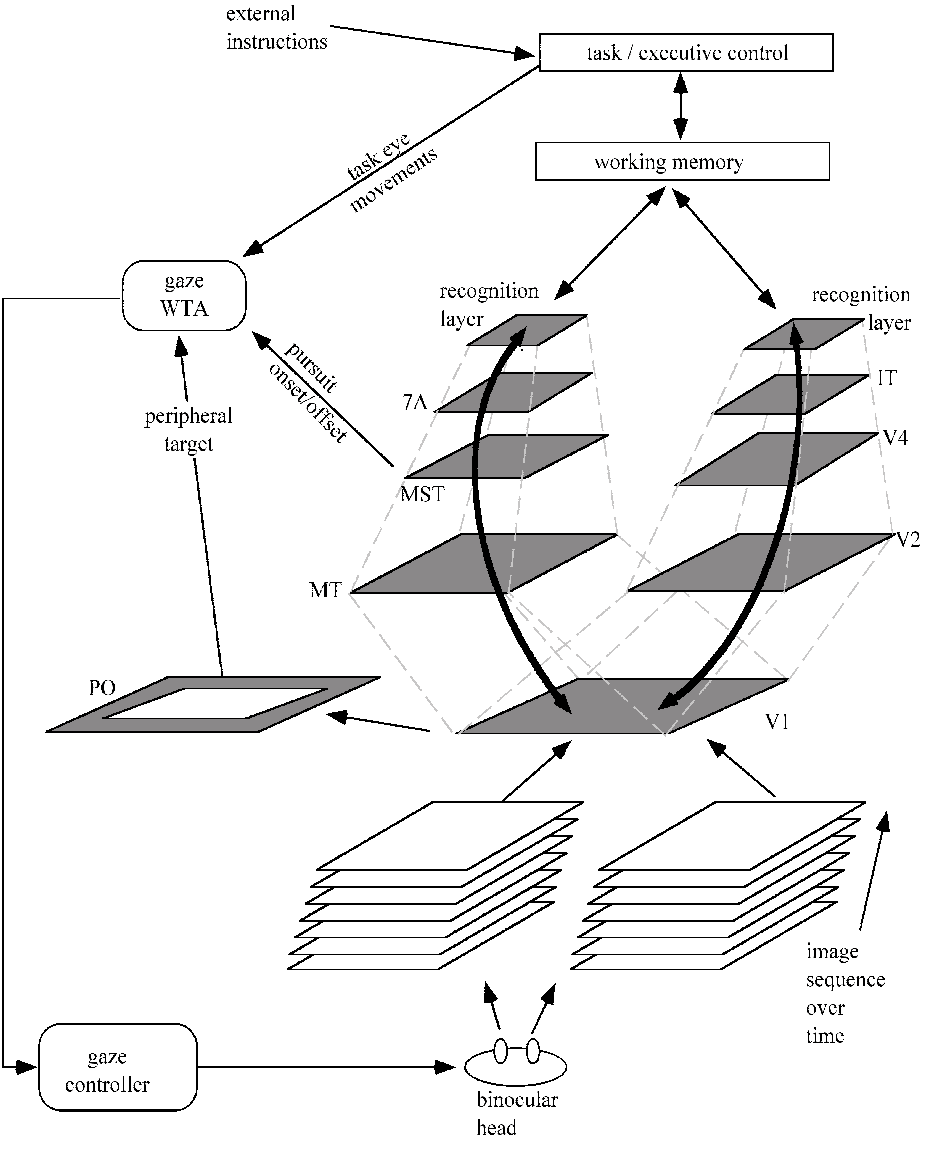

Fig. 3.2. The full model

3.3.4 Feature Binding: The Link to Recognition

A major contribution of the demonstration of how STM can operate within a complex visual hierarchy is the method of grouping features (known as the binding problem in computational neuroscience, Roskies 1999 [381]) into wholes. It is not claimed that this particular strategy has sufficient generality to solve all possible issues within the binding

problem. Nevertheless, it is the first instance of such a solution and further work will investigate its generality.

Following Roskies, "the canonical example of binding is the one suggested by Rosenblatt in which one sort of visual feature, such as an object's shape, must be correctly associated with another feature, such as its location, to provide a unified representation of that object". Such explicit association 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 [455]). At least some authors (Ghose & Maunsell 1999 [146], von der Malsburg 1999 [481]) suggest that specialized neurons that code feature combinations (introduced as cardinal cells by Barlow 1972 [28]) may assist in binding. The STM solution does indeed include such cells; however, they do not suffice on their own as will be described because they alone cannot solve the localization problem.

Using the classical view of the binding problem, it is straightforward to 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, Tsotsos 1989 [457]). In simple detection problems the complexity is manageable by simple strategies because there are not too many choices and the task is simply detection of a target. However, in the general case, a top-down attentional selection mechanism is needed to reduce the complexity of the search problem. It is for this reason that attention constitutes the link between sensing and recognition.

The use of localized saliency and WTA decision processes is precisely what the binding problem requires: neurons in different representations that respond to different features and in different locations are selected together, the selection being in location and in feature space, and are thus bound together via the 'pass' zone(s) of the attention mechanism. Even if there is no single neuron at the top of the pyramid that represents the concept, the WTA allows for multiple threads bound through location by definition in Eq. 1–4.

Part of the difficulty facing research on binding is the confusion over definitions and the wide variety of tasks included in binding discussions. For example, in Feature Integration Theory (Treisman and Gelade 1980 [454]), location is a feature because it assumes it is faithfully represented in a master map of locations. But this cannot be true; location precision changes layer to layer in any pyramid representation. In the cortex, it is not accurate in a Euclidean sense almost anywhere, although the topography is qualitatively preserved (Felleman & Van Essen 1991 [118]). The wiring pattern matters in order to get the right image bits to the right neurons. Thus binding needs to occur layer to layer and is not simply a problem for high-level consideration. Features from different representations with different location coding properties converge onto single cells and this seems to necessitate an active search process.

For the purposes of this argument, consider the following:

1. Location is not a feature, rather, it is the anchor that permits features to be bound together. Location is defined broadly and differently in each visual area and in practice is considered to be local coordinates within a visual area (think of an array of hypercolumns, each with its own local coordinates);

2. A grouping of features not coincident by location cannot be considered as a unitary group unless there is a unit to represent that group;
3. Features that compose a group may be in different locations and represented in different visual areas as long as they converge onto units that represent the group;
4. If the group is attended, then the WTA of Section 3.3.3 will find and attend to each of its parts regardless of their location or feature map representation.

This strategy is sufficient to handle complex recognition tasks such as multiple patterns, overlapping objects, or even transparent motions. As such, it is a solution to the aspect of binding that attends to groups and finds parts of groups.

3.4 Discussion

How can attentional selection be integrated into a cognitive computer vision system? It is certainly true that most if not all such systems have some early vision processing stages. STM provides a skeleton within which one can include layers of early vision filters and organize them into meaningful hierarchies. It is also true that somewhere in the processing stages, the need to segment an image into regions or events is important. STM's selection strategy may assist with this. If a target object is specified in advance, STM can be shown this in advance, that particular image can be processed and then stored in working memory, and used to guide visual search making search for that target more efficient than without guidance.

It is straightforward to 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. In simple detection problems the complexity is manageable by simple strategies because there are not too many choices and the task is simply detection of a target. However, in the general case, a top-down attentional selection mechanism is needed to reduce the complexity of the search problem. Thus, attention is an important link connecting sensing and recognition for realistic images and world models.

It is clear that STM can provide selection of visual field, selection of detailed subregions for analysis, selection of spatial and feature dimensions of interest, and selection of parameters for low-level operations. It cannot select relevant objects or events from a knowledge base with respect to a particular task, select tasks relevant for a domain, select the world model appropriate for solving the current task, and so on. In other words, the machinery described seems appropriate for early and intermediate levels of visual processing but has not yet advanced to a stage to be as useful for higher levels of visual processing or for the task levels of processing. These must remain topics for future research.