Hierarchical, Localized Saliency Computations Solve the Visual Feature Binding Problem

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

Abstract

Computational vision has a long history of proposing methods for decomposing a visual signal into components. What has been far more elusive is how to recombine those components into a whole, a problem known as the binding problem. *Although several proposals have appeared, the approaches and their demonstrations seem weak at best. This paper proposes a novel solution for a significant portion of the binding problem, namely, the re-combination of visual features into larger patterns and their localization in the image. The solution requires the abandonment of the nearly ubiquitous single saliency map and the adoption of a hierarchical, localized computation of saliency that is dependent on local neural selectivity constraints. This strategy has been demonstrated within a fully implemented model that attends to simple motion patterns in image sequences.*

1. Introduction

As described by Roskies [7], "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" [6, 22]. Such explicit association ("binding") is particularly important when more than one visual object is present, in order to avoid incorrect combinations of features belonging to different objects, otherwise known as "illusory conjunctions" [16]. At least some authors [3, 15] suggest that specialized neurons that code feature combinations (introduced as cardinal cells by Barlow [1]) may assist in binding. The solution in this paper 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.

This contribution presents the strategy used by the Selective Tuning Model to solve binding of features into wholes. It has been demonstrated using motion patterns (the complete motion system appears elsewhere [13, 14]). It is not claimed that this particular strategy has sufficient generality to solve all possible issues within the binding problem; however it seems to solve the limited cases that occur in real image sequences of simple motion patterns. As such, it is the first instance of such a solution and further work will investigate its generality.

Before the details of the binding solution may be presented an overview of the neural computation machinery that acts as the foundation must be overviewed.

2. The Selective Tuning Model (STM) of Visual Attention

STM features a first-principles, theoretical foundation of provable properties based in the theory of computational complexity [8, 9, 10, 11]. 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 [10, 12, 14]).

2.1 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 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 below 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 major goals of attention is to reduce or eliminate signal interference [10]. 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 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 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 hierarchy 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.

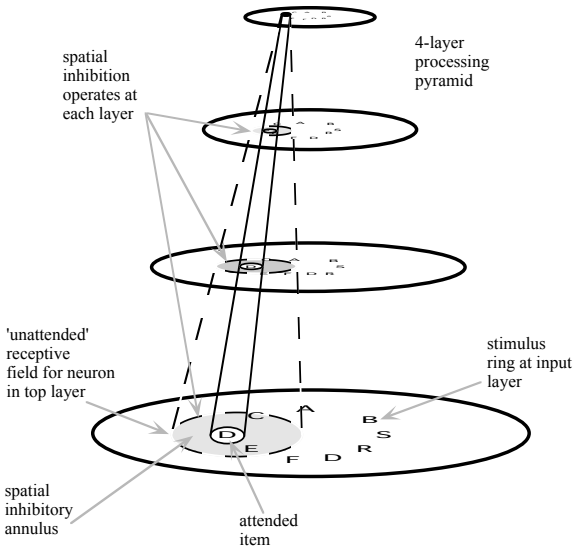

Figure 1. Attentional beam

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 Section 2.2, 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 has similar requirements, although this is not within the scope of the model at this point.

2.2 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. Saliency and Hierarchical, Local Winner-Take-All Computations

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 partitioning 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 include the representation of the strongest responding contiguous region in the image. Formal performance proofs appear in [12].

First, the WTA implementation uses an iterative algorithm with unit response values updated with 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 not affect 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 [12]. 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 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. 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 following equations define the sets of neurons that participate in WTA competitions and in what way. The "max" function used below is implemented using the iterative process just described. The definition for the competition at the output layers is described first followed by the competitions at all subsequent layers.

At the top level WTA, the competition is global, across the entire visual field as argued earlier. Task biases enter the process here by inhibiting task irrelevant features/units. Let F be the set of feature maps at the output layers, and F^i, i=1 to n, be particular feature maps. There is no requirement that there exists a single output representation. Values at each x,y location within map i are represented by $M^i_{x,y}$. Features are either mutually exclusive (the set A) at each location across the entire visual field or can co-exist (the set B). The winning units are those with value defined by

$$W = \max_{x,y}\left[\sum_{b\in B} M^b_{x,y} + \max_{a\in A}\left(M^a_{x,y}\right)\right] \quad (1)$$

The units in the winning bin then activate WTA's across their inputs, and those competitions are governed as follows. To allow full generality, define a receptive field R as a 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. The receptive field may be composed of a set S of k arbitrarily shaped, contiguous, possibly overlapping location sub-fields, S = $\{f_j = \{(x_{j,a}, y_{j,a})$, a=1..$b_j\}$, j=1...k$\}$, such that $\bigcup_{j=1,k} f_j = R$. Sub-fields are defined based on the different features that the neuron requires to be selective for some object/event. The size and shape of the sub-fields are set by the variability in position for each object/event feature. Each sub-field connects to a retinotopic representation of a single particular feature. This does not restrict features to adjacent layers of the hierarchy; rather, they may be in any other appropriate representation. The WTA competitions are defined on the sub-fields f_i. For spatially overlapping parts of these sub-fields, the features represented can be either mutually exclusive or can co-exist and a separate WTA is set up for those regions. As a result the number of separate WTA competition threads is the number of sub-fields in the RF plus the number of overlapping regions among them. The winning value in each case is represented by W, (i.e., the value of the winning bin described above), and this is characterized by:

1. For one sub-field that represents feature f then

$$W = \max_{x,y} M^f_{x,y} \quad (2)$$

2. If z sub-fields partially overlap in some single region, then there are z possible features for each location in that overlap region. Either all z features can co-exist at each point or they are all mutually exclusive at each point or some can co-exist (set B) while others are mutually exclusive (set A). If all are mutually exclusive, then

$$W = \max_{x,y}\left(\max_{i\in F} M^i_{x,y}\right) \quad (3)$$

If all can co-exist, then

$$W = \max_{x,y}\left(\sum_{i\in F} M^i_{x,y}\right) \quad (4)$$

If there is a combination, then Rule 1 applies.

3. For sub-fields that are fully overlapping each representing different features, then rule 1 applies.

The winning values determined by these rules are grouped into the same value bin and into the same spatially contiguous unit using the method described earlier in this section. Those units representing those winning values are thus bound together.

It is clear that there is no single saliency map in this model as there is in most other models. Indeed, there is no single WTA process necessarily, but several simultaneous WTA threads. Saliency is a dynamic, local, hierarchical 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.

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.

4. Feature Binding

What is described here through 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 [24] 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 [2]. 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.

4.1 Binding Strategy

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 will find and attend to each of its parts regardless of their location or feature map representation.

This is a solution to the aspect of binding that attends to groups and finds parts of groups. In the demonstrations shown in detail in [13] and [14], the groups are motion patterns. There are several components to this solution for motion. The first has to do with the particular representations chosen for motion patterns. Our representation is hierarchical with each layer being defined using components from the previous. For example, a constant speed, rotating object exhibits constant velocity gradient across location with respect to local motion. A neuron higher in the hierarchy then can be selective to regions that are homogeneous for this value and this is an easy selectivity to define and implement. A motion pattern detector in layer MST simply sums responses of the corresponding MT units that feed it. Consider a simple rotating textured square. In layer MT, neurons sensitive to local motion direction within the object select gradients perpendicular to that direction. Across all directions in the representation, the responses of the neurons make it appear as if the object has been 'cut into pie pieces', one for each local motion direction. That is, the tuning properties of the neurons have decomposed the flow field into distinct areas of constant velocity gradient. These are also partitioned depending on speed. Then, at the MST layer, the neurons whose selectivity is for rotation within each particular speed band will receive input from these MT representations (and not from the others). The MST neuron whose receptive field is best centered on the object will fire strongest if it receives sufficient stimulation, which in this case means that it sees all pieces of the pie. That best responding neuron can now be considered as having grouped the pie pieces and re-assembled the pie, that is, to have bound together the representations at the MT layer which otherwise are neither co-incident by location nor feature type. This is the feed-forward part of this process - an *implicit* binding action. If the task of the system were to simply detect the presence of a particular motion pattern, this representation would suffice as long as the top-level global WTA selects this region (this is the aspect of binding that models such as Reisenhuber and Poggio [17] address and thus erroneously conclude binding and attention are not required). However, if the system's task is to localize or recognize, then the job is not complete. There are many MST neurons that respond. The feedback process of top-down attention selects the best of these responses, and actively sub-selects the particular regions of MT neurons that correspond to that best firing, and thus best fitting the pattern selectivity of the neuron. The unique aspect here is that the receptive field of the MST neuron is defined by a spatial region as well as a subset of features computed within that spatial region, each feature contributing a component across that spatial region. This shows the need for a more flexible view on saliency and WTA computations than has been previously shown in other models. The binding is thus

complete both in feature as well as location dimensions. No other model currently includes such a distributed definition of saliency and in fact the bulk of models follow the lead of Koch and Ullman [4] and use a single overall map.

What if a more complex binding problem is considered, one where multiple motion patterns appear in an image sequence? The strategy presented works equally well for multiple objects, separated in space or even overlapping and examples are shown in [14].

4.2 Illusory Conjunctions

STM allows a clear demonstration of how illusory conjunctions may arise [16]. Following the experiment of Treisman and Schmidt carefully, two separate stimuli are attended (black coloured digits), the two stimuli being separated by some retinal distance that is occupied by other stimuli (coloured shapes or letters). The observation is that if attention is directed to the separated stimuli due to some task, then subjects make errors in reporting what they see in the space in between, those errors being conjunction errors such as assigning the colour of one stimulus to the other or the shape of one to another. How can this be explained by STM?

This locus of attentional interest is imposed on the model from task instructions (as it would be in the original experiment). This can be accomplished within STM by first showing a cue image with two cue locations, the same size and colour of the targets (the black digits). Task instruction would include the fact that both must be attended - it does not matter for the purposes of this experiment where they be attended simultaneously or in sequence, however, for the diagrams only the simultaneous version is shown.

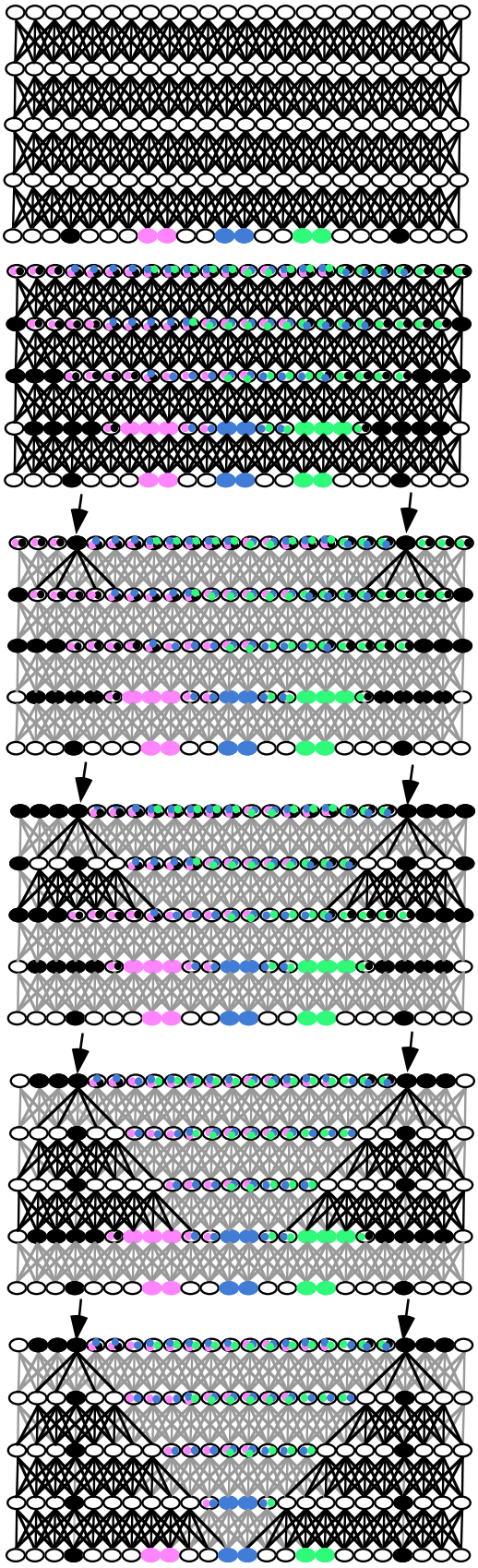

Figure 2. The STM explanation for illusory conjunctions. The top element shows a hypothetical visual processing network, with 5 stimulus items in the input layer, similar to the test stimuli in [16]. The next four show the sequence of network changes as a result of attention beginning with the network feed-forward activation, and then the application of the location foci followed by the deployment of the attentional beams. The bottom element shows the final configuration of responses once attention has been fully deployed.

As can be seen in the sequence of processing steps of Figure 2, the two black stimuli can be easily localized, while the coloured stimuli in between them lead to a confused representation at the top. No neuron in the top layer represents only one of those stimuli; each is activated by a combination and thus a query about those stimuli cannot possibly be answered correctly.

The true test of this explanation rests with future experiments that test two predictions that arise here. The first prediction has to do with the spacing of the stimuli. The second deals with the frequency of error types. As can be seen in the diagrams, if the two black items are sufficiently close in location, their inhibitory attention zones may overlap causing the intervening stimuli to appear to disappear. This would be true only if both were simultaneously attended, not sequentially. An experiment that tests the dependence of illusory conjunction performance retinal separation would illuminate this? Second, as is seen by the distribution of colours at the top of the hierarchy, the most common error is to confuse a characteristic of the middle stimulus with the others. In fact, the distribution of possible conjunctions in this example is: pink-blue-black - 2; pink-blue - 1, pink-green-blue - 10, blue-green - 1, blue-green-black - 2. Therefore it might be possible to collect statistics on specific conjunction errors to see if such a pattern that is strongly dependent on relative location is borne out.

5. Discussion

The binding mechanism described above differs significantly from all prior proposals, not the least difference being that of an experimental demonstration. It is not assumed away as by Reisenhuber & Poggio [17] by solving only a detection task, and it is not postponed to some unspecified computation involving the recognition of synchronized neural firing patterns [21, 22]. As well stated by Shadlen & Movshon [23], the synchronized firing theory is incomplete in that it describes the signature of binding without detailing how binding is computed. A valiant attempt at developing a detailed model and simulation of temporal synchrony solution to binding is due to Hummel & Biederman [19]. Their scheme used a constraint propagation paradigm on early neural representations (lines, endpoints, etc.). The set of co-activated neurons based on local grouping constraints fires in phase and is considered as arising from a single geon. This was a good idea to try but suffers from a few problems. First, the authors sensibly distinguish their work from that of closely related line-labeling efforts in computer vision. It is true they do not consider labelling on the lines. However the other part of the problem that is also part of line-labelling is to determine whether or not the overall figure is a physically realizable one; this overall problem is known to be exponential in nature [5] for polyhedral scenes. Since geons are a more general class of objects than polyhedra one may surmise that the complexity involved is at least as great. They also neglect to note that constraint propagation will not necessarily yield unique hypotheses for each element (that is why the line-labelling part of the problem is so important); they have no subsequent computation to deal with this. Finally, even though they assume that propagation time is very fast so that all the neurons fire in phase, this necessarily has limits and can only be considered reasonable for objects whose retinal representation has very small spatial extent.

Reynolds & Desimone [20] have a view that attentional mechanisms eliminate illusory conjunctions by filtering out unattended stimuli whose features could be mistakenly conjoined with those of the attended stimulus. They say that this selection process occurs in several stages and depends on attentionally induced increases in the effective salience of the attended object in earlier stages of processing, where it appears alone within a receptive field. As signals from multiple stimuli progress forward into higher-order areas with larger receptive fields, stimuli compete to control neuronal responses. The added strength of the signals from the attended stimulus resolves this competition in its favor. As a result, the responses of higher-order neurons with large receptive fields encode only the attended stimulus, implicitly binding together its features. Although this explanation by Reynolds & Desimone has many elements in common with the STM explanation, it relies too strongly on the bottom-up nature of their biased competition perspective; it further has not been shown to actually work in practice. The arguments presented in Section 2.2 work against this perspective.

The STM explanation is consistent with the views on binding presented in Itti & Koch [18], although those authors did not specify any mechanism that might accomplish binding using feedback. They do however say that it may be done by a feedback modulation of neural activity for the visual attributes and at the location of desired or selected targets.

The main contribution of this paper rests with the specification of how a distributed saliency mechanism might function. It begins with the key observation that all features are not necessarily mutually complimentary nor should they necessarily be participants in a single, globally defined saliency computation on a location-by-location basis. Features are components of the definition of objects or events on an individual basis; some may be totally irrelevant, some may co-exist at the same location, some may be mutually exclusive at a given location, some may be replicated identically or with minor variations over several locations, and so on. A set of rules was presented that, when used in

conjunction with the hierarchical WTA process of STM, allow for the representation of the variations in saliency definition required by individual objects or events. It is argued here (and demonstrated experimentally elsewhere [12, 14]) that this mechanism suffices for the complex binding tasks inherent in the detection, localization and recognition of motion patterns. The strategy also makes concrete non-intuitive predictions that warrant further investigation. The first prediction has to do with the spacing of the stimuli. The second deals with the frequency of error types. One additional implication concerns the locus of saliency representations. Evidence has been found for the representation of saliency in almost every visual area in brain. This may be because, according to the STM strategy, salience is a computation that appears within each area through the WTA process wherever there is a many-to-one neural convergence.

Roskies describes the breadth of the binding problem and gives examples of the different types [7]. These types span visual binding, auditory binding, binding across time, cross-modal binding, cognitive binding of a percept to a concept, cross-modal identification and memory reconstruction (the linking of previously encoded information to form a structured representation). There is no claim here that the STM strategy for binding was designed to address each of these, although it does appear appropriate for a significant subset within the visual binding group. It is clear that much more work and new insights are needed before binding is completely understood.

Acknowledgements
Funding for this research was gratefully received from the Natural Sciences and Engineering Council of Canada, Communications and Information Technology Ontario, a Province of Ontario Centre of Excellence, and from the Institute for Robotics and Intelligent Systems, A Network of Centers of Excellence of the Government of Canada. JKT holds the Canada Research Chair in Computational Vision.

References

[1] Barlow, H.B. (1972). Single units and cognition: a neurone doctrine for perceptual psychology, *Perception 1*, 371–394.

[2] Felleman, D., Van Essen, D. (1991). Distributed Hierarchical Processing in the Primate Visual Cortex, *Cerebral Cortex 1*, p 1-47.

[3] Ghose, G., Maunsell, J. (1999). Specialized Representations Review in Visual Cortex: A Role for Binding?, *Neuron*, Vol. 24, 79–85, September.

[4] Koch, C., Ullman, S. (1985). Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry, *Hum. Neurobiology 4*, p 219 - 227.

[5] Kirousis, L., Papadimitriou, C. (1988). The Complexity of Recognizing Polyhedral Scenes, *J. of Computer and System Sciences* 37, 14-38.

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

[7] Roskies, A. (1999). The Binding Problem Review Introduction, *Neuron*, Vol. 24, 7–9, September

[8] Tsotsos, J.K. (1987). A `Complexity Level' Analysis of Vision, Proceedings of International Conference on Computer Vision: London, England.

[9] Tsotsos, J.K. (1989). The Complexity of Perceptual Search Tasks, Proc. International Joint Conference on Artificial Intelligence, Detroit, pp1571 - 1577.

[10] Tsotsos, J.K. (1990). Analyzing vision at the complexity level, *Behavioral and Brain Sciences 13-3*, p423 - 445.

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

[12] Tsotsos, J.K., Culhane, S., Wai, W., Lai, Y., Davis, N., Nuflo, F. (1995). Modeling visual attention via selective tuning, *Artificial Intelligence 78*, p507-547.

[13] Tsotsos, J.K., Pomplun, M., Liu, Y., Martinez-Trujillo, J., Simine, E., (2002) Attending to Motion: Localizing and Labeling Simple Motion Patterns in Image Sequences, Conference on Biologically-Motivated Computer Vision, Tuebingen, Germany.

[14] Tsotsos, J.K., Liu, Y., Martinez-Trujillo, J., Pomplun. M., Simine, E., Zhou, K. (2004). Attending to Visual Motion, (pre-print available at www.cs.yorku.ca/~tsotsos)

[15] von der Malsburg, C., (1999). The What and Why of Binding: Review The Modeler's Perspective, *Neuron*, Vol. 24, 95–104.

[16] Treisman, A., Schmidt, H. (1982). Illusory conjunctions in the perception of objects, *Cognitive Psychology*, 14, 107–141.

[17] Riesenhuber M, Poggio T (1999). Are cortical models really bound by the 'binding problem'? *Neuron*, 24:87-93.

[18] Itti, L., Koch, C. (2001). Computational Modelling of Visual Attention, *Nature Neuroscience Reviews*, 2, p194 - 203.

[19] Hummel, J. E., Biederman, I. (1992). Dynamic binding in a neural network for shape recognition, *Psychol. Rev.* 99, 480–517.

[20] Reynolds, J. H., Desimone, R. (1999). The role of neural mechanisms of attention in solving the binding problem, *Neuron* 24, 19–29.

[21] Singer W., Gray CM. (1995). Annu. Rev. Neurosci 18, 555 - 586.

[22] von der Malsburg, C. (1981). The correlation theory of brain function, MPI Biophysical Chemistry, Internal Report 81–2.

[23] Shadlen, M., Movshon, A. (1999). Synchrony Unbound: Review A Critical Evaluation of the Temporal Binding Hypothesis, *Neuron*, 24, 67–77.

[24] Treisman, A., and Gelade, G. (1980). A feature-integration theory of attention, *Cogn. Psychol. 12*, 97–136.