Selective Tuning: Feature Binding Through Selective Attention

Albert L. Rothenstein and John K. Tsotsos

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

Abstract. We present a biologically plausible computational model for solving the visual binding problem. The binding problem appears due to the distributed nature of visual processing in the primate brain, and the gradual loss of spatial information along the processing hierarchy. The model relies on the reentrant connections so ubiquitous in the primate brain to recover spatial information, and thus allow features represented in different parts of the brain to be integrated in a unitary conscious percept. We demonstrate the ability of the Selective Tuning (ST) model of visual attention [1] to recover spatial information, and based on this propose a general solution to the binding problem. The solution is demonstrated on two classic problems: recovery of form from motion and binding of shape and color. We also demonstrate how the method is able to handle difficult situations such as occlusions and transparency. The model is discussed in relation to recent results regarding the time course and processing sequence for form-from-motion in the primate visual system.

1 Introduction

Convergent evidence from many different kinds of studies suggests the visual cortex is divided into a large number of specialized areas processing different feature dimensions, organized into two main processing streams, a dorsal pathway, responsible for encoding motion, space, and spatial relations for guiding actions, and a ventral pathway, associated with object recognition and classification, conclusions supported by functional imaging, neurophysiology, and by strikingly selective localized lesions. This high selectivity of the various cortical areas has led to the obvious questions of how, despite this specialization, the visual percept is unitary, and what are the mechanisms responsible for, in effect, putting all this distributed information together. Following Roskies [2], “the canonical example of binding is the one suggested by Rosenblatt [3] 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” [4]. Limiting the resources available for visual processing through increased loads and/or reduced time leads observers to erroneously associate basic features present in the image into objects that do not exist, e.g. a red X and a blue O are sometimes reported as a blue O and a red X. Studies have shown that these are real conjunction errors, and can not be attributed to guessing or memory. A general discussion of the binding problem appears in Neuron 24(1) (1999).

Three classes of solutions to the binding problem have been proposed in the literature. Proponents of the *convergence* solution suggest that highly selective, specialized neurons that explicitly code each percept (introduced as cardinal cells by Barlow [5] – also known as gnostic or grandmother neurons) form the basis of binding (e.g. [6, 7]). The main problem with this solution is the combinatorial explosion in the number of units needed to represent all the different possible stimuli. Also, while this solution might be able to detect conjunctions of features in a biologically plausible network (i.e. a multi-layer hierarchy with pyramidal abstraction) it is unable to localize them in space on its own [8], and additional mechanisms are required to recover location information. *Synchrony*, the correlated firing of neurons, has also been proposed as a solution for the binding problem [9–11]. Synchrony might be necessary for signaling binding, but is not sufficient by itself, as it is clear that this phenomenon can at most tag bound representations, but not perform the binding process. The *colocation* solution proposed in the Feature Integration Theory (FIT) [12] simply states that features occupying the same spatial location belong together. Due to its purely spatial nature, this solution can not deal with transparency and other forms of spatial overlap. Also, since detailed spatial information is only available in the early areas of the visual system, simple location-based binding is agnostic of high-level image structure, which means that it can not impose boundaries (obviously, the different edges of an object occupy different spatial locations), and arbitrary areas that belong to none, one or more objects can be selected.

Selective Tuning (ST) [1] is a computational model of visual attention that integrates feedforward and feedback pathways into a network that is able to take high level decisions, and, through a series of winner-take-all (WTA) processes, identify all the neurons that have participated in that decision. This identification satisfies the key requirement for a kind of visual feature binding that ST was demonstrated to solve [13], despite the loss of spatial information inherent in a pyramidal system. The ST feedback process does not need collocation if neural convergence is guaranteed, so ST is able to select all parts of a stimulus, even if they do not share a location (e.g. stimuli with discontinuities due to overlap, or stimuli that are separated spatially due to the nature of the cortical feature maps). The partial solution to binding proposed in [13] is able to correctly bind all the activations that have contributed to a high level decision (*convergence*) and even non-convergent representations if the problem can be solved at the spatial resolution of the top level of the pyramid (a weak form of *collocation*) – i.e. there is sufficient spatial separation between the target and the distractors (see [14] for the importance of spatial separation in attention and recognition). Note that the feedback process will select only the units responding to the selected stimulus, and not units that just happen to share locations with it, thus ensuring that overlapping and transparent stimuli will be handled correctly.

2 Proposed Solution

This section motivates and introduces an original approach to the binding problem. FIT [12] considers location as a feature that is faithfully represented in a "master map" of locations but, as Taraborelli [15] points out: "the idea of binding itself is nothing but a spatial conjunction of information concerning visual attributes of the same item." Tsotsos et al. [13] note that considering location as a feature can not be valid as location precision (resolution) changes layer to layer in any pyramid representation, and propose that location should be considered as the anchor that permits features to be bound together. At the same time, Robertson lists three phenomena that demonstrate the special role of spatial attention in binding [16]: illusory conjunctions under divided attention, dependence on number of distractors for conjunction searches, and the elimination of the influence of distractors with spatial cueing. In effect, a solution to the binding problem must address this seemingly incompatible requirement: binding is ultimately only a spatial conjunction, but at the same time it must be based on high-level information, allowing for object and feature-based selection.

The solution proposed is based on the general *collocation* principle of FIT, but using Selective Tuning to integrate high-level representations in the spatial selection process, and performing the spatial selection without requiring a "master map" of locations. The proposal, illustrated in Fig. 1 is to allow one feedforward pass through the network (arrow A in the figure), detect and select one salient high-level representation (in this case, one motion representation), and proceed backwards through the system in Selective Tuning manner (arrow B), selecting compatible representations that have contributed to the winning units, and inhibiting all the activations that are incompatible. As this feedback proceeds, lower level representations that participated in the salient activation are selected, and distractors inhibited, all the way to the first layer of processing. This allows further feedforward processing to be selectively directed towards the selected object (arrow C), eliminating the influence of spatially near competing stimuli and allowing the ventral pathway to detect the shape corresponding to the motion signal. When processing ends, the remaining active high-level representations all describe the selected stimulus in a sparse, distributed fashion ideal for maximizing the capacity of associative memories [17]. At the same time all the components of the attended stimulus will be selected throughout the visual system for recognition, and the location information can be used for the planning of actions towards the selected stimulus.

3 Examples

The general structure of the neural network used in the following examples is presented in Fig. 1, consisting of two biologically inspired processing path-

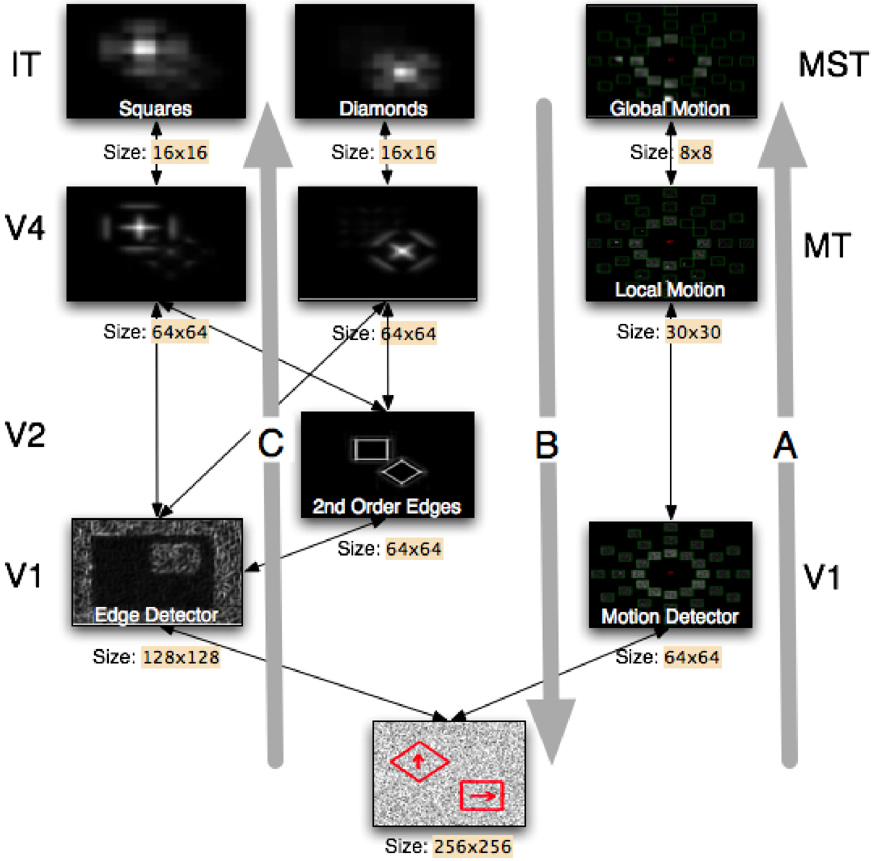


Fig. 1. Diagram of the network. On the left side is the shape recognition pathway, while on the right side the motion pathway. The arrows show the flow of information. See the text for details.

ways, corresponding to the ventral and dorsal visual pathways. The pathways are pyramidal, meaning that successive layers represent more and more abstract concepts, and location, size and (direct) shape information is lost. The motion pathway recognizes affine motions, and is described in detail in [13]. The form pathway is an abstraction of the primate object recognition stream, and consists of layers that combine the edge information to detect simple geometric shapes.

As shape and motion are processed in different areas of the brain, the recovery of shape from motion is a particularly good illustration of binding. The subset of the motion processing hierarchy in Fig. 1 consists of a layer of motion sensitive neurons, followed by two layers of translation detection neurons, corresponding to visual areas V1, MT (local motion) and MST (global motion), respectively. The simplified shape processing hierarchy, detecting square and diamond shapes, consists of two layers of first and second order edge detectors (four directions, one scale), and two layers of shape detectors, corresponding to visual areas V1 and V2 (edges), V4 (local shape) and IT (global shape), respectively. All the weights in the neural network are preset, no learning was used. Our system will process the image in parallel, along all the independent processing pathways, detecting the presence of the different shapes and motion patterns. The attentional process will select one top-level representation for further analysis, and the ST process will localize the corresponding pixels in the input image through feedback. ST will also inhibit pixels in the surround of the attended item, thus enhancing the relative saliency of the attended stimulus and introducing the contour information needed by the shape pathway. A second feedforward pass through the pyramids will refine the representation of the selected object, and at the same time select all the (distributed) representations that belong to it, thus achieving binding. The process can be repeated until the desired target is found, implementing a visual search mechanism. In the following experiments, the stimuli consists of random dot kinematograms with the dots in one or two windows performing translation motion in one or two different directions. The window can be square or diamond shaped – Fig. 2.

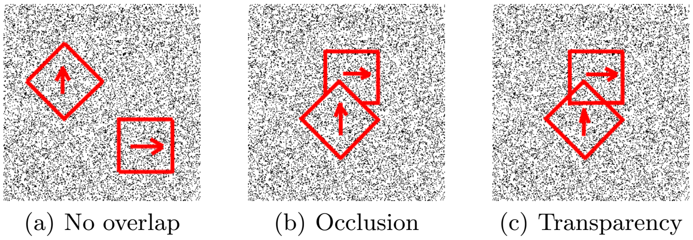

(a) No overlap (b) Occlusion (c) Transparency

Fig. 2. Random dot kinematograms used as stimuli in the form-from-motion experiments. Dots in one or two windows perform translation motion. The window can be square or diamond shaped.

3.1 Example 1 – Form from motion

The stimulus consists of a random dot kinematogram with the dots in a square window performing translation motion to the right. After the initial feedforward and feedback passes, the moving dots will be localized in the input layer, and neighboring dots will be inhibited, as illustrated in Fig. 3(a). Once the neighboring neurons have been inhibited, the V1 edge detectors will detect these pseudo-edges, and the shape recognition pathway will become activated, detecting the presence of the square in the input sequence. Similarly, Fig. 3(b) illustrates the result of detecting motion in a diamond shaped window. To test the capabilities of the attentional system to deal with multiple input patterns we used an image sequence containing two moving regions: a square region of rightward moving dots and a diamond shaped region of upward moving dots. MST neurons will fire indicating the two incompatible motion patterns, and the attentional system will select the top level movements one after the other, allowing the system to detect first one, then the other shape – Fig. 3(c) and Fig. 3(d).

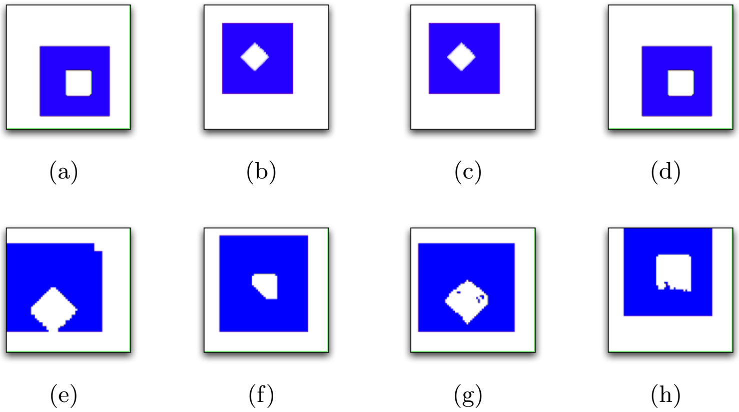


Fig. 3. Attentional selection windows for random-dot kinematograms. (a) Single stimulus, Square (b) Single stimulus, Diamond (c) Two objects, fixation 1 (d) Two objects, fixation 2 (e) Occluder, fixation 1 (f) Occluded, fixation 2 (g) Transparent, fixation 1 (h) Transparent, fixation 2. The internal white area represents the localized stimulus, blue (dark) the inhibitory surround.

An important test that must be passed by any artificial vision system is its handling of occlusions and transparency. The two stimulus sequences from Example 1 were modified by making the two stimuli overlap spatially, either through occlusion or through transparency.

3.2 Example 2 – Occlusion

In this case, the diamond partially occludes the square. The process follows as illustrated above for the occluder – Fig. 3(e), but it is important to observe that for the occluded stimulus only the visible portion is selected Fig. 3(f). This is a very important point that highlights a key difference between ST and the Neocognitron [18] system. While in both systems, feedback "tunes" the processing pyramid, the latter increases the activation of units that correspond to the selected hypothesis by reducing their firing thresholds, thus introducing the risk of hallucinations. In ST the only manipulation permitted is the reduction of the activations of units that do not match the hypothesis, and so the risk of hallucination is eliminated, and if the hypothesis turns out to be incorrect, the responses of the selected output units should decrease, indicating the failure to find support for the hypothesis. This example clearly shows that based on the detected motion patterns, the object recognition pathway will get as input a correct representation of the shape present in the stimulus.

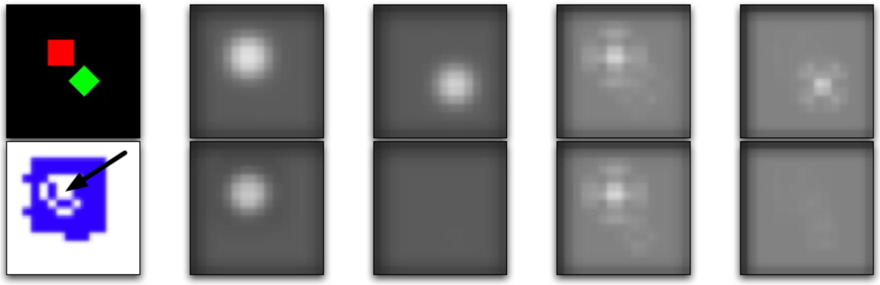
3.3 Example 3 – Transparency

In this case, the square partially overlaps the diamond, but the moving dots in the diamond shaped window remain visible. Again, the high level detection combined with the ST feedback process are able to highlight the correct areas in the input, thus allowing the object recognition pathway to correctly recognize the shapes – Fig. 3(g) and 3(h). While not an issue in this example, note that in some cases humans perceive two overlapping and different motion patterns as a single motion in a third direction (e.g. plaid motion), functionality currently missing from our motion model. Due to the fact that the ST process selects all inputs that have contributed to a high level decision as belonging to the stimulus, we expect the system to function correctly once this functionality is added.

3.4 Example 4 – Binding color and shape

The binding problem is quite often illustrated with images consisting of different geometric shapes, each of a different color. Similar to the previous examples, we will use red and green diamond and square objects (see Fig. 4(a) top), and we add a color detection pyramid (simple Gaussian blurring and downsampling). The system will initially detect the presence of the different shapes and colors by processing the whole image in parallel – Figs. 4(b) - 4(e) top. The attentional WTA process will select one top-level representation (the red representation, in this case), the ST process will localize the corresponding pixels in the input image and inhibit all nearby pixels, thus enhancing the relative saliency of the attended stimulus.

A second pass through the pyramids will refine the representation of the selected object, and select all the (distributed) representations that belong to it, while the green and diamond representations are strongly inhibited, thus achieving binding – Figs. 4(b) - 4(e) bottom. Fig. 4(a) bottom represents the



(a) (b) Red (c) Green (d) Square (e) Diamond

Fig. 4. Effects of attentional selection on colored shape stimuli. See text for details.

difference between the activation of the red detector with and without attention, and it can be observed that in the attended condition the representation that was initially distributed (the dark inhibited area) is much more focused, as indicated by the arrow. The initial representation corresponded to all the items in the field, but as a result of attention only the representations corresponding to red square remain active, making binding possible.

4 Discussion

While the importance of space in binding is captured in the Feature Integration Theory, high level representations, object- and feature-based attention mechanisms are not easily integrated into FIT. In this paper we have presented an original solution to the binding problem in visual perception, by recovering spatial information from high level representations through Selective Tuning style feedback. Another important contribution of this research is a process of recovering spatial information that does not require a “master map” of locations or any other separate representation of space. We have demonstrated this solution through a number of examples, including the difficult cases of occlusion and transparency. While these preliminary results are encouraging, the representations used (especially the shape recognition pathway) are very simplistic, and significant work needs to be done to prove the generality of the solution. It is important to observe that the mechanisms employed are very general, and could potentially be applied in the context of very different object recognition schemes, including structural and view based, as long as they have a multi-layer hierarchy with pyramidal abstraction structure (e.g. [19, 20]).

A recent study regarding the time course and processing sequence of form-from-motion in humans using similar stimuli [21] concludes that dorsal activation (area V5/MT) precedes ventral activation (areas LO and IT) by 50-60ms. This time interval is consistent with the proposed mechanism [22]. The strongest indications about the nature of the internal representations used in object/shape perception comes from a series of imaging and neurophysiology experiments [23].

These results consistently show that for each in a very broad categories of stimuli, a small number of regions in IT become active, pointing to a sparse coarse coding. In order to solve the binding problem for object recognition, the various areas of activity corresponding to one stimulus must be selected together [13], but this is very difficult if not impossible at the level of IT due to the loss of spatial information. If, as indicated by recent studies (e.g. [24, 25]), categoric information is available early in visual processing, the holistic nature of this representation might be able to act as anchor for the mechanism proposed in this paper, and select all the individual activations, while at the same time inhibiting competing representations, thus increasing the signal-to-noise ratio of the selected stimulus. The idea that attention binds together distributed cortical activations at multiple levels of the visual hierarchy involved in processing attended stimuli has recently received significant experimental support [26], and reentrant connections between extrastriate areas and V1 are gaining support as the substrate for attention and conscious perception – see [27] for a review.

Object recognition is one of the most important problems in computer vision, and in this context probably the biggest challenge is the representational gap between low-level image features and high-level concepts such as generic models [28]. The proposal presented here allows a system to extract intermediate level representations based on available/extractable information and perceptual grouping, and use the feedback process to refine and bind together those intermediate level representations that belong to the same object. This would create a distributed sparse representation similar to that present in IT, representation know to be optimal for maximizing the capacity of associative memory networks [17]. Given the representational gap that computer vision systems must bridge, this method has the potential to make important contributions to the field.

References

1. Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., Lai, Y.H., Davis, N., Nuflo, F.: Modeling visual-attention via selective tuning. Artif. Intell. **78**(1-2) (1995) 507–545
2. Roskies, A.L.: The binding problem. Neuron **24**(1) (1999) 7–9
3. Rosenblatt, F.: Principles of Neurodynamics: Perceptions and the Theory of Brain Mechanisms. Spartan Books (1961)
4. Treisman, A., Schmidt, H.: Illusory conjunctions in the perception of objects. Cognit Psychol **14**(1) (1982) 107–41
5. Barlow, H.B.: Single units and sensation: A neuron doctrine for perceptual psychology? Perception **1**(4) (1972) 371–394
6. Ghose, G.M., Maunsell, J.: Specialized representations in visual cortex: a role for binding? Neuron **24**(1) (1999) 79–85
7. von der Malsburg, C.: The what and why of binding: the modeler's perspective. Neuron **24**(1) (1999) 95–104
8. Rothenstein, A.L., Tsotsos, J.K.: Attention links sensing to recognition. Image and Vision Computing (in press doi:10.1016/j.imavis.2005.08.011) (2006)
9. Milner, P.: A model for visual shape recognition. Psychol. Rev. **81** (1974) 521–535
10. von der Malsburg, C.: Nervous structures with dynamical links. Ber. Bunsenges. Phys. Chem. **89** (1985) 703–710

11. Singer, W.: Neuronal synchrony: a versatile code for the definition of relations? Neuron **24** (1999) 49–65
12. Treisman, A.M., Gelade, G.: Feature-integration theory of attention. Cognitive Psychology **12**(1) (1980) 97–136
13. Tsotsos, J.K., Liu, Y., Martinez-Trujillo, J.C., Pomplun, M., Simine, E., Zhou, K.: Attending to visual motion. Comput. Vis. Image Und. **100**(1-2) (2005) 3–40
14. Cutzu, F., Tsotsos, J.K.: The selective tuning model of attention: psychophysical evidence for a suppressive annulus around an attended item. Vision Research **43**(2) (2003) 205–219
15. Taraborelli, D.: Feature binding and object perception. Does object awareness require feature conjunction? In: 10th Annual Meeting of the European Society for Philosophy and Psychology - ESPP 2002, Lyon. (2002)
16. Robertson, L.: Space, Objects, Brains and Minds. Essays in Cognitive Psychology. Psychology Press (2004)
17. Foldiak, P., Young, M.: Sparse coding in the primate cortex. In Arbib, M.A., ed.: The Handbook of Brain Theory and Neural Networks. MIT Press (1995) 895–898
18. Fukushima, K., Imagawa, T., Ashida, E.: Character recognition with selective attention. In: International Joint Conference on Neural Networks. Volume 1., Seattle (1991) 593–598
19. Hummel, J.E., Stankiewicz, B.J.: An architecture for rapid, hierarchical structural description. In Inui, T., McClelland, J., eds.: Attention and Performance XVI: Information Integration in Perception and Communication. MIT Press. (1996) 93–121
20. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nature Neuroscience **2**(11) (1999) 1019–1025
21. Schoenfeld, M.A., Woldorff, M., Duzel, E., Scheich, H., Heinze, H.J., Mangun, G.R.: Form-from-motion: MEG evidence for time course and processing sequence meg evidence for time course and processing sequence. Journal of Cognitive Neuroscience **15**(2) (2003) 157–172
22. Bullier, J.: Integrated model of visual processing. Brain Research Reviews **36**(2-3) (2001) 96–107
23. Tsunoda, K., Yamane, Y., Nishizaki, M., Tanifuji, M.: Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. Nature Neuroscience **4** (2001) 832–838
24. Grill-Spector, K., Kanwisher, N.: Visual recognition: as soon as you see it, you know what it is. Psychological Science **16**(2) (2005) 152–160
25. Wolfe, J.M.: Moving towards solutions to some enduring controversies in visual search. Trends in Cognitive Sciences **7**(2) (2003) 70–76
26. Haynes, J.D., Tregellas, J., Rees, G.: Attentional integration between anatomically distinct stimulus representations in early visual cortex. Proc. Natl. Acad. Sci. USA **102**(41) (2005) 14925–30
27. Pollen, D.A.: Explicit neural representations, recursive neural networks and conscious visual perception. Cerebral Cortex **13**(8) (2003) 807–14
28. Keselman, Y., Dickinson, S.J.: Generic model abstraction from examples. IEEE T. Pattern Anal. **27**(7) (2005) 1141 – 1156