THE SELECTIVE TUNING MODEL FOR VISUAL ATTENTION

John K. Tsotsos

Department of Computer Science, and
Centre for Vision Research
York University
Toronto, Ontario, Canada

INTRODUCTION

Complexity analysis leads to the conclusion that if attention tunes the visual processing architecture task-directed processing is enabled and a solution to signal interference otherwise present in the converging feedforward pathways is provided. Selective tuning takes two forms: spatial selection is realized by inhibition of irrelevant connections; and feature selection is realized by inhibition of the units, which compute irrelevant features. Only a very brief summary is presented here (a more detailed account is in Tsotsos et al.[1]).

One role of attention in the image domain is to localize a stimulus in retinotopic as well as feature space in such a way so that any interfering or corrupting signals are minimized. In doing so, attention also seeks to increase the discriminability of a particular image subset. The search process that localizes the image subset is as follows. The visual processing architecture is assumed to be pyramidal in structure with units within this network receiving both feed-forward and feedback connections (the model has this in common with the architecture developed by Van Essen et al.[2]. When a stimulus is first applied to the input layer of the pyramid, it activates in a feed-forward manner all of the units within the pyramid to which it is connected; the result is that an inverted sub-pyramid of units and connections is activated. It is assumed that response strength of units in the network is a measure of goodness-of-match of stimulus to model and of relative importance of the contents of the corresponding receptive field in the scene.

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: it computes the global winner, i.e., the units with largest response. The WTA can accept guidance for areas or stimulus qualities to favour if that guidance was 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. This localizes the largest response units within the top-level winning receptive field. Next, all of the connections of the visual pyramid that do not contribute to the winner are pruned. This strategy of finding the winners within successively smaller receptive fields layer by layer in the pyramid and then pruning away irrelevant connections 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 while the pruned paths form the inhibitory zone of an attentional beam. The WTA does not violate biological connectivity or time constraints.

AN EXAMPLE

Figure 1 shows a hypothetical visual processing pyramid. There are 4 layers, each unit connected to 7 units in the layer above it and 7 units in the layer below it. The input layer (bottom layer) is numbered 1, while the output layer (top layer) is numbered 4. The two examples that follow are intended to illustrate the structure and time course of the application of attentional selection in the model. The first example shows the structure that results if a single stimulus is placed in the visual field. The second example shows the time course of attentional selection if two stimuli are placed in the visual field. In Figure 1, only the feed-forward connections are shown; the feed-back connections are analogous.

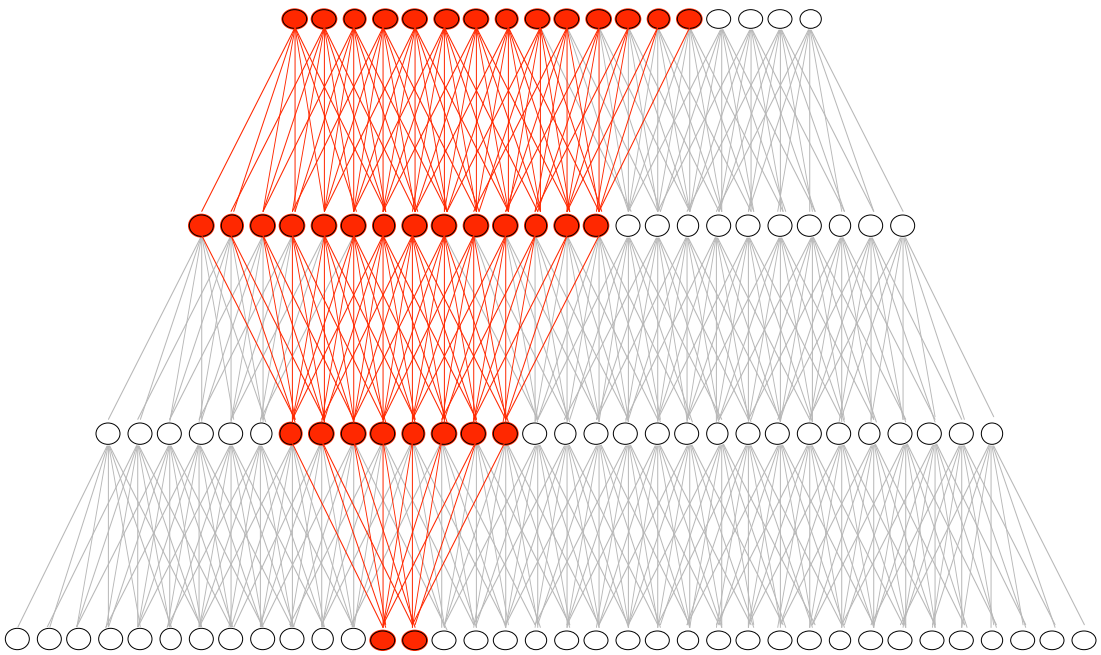


Figure 1. A hypothetical visual processing pyramid showing the portion of the pyramid activated due to the initial feedforward stimulation of a single input stimulus.

A stimulus that spans 2 units in the input layer is to be attended by the system; the resulting attentional beam is shown in Figure 2. The grey lines represent inactive connections, the black lines represent connections whose feedforward flow is inhibited by the attentional beam, and the red lines represent feedforward connections activated by the stimulus. Red units are activated solely by the red stimulus.

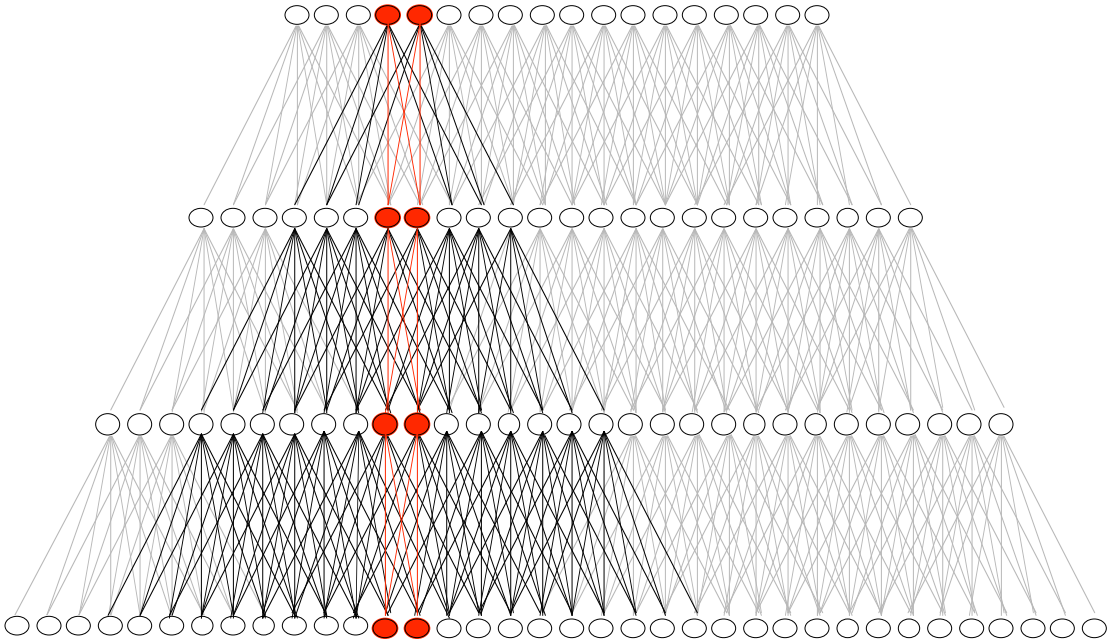


Figure 2. The final configuration of the attentional beam reacting to a single input stimulus.

The WTA mechanism locates the peaks in the response the output layer of the pyramid, the two remaining red units in Figure 2. The inhibitory beam then is extended from top to bottom, in a top-down coarse-to-fine manner, pruning away the connections that might interfere with the selected units. Eventually, the two stimulus units are located in the input layer and isolated within the beam.

How does the WTA process locate two winners in the output layer? On the assumption that each of the units in the pyramid computes some quantity using a Gaussian weighted function across its receptive field, then the maximum responses of these computations (whatever they may be) will be exactly the two units selected in the output layer (see Tsotsos et al. [1], for more detail on this). More importantly, with respect to attention, how does the mechanism function if there is more than one stimulus in the input, that is, with target as well as distractor elements in the visual field?

Figure 3 shows the first of a five-step sequence depicting the changes that the visual processing pyramid undergoes in such a situation. Using the same network configuration as in the previous figure and again showing only the feed-forward connections, two stimuli are

placed in the visual field (input layer). They are colour-coded red and blue as are the connections and units that are activated solely by them. The mauve coloured units and connections are those which are activated by both stimuli regardless of proportions. Note that much of the pyramid is affected by both stimuli and as a result, most of the output layer gives a confounded response.

The mauve units' response is a weak one due to the conflict that arises since each of those units 'sees' two different stimuli within its receptive field. Now the subject is directed to attend to the location of the red stimulus. Location is determined by a mechanism outside the network shown here; appropriate units corresponding to the location in the output layer are marked as shown in Figure 3.

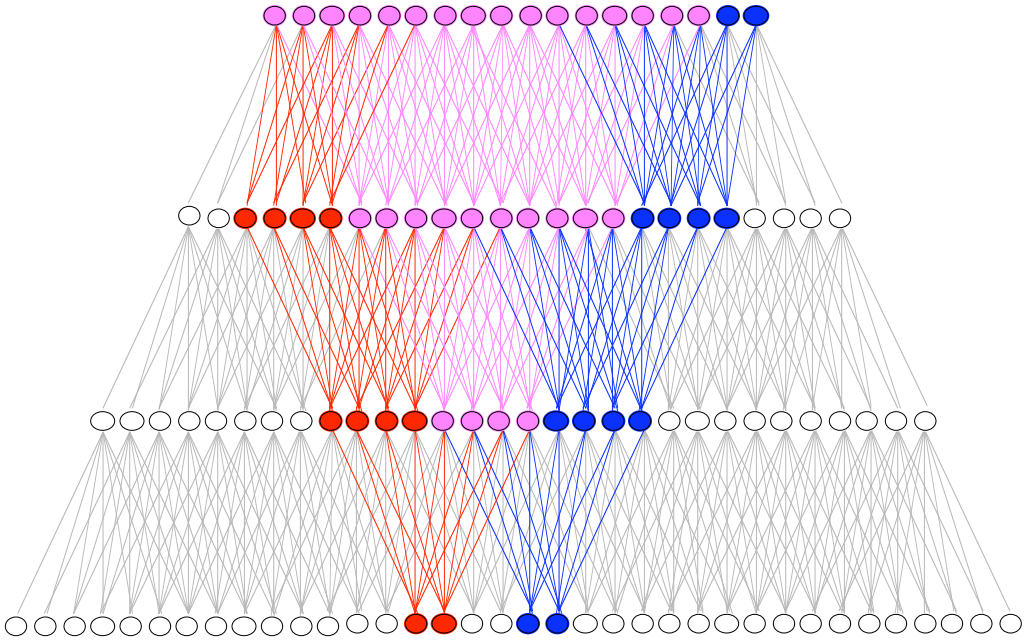
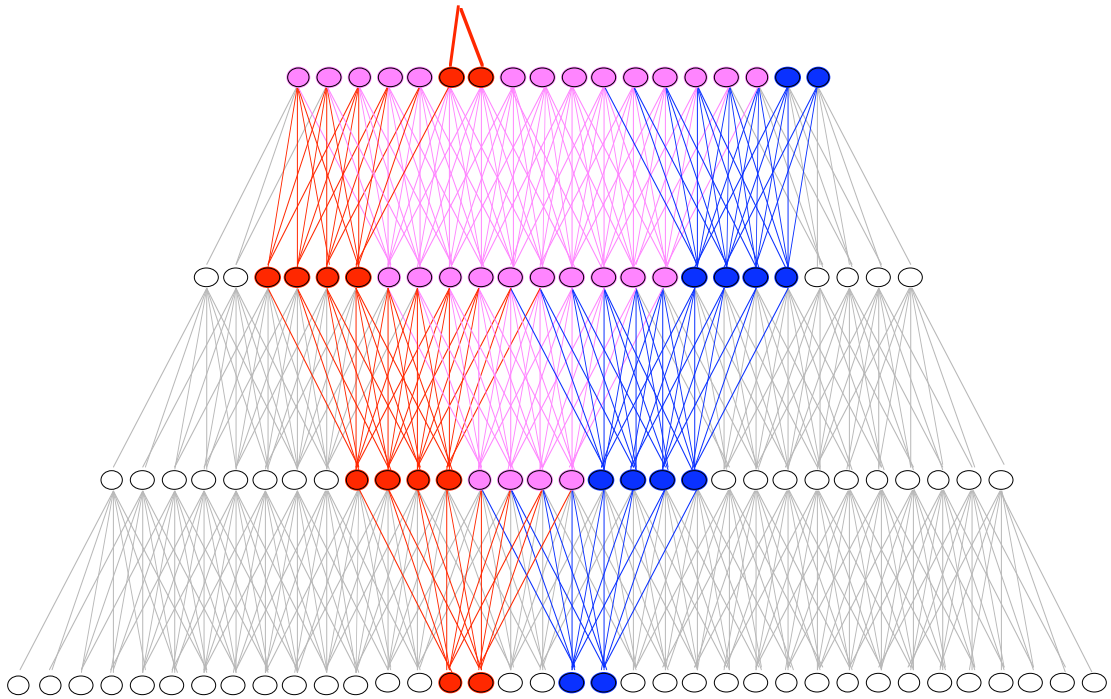


Figure 3. The visual processing pyramid at the point where the activation due to two separate stimuli in the input layer has just reached the output layer. No attentional effects



are yet in evidence.

Figure 4. Attention is focussed at the location of the output layer corresponding to the location of the selected input.

Attention is focused at the red units in the output layer. This is not to say that the output of those units reflects the desired input at this point in time. Rather, these units form the root of the attentional beam as it begins its downward traversal through the visual pyramid.

The next phase of the computation is to push the effect of attention down one level further, locating the units that will be the attended ones. Simultaneously, the feedforward connections from all units in the layer 3 which feed the attended units in the output layer are inhibited and are not part of the pass zone

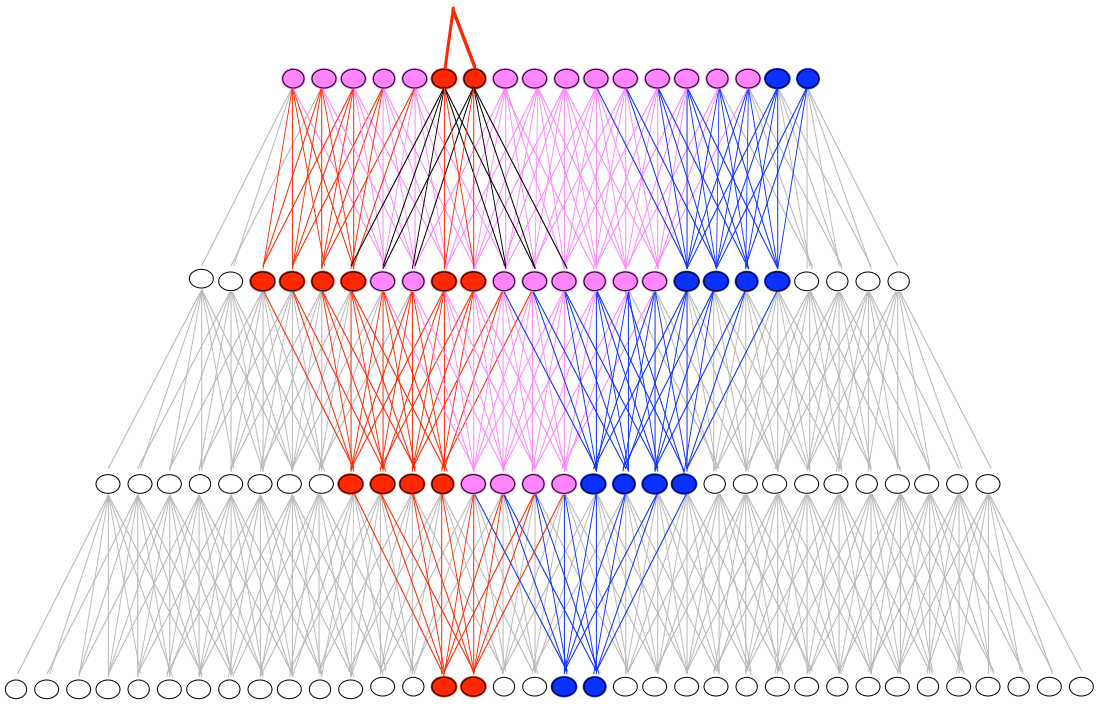


Figure 5. The first level of inhibition due to the attentional beam. The feedforward flow of the black connections is inhibited.

Figure 5 shows the connections whose feed-forward flow is explicitly inhibited by the attentional mechanism between layers 3 and 4. The interesting thing to note is that at this early stage of the application of the attentional beam very little seems to be changing. The large scale changes come later as more of the visual pyramid is affected by the flow of the attentional beam through it. The selection of units also moves down one level to layer 3.

The next major milestone in this process is shown in Figure 6. At this point of the beam traversal, major changes can be seen. The next set of connections, those between the middle layers are inhibited. In turn, those inhibitions cause several units in 3 layer to have no active input and thus they provide no signal to the output layer. Those connections are coloured grey, the same as the other inactive connections in the pyramid.

This change in turn causes several units in the output layer previously coded mauve to be coded red, that is, they receive signals originating only from the red input stimulus.

The final stage of the process leads to the network shown in Figure 7. After this point, the selected units in the output layer receive input only from the selected stimulus in the input layer. Note that several units in the output layer are coded in blue, showing that the effect of the blue stimulus still gets through the beam structure, in fact stronger than in the unattended case of Figure 3.

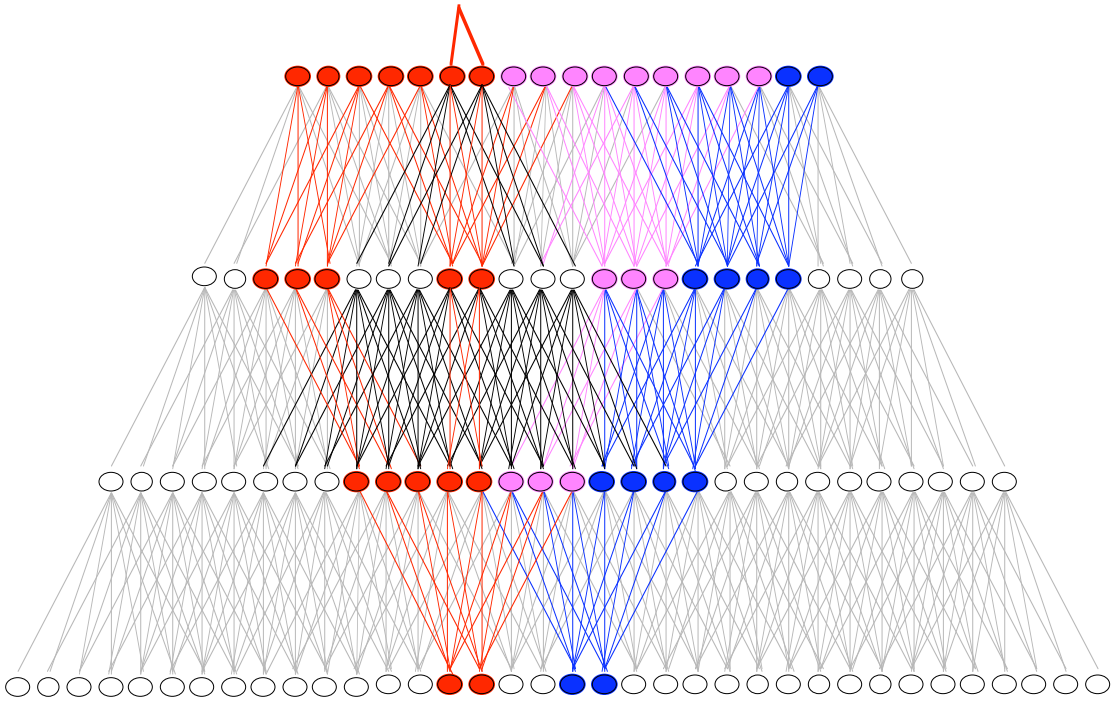


Figure 6. The second level of attentional inhibition. Several units in layer 3 now receive no input and thus do not provide signals to the output layer.

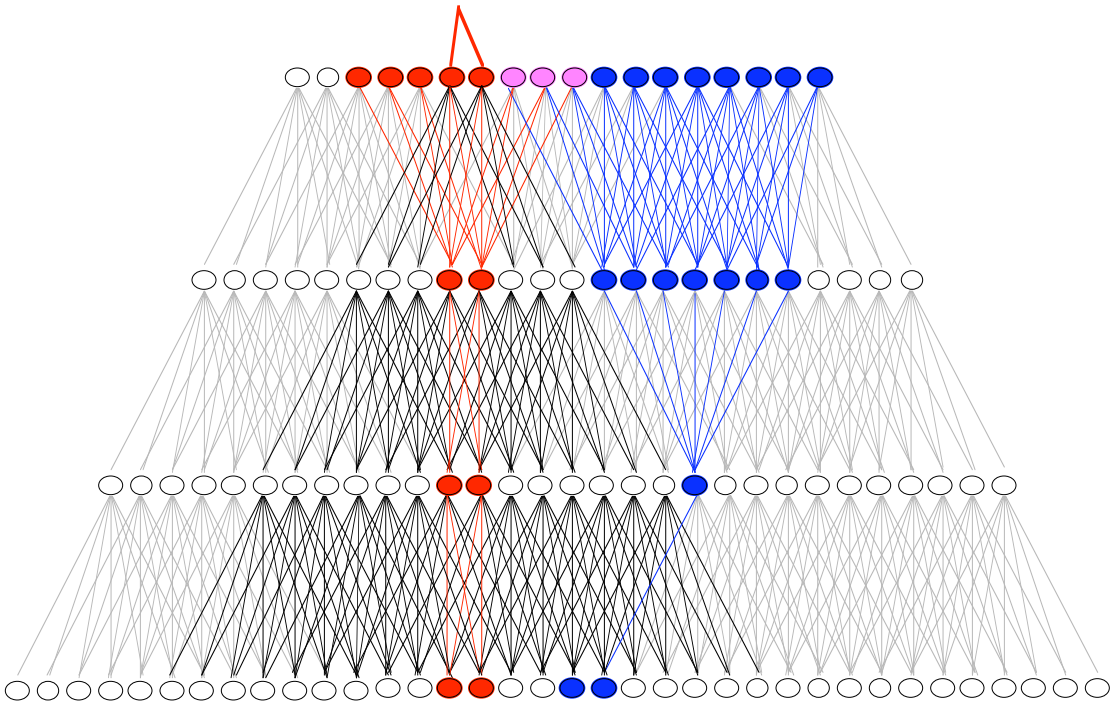


Figure 7. The third and final level of inhibition due to the attentional beam.

As well, a few of the weakly responding mauve units still remain. The interference between stimuli evident in Figure 3 is eliminated completely with respect to the item attended and much reduced for the unattended item. The events depicted with this set of figures would occur in the 100 to 200 ms after stimulus onset (for example, as shown by Chelazzi et al. [3]). Note the difference between the pattern of activations in Figures 2 and 7. In the former case, no location cue is given; the winner-take-all mechanism chooses the strongest responses in the output layer and inhibits the rest. Thus the set of connections and units attended forms the structure shown with the active connections being strictly those permitted by the selection mechanism. In the latter case, a location cue is given; thus, there is no inhibition within the output layer (if there were, none of the blue or mauve units would survive).

Suppose one records from one of the mauve 'neurons' in layer 3 of Figure 6 during the entire process. What will the response over time look like? On stimulus onset, the response will rise from zero to some level; then, as the beam is applied, will remain steady until the time corresponding to Figure 6 when the response will change. The fact that the WTA process is not a binary one and that changes occur gradually in an iterative fashion for each level also has impact on the time course of the response. The changes in the unit's response will begin when the WTA is first applied to the configuration of Figure 5 and not only once it completes. Suppose one recorded from one of the units coloured red in layer 3 of Figure 7 throughout (this would correspond to a receptive field that was being attended). One would observe an increment in response late in the response time course here. This kind of increment in response well into the attentive process is observed experimentally (for example, Motter[4]; also inhibitions were reported); this is the first explanation proposed for Motter's observations. The inhibitory effects are most clearly seen by observing the time course of units coloured white in layers 2 and 3 within the inhibitory beam of Figure 7. The changes due to attentional modulation during the top-down traversal of the attention process are summarized in Figure 8. It is apparent that both enhancement and suppression can be observed but in specific portions of each visual processing layer. This spatio-temporal structure of attentional modulation represents a strong prediction of the model.

From the example above, it should be clear that the retinotopic distance between the two stimuli in the input layer is important. If the blue stimulus were one unit closer to the red, none of its signal would reach the output layer after the application of the attentional beam. On the other hand if it were one unit farther away, the conflict region is smaller.

The above does not by itself account for the serial search observed in experiments of visual search. It does form, however, a part of the explanation. The rest of the explanation includes a method for inhibition of return, and for the re-deployment of the beam to the next most salient target. The model does in fact accomplish serial search well and these examples can be found elsewhere[1].

However, there is a body of single-cell recording work that this example does explain. One of those key experiments was that described by Moran and Desimone[5]. What they showed, which was surprising at the time, was that even though an effective stimulus was within a neuron's receptive field, it did not cause the neuron to fire well if the monkey was cued to attend a non-effective stimulus within the same receptive field. A more specific demonstration using the selective tuning model of that exact experimental set-up appears in

Tsotsos et al.[1].

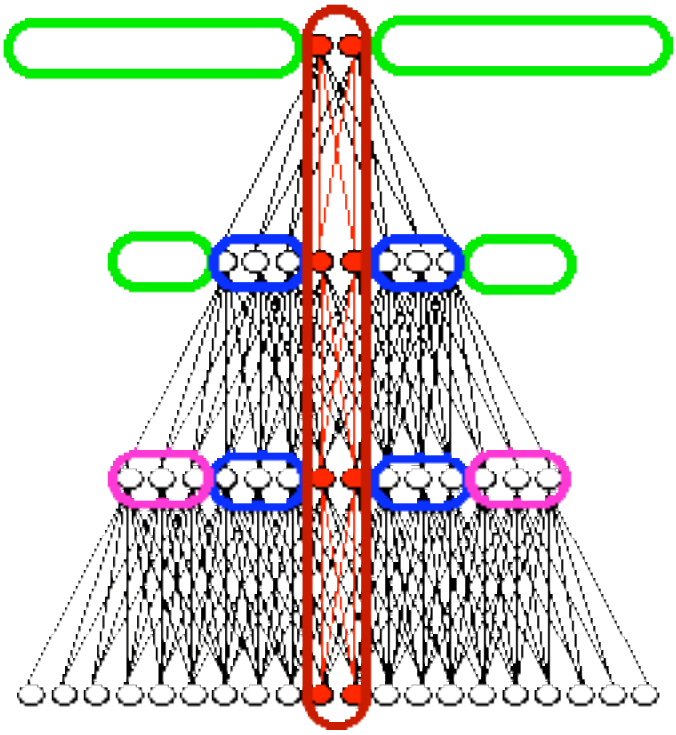


Figure 8. Modulation predictions. Following the changes of a particular unit of the pyramid through the sequence of Figures 3 through 7 leads to the overall changes depicted in this diagram. Specific portions of each layer undergo systematic changes as indicated; the changes depend on whether distractors are present or not and their strengths may differ depending on the distance separating the attended stimulus and the distractors. The best way to relate this figure to those of Figures 3-7 is to select a specific unit in the pyramid in Figure 3 and track its changes over time as depicted in the sequence. The colour-coded ovals represent the following: green ovals depict the portion of a layer where increases due to attention can be observed but only if distractors are present; the blue ovals represent the portion of the layer where decreases due to attention can be observed with or without distractors; purple ovals represent portions of the layer where decreases can be observed only if distractors are present; and the red oval represents the pass zone of the attentional structure and no change is observed if there are no distractors and may increase if distractors are present.

Since distance between stimuli is important, if the attended stimulus is near but not in the receptive field studied, the inhibitory effect of attention on the recorded neuron should be large. If it is far, the effect should disappear, and in between the inhibitory effect of attention will gradually decrease with increasing distance. This should be clear from the figures above. This corresponds very closely to the kind of activity observed in visual-movement neurons in the frontal eye fields when tested with visual search tasks that include distractors [6]. They found that neural activity peaked when the target was in the response field and was suppressed when the target was beside but not distant from this field. The magnitude of these effects are also affected by the position of the neuron within the hierarchy. Schall and Hanes hypothesized that this might be due to a lateral inhibition

mechanism; the selective tuning model is an alternative explanation. There is insufficient information to accomplish such a task-specific inhibition if only lateral connections are considered.

Motter[8] concluded that the topographic representation of the neural activity in area V4 highlights potential candidates for matching to targets while minimizing the impact of any background items. In other words, the computations that create this representation seem to maximize signal-to-noise ratios for the features that are relevant to the task. Neural activity was attenuated when the stimulus did not match a cue, independent of spatial location, but was about twice as large as the attenuated value if the stimulus and cue did match. He used colour and luminance as features. Interestingly, he found that neural activity was not affected due to the cueing conditions prior to presentation of stimulus arrays. This is consistent with a model which de-emphasizes connections which are not of interest. Motter goes one step further[9] and concludes that the attentional control system seems to be able to "shut down" the synaptic impact of all but one of many colour inputs. This too is consistent with the selective tuning model and was suggested by Tsotsos[7] as an important search optimization. Finally, Motter suggests that a sequential combination of a full field pre-attentive selection based on features which identify candidate targets, followed by a spatially restrictive focal attentive process which localizes targets, would be an interesting explanation of both his and Moran and Desimone's results; this is exactly the concept initially sketched out and embodied in the selective tuning model presented here[9].

MODEL PREDICTIONS

The model described displays performance compatible with experimental observations. The predictive power of the model seems broad:

• An early prediction[7] was that attention seems necessary at any level of processing where a many-to-one mapping of neurons is found and the potential exists for stimulus interference[7]. This was disputed at first[20]; however, more recent experimental work would appear to be supportive [10, 11]. Further, attention occurs in all the areas in concert. The prediction was made at a time when good evidence for attentional modulation was known for area V4 only[5]. Since then, attentional modulation has been found in many other areas both earlier and later in the visual processing stream, and that it occurs in these areas simultaneously[10]. Vanduffel and colleagues[11] have shown that attentional modulation appears at early as the LGN. The prediction that attention modulates all cortical and perhaps even subcortical levels of processing has been borne out by recent work from several groups[12, 13].

• The notions of competition between stimuli and of attentional modulation of this competition were also early components of the model[7] and these too have gained substantial support over the years [14, 10, 15].

• The model predicts an inhibitory surround that impairs perception around the focus of attention[7]. This too has recently gained support [11, 16, 17].

• The model further implies that pre-attentive and attentive visual processing occur in the same neural substrate, which contrasts with the traditional view that these are wholly independent mechanisms. This point of view has been gaining ground recently [18, 19].

• A final prediction is that attentional guidance and control are integrated into the visual processing hierarchy, rather than being centralized in some external brain structure. This implies that the latency of attentional modulations *decreases* from lower to higher visual areas.

CONCLUSIONS

The selective tuning model was derived in a first principles fashion. The major contributor to those principles derives from a series of formal analyses performed within the theory of computational complexity, the most appropriate theoretical foundation to address the question "why is attention necessary for perception?" The model not only displays performance compatible with experimental observations but also does so in a self-contained manner. That is, input to the model is a set of real, digitized images and not pre-processed data. Several examples using this paradigm using the implemented model with real images obtained from a robot head have appeared[1]. The predictive power of the model seems broad.

A new example of the performance of that model is given which shows how attention might function in the face of conflicting stimuli from a simulated single-cell recordingperspective. What remains? A great deal! Among other problems, complexity level analysis is expected to provide insights into:

- how much information can be extracted from a given attentional fixation?
- how are successive fixations integrated?
- how many successive fixations may be processed simultaneously?
- how does the eye movement system interact with covert attentional system?
- how is task information represented and how much of it can be used to tune the visual processing pyramid?
- how large is the visual processing pyramid in terms of numbers of layers, sizes of layers and number of units computing different visual qualities at each position?

Note that all of these open problems depend strongly on the amount of computation that can be performed in a given amount of time, or how much memory is required to store information; thus, complexity is an appropriate tool for their analysis.

Complexity analysis is by no means the most useful tool in the repertoire of the visual scientist. It is however a long neglected one, and a critical tool that can predict the realizability and performance of a given perceptual theory with respect to its neural or silicon implementation.

REFERENCES

1. J.K. Tsotsos, S. Culhane, W. Wai, Y. Lai, N. Davis, F. Nuflo. Modeling visual attention via selective tuning, *Artificial Intelligence,* 8(1-2),p 507 - 547. (1995).
2. D. Van Essen, C. Anderson, D. Fellowman. Information processing in the primate visual system: An integrated systems perspective, *Science 255(5043),* p 419 - 422. (1992).
3. L. Chelazzi, E. Miller, J. Duncan, R. Desimone. A neural basis for visual search in inferior temporal cortex, *Nature* , Vol. 363, p 345 - 347. (1993).
4. B. Motter. Focal attention produces spatially selective processing in visual cortical areas V1, V2 and V4 in the presence of competing stimuli, *J. Neurophysiology 70(3)*, p 909 - 919. (1993).
5. J. Moran, R. Desimone. Selective attention gates visual processing in the extrastriate cortex, *Science 229,* p 782 - 784. (1985).
6. J. Schall, H. Hanes. Neural basis of saccade target selection in frontal eye field during visual search, *Nature 366*, p 467 - 469. (1993).
7. J.K. Tsotsos. A Complexity Level Analysis of Vision. *Behavioral and Brain Sciences*, 13, 423-455. (1990).
8. B. Motter. Neural correlates of attentive selection of color or luminance in extrastriate area V4, *J. Neuroscience.14(4)*, p 2178 - 2189. (1994).
9. B. Motter. Neural correlates of feature selective memory and pop-out in extrastriate area V4, *J. Neuroscience.14(4),* p 2190 - 2199. (1994).
10. S. Kastner, P. De Weerd, R. Desimone, L. Ungerleider. Mechanisms of Directed Attention in the Human Extrastriate Cortex as Revealed by Functional MRI, *Science 282*, p108 - 111. (1998).
11. W. Vanduffel, R. Tootell, G. Orban. "Attention-dependent suppression of metabolic activity in the early stages of the macaque visual system, *Cerebral Cortex*. (2000).
12. J. Brefczynski, E. DeYoe. A physiological correlate of the 'spotlight' of visual attention. *Nat Neurosci*. Apr;2(4):370-4 . (1999).
13. S. Gandhi, D. Heeger, G. Boynton. Spatial attention affects brain activity in human primary visual cortex, Proc. Natl Acad Sci U S A 1999 Mar 16;96(6):3314-9. (1999).
14. R. Desimone, J. Duncan. Neural Mechanisms of Selective Attention, *Annual Review of Neuroscience 18*, p193 - 222. (1995).
15. J. Reynolds, L. Chelazzi, R. Desimone. Competitive Mechanisms Subserve Attention in Macaque Areas V2 and V4, *The Journal of Neuroscience, 19(5)*, p1736–1753. (1999).
16. G. Caputo, S. Guerra. Attentional Selection by Distractor Suppression, *Vision Research 38(5)*, p. 669 - 689. (1998).
17. D. Bahcall, E. Kowler. Attentional Interference at Small Spatial Separations, *Vision Research 39(1)*, p 71 - 86. (1999).
18. J. Joseph, M. Chun, K. Nakayama. Attentional Requirements in a 'Preattentive' Feature Search Task, *Nature 387*, p. 805 - 807. (1997).
19. Y. Yeshurun, M. Carrasco. Spatial attention improves performance in spatial resolution tasks, *Vision Res*. Jan;39(2):293-306. (1999).

20. R. Desimone. Complexity at the Neuronal Level, *Behavioral and Brain Sciences 13(3)*, p 446. (1990).