Appearing in **Visual Attention and Cortical Circuits**,
ed. by J. Braun, C. Koch, J. Davis, MIT Press, 2001 p 285 – 306.

From Foundational Principles to a Hierarchical Selection Circuit for Attention

John K. Tsotsos, Sean M. Culhane, Florin Cutzu

Dept. of Computer Science
University of Toronto

1. Introduction

The modeling efforts described in this paper span theoretical considerations, computer simulations applied to real-world scenes and human psychophysics. The theoretical work initially addressed the question "Is there a computational justification for attentive selection?". The obvious answer that has been described many times since at least Broadbent, namely that the brain is not large enough to process all the incoming stimuli, is not satisfactory since it is not quantitative and provides no constraints on what processing system might be sufficient. Methods from computational complexity were employed to formally prove for the first time that purely data-directed visual search in its most general form is an intractable problem in any realization (Tsotsos, 1989). 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. The constraints arose because vision was cast as a search problem; the combinatorics of search are too large at each stage of analysis and attentive selection is a powerful heuristic to limit search and make the overall problem tractable.

Attention is an important mechanism at any level of processing where one finds a many-to-one convergence of neural inputs and thus potential stimulus interference, a conclusion reached in (Tsotsos, 1990). This was disputed by (Desimone, 1990); recently, however, experimental work seems supportive (for example, Kastner et al., 1998, Vanduffel et al. in press).

The basic component of the attentional mechanism is a hierarchical neural network that implements a task-dependent, top-down, directed competition among conflicting neural elements, a circuit first described by Tsotsos (1993). As such, the mechanism implements a selective tuning of the visual processing hierarchy (the model is thus named the 'selective tuning model'). In contrast to theories that claim similar conceptual strategies for attention (such as Desimone and Duncan, 1995), our model has been fully detailed and simulated on a computer. It provides attentive control to a robotic camera system and attends both overtly and covertly to task-directed features and objects using real-world image sequences acquired from video cameras. As such, it is an existence proof that the key elements of the model are realizable and perform as expected.

The exposition will proceed in two parts. The first part will overview the selective tuning model. In particular, the issue of attentional control will be highlighted. The second section will detail an experimental investigation that tests a basic prediction of the model, namely, that with attention, perception is impaired near the attended stimulus.

2. The Selective Tuning Model

Complexity analysis leads to the conclusion that attention must tune the visual processing architecture to permit task-directed processing (details in Tsotsos 1990). Selective tuning takes two forms: spatial selection is realized by inhibition of irrelevant connections in the neural network, and feature selection is realized by inhibition of the neurons in that network which compute non-selected features. Only a brief summary is presented here since the model is detailed elsewhere (Tsotsos et al. 1995).

The role of attention in the image domain is to localize a subset of the input image and its path through the processing hierarchy in such a way so that any interfering or corrupting signals are minimized. The visual processing architecture is a pyramidal network with units within this network receiving both feedforward and feedback connections (the model has this in common with the architecture developed in Van Essen et al. 1992). When a stimulus is first applied to the input layer of the pyramid, it activates in a feedforward 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 as shown in Figure 1. It is assumed that the activation value of units in the network is a measure of goodness-of-match of stimulus to the model that the unit represents. Figure 1 shows such a visual processing pyramid. There are four layers, each unit connected to seven units in the layer above it and seven units in the layer below it. The input layer (bottom layer) is numbered 1, while the output layer (top layer) is numbered 4.

**********Figure 1 ********

In the figure, only the feedforward connections are shown; the feedback connections are identical but conduct information in a feedback direction.

2.1 Hierarchical Winner-Take-All Processes

Selection relies on a hierarchy of winner-take-all (WTA) processes. WTA is a parallel algorithm for finding the maximum value in a set of variables (first proposed in this context by Koch and Ullman 1985). The WTA can accept guidance for areas or stimulus qualities to favor if that guidance were available but operates independently otherwise. Processing an input using this model proceeds in three main stages. The first stage of processing beings when stimuli are applied to the input layer of the network. They propagate upwards to the output layer in a feed-forward manner and are processed by the neurons in the network according to their particular selectivities (perhaps biased for task specific stimulus qualities). The second stage applies a hierarchy of WTA processes in a top-down, coarse-to-fine manner. The first WTA process operates across the entire visual field at the top layer: it computes the global winner, i.e., the units in the output layer with largest response. The global winner activates a WTA that operates only over its input units from the layer below. This localizes the largest response units within

the top-level winning receptive field. All of the connections of the visual pyramid that do not contribute to the winner are 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. Thus, the perceptual origin of that largest response in the output layer is eventually localized in the input layer. The remaining paths may be considered the pass zone while the pruned paths form the inhibitory zone of an attentional beam. While we are not claiming biological accuracy for the WTA process, we are claiming plausibility, as it does not violate biological connectivity or time constraints. The final stage allows the selected stimuli in the input layer to re-propagate through the network, being processed again by the neurons of the network but this time as if they are found on a 'blank' background (since any distractors within the receptive fields have been inhibited). Note that there is no change in identity of the winning neurons in the output layer; the winner initially selected remains the winner but its value is refined by this process.

2.2 Examples

The process described above is shown in the examples below. The first example shows the start and end stages for the processing of a single stimulus item and is shown in Figure 1. If the system attends the stimulus in the input layer of Figure 1a, the configuration of Figure 1b results. The gray lines represent inactive connections, the black lines represent connections whose feedforward flow is inhibited (pruned) by the attentional beam, and the red lines represent feedforward connections activated by the stimulus. Red units are activated solely by the red stimulus.

The WTA mechanism locates the peaks in the response of the output layer of the pyramid, the two remaining red units in Figure 1b. The attentional beam is then extended from top to bottom, 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.

The important missing link is the mechanism for localizing the two winners in the output layer. On the assumption that each of the units in the pyramid computes some quantity using a Gaussian-shaped weighting function across its receptive field, the maximum responses of these computations (whatever they may be) will correspond exactly to the two units selected in the output layer (see Tsotsos et al. 1995, for more detail on this including the mathematical formulation and proofs of its properties). The general question is thus 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 2 shows the four-step sequence of the changes that the visual processing pyramid undergoes in such a situation.

******** Figure 2 ******

Using the same network configuration as in the previous figure and again showing only the feedforward connections, two stimuli are placed in the visual field (input layer). They are color-coded red and blue as are the connections and units which are activated solely

by them[1]. The mauve colored units and connections are those which are activated by both stimuli (the relative proportion does not matter here). 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 weak due to the conflict that arises when each of those units 'sees' two different stimuli within its receptive field. Whatever the optimal tuning properties of a unit may be, non-optimal input will lead to a reduced response. This manner of response is a general characteristic of tuned filters commonly used in computer vision for decades; it now appears as if there is experimental evidence supporting this characteristic (Reynolds et al. 1999).

How is the location cue provided to the model (units corresponding to the location in the output layer are marked as shown in Figure 2b)? When a human subject in a psychophysical experiment is given a location cue, this is commonly done by displaying a spot/cross on the image and then the remaining test images are displayed in spatial register with the cue image. This is used in the model as well. The system attends to the cue when it is presented, and then remembers the locations of that cue stimulus at the output layer of the network for use when the test display is presented. Figure 2b shows the changes in connection strength and unit strength after the first WTA stage is completed, the largest responses within the next layer receptive fields of the selected units are found. The connections not corresponding to those largest response units are inhibited. Figures 2c and 2d show the final two stages of the process.

2.3 Spatial Structure of Attentional Modulation through the Network

The example leads to a particular question: if there are is than one stimulus in the input, what is the spatial and temporal structure of the attentional changes in the network as the hierarchical WTA processes are applied? Figure 3 summarizes the changes that specific portions of each layer undergo during the application of attentive selection; the changes may differ depending on whether distractors are present or not. Further, the changes and their intensities may depend on the distance separating the attended stimulus and the distractors. Motter (1993) observed both increases and decreases in response when attention was required for images containing more than one stimulus. Whether one observes an increase or decrease depends on exactly where the neuron one records is found in relationship to the attended stimulus path. Due to the top down application of the WTA processes, the temporal order of response change is from top to bottom.

******* Figure 3 ********

The proposed circuit that may accomplish selective tuning described is now presented. Figure 4 shows a small portion of the 3 adjacent layers of the pyramid.

********Figure 4 ********

1. This network is a simplified representation. All the units are assumed to have the same tuning selectivity, specific to a unitary stimulus (whose type is irrelevant). There is no loss of generality as a result of these simplifications; the implementation of the model does not incorporate them.

2.4 Network Structure and Function

The model requires several different types of computing units arranged in a pyramid. *Interpretive units* compute the visual features. *Gating units* compute the WTA result across the inputs of a particular interpretive unit and then gate the winning input through to the interpretive units in the next feedforward layer of the pyramid. *Gating control units* control the downward flow of selection through the pyramid and are responsible for the signals which either activate or shut down the WTA processes. *Bias units* provide top-down, task-related selection via multiplicative inhibition. A group consisting of one interpretive unit, its associated gating control and bias unit, the set of WTA gating units on the inputs of the interpretive unit and associated connections is termed an *assembly*.

The control signal which turns the selection process on and off provides top-down control of the WTA processes. It selects the path of the beam's pass zone depending on the winning WTA units in the next higher layer. If the gating control unit has value 1.0, then the WTA process is turned on, otherwise it is turned off. This is implemented by multiplicatively modifying the iterative rule so that if the WTA is off, all updated values are zero. In this way, the gating units are affected but the interpretive units are not; only the relevant connections are closed down thus allowing the interpretive unit to participate in other computations as needed. The value of the control signal is zero for all units during the first phase of the process. During this first phase, the gating units are open and the WTAs are all disabled so that the responses computed by the interpretive units based on the stimulus in a bottom-up fashion can pass through the pyramid. Then, the value of the control signal becomes 1.0 for all the units at the top layer, turning on the top WTA process. The results of this WTA process then determine the values of the control signal for the successively lower layers; winning units propagate a value of 1.0 downwards while losing units propagate 0.0. As the pruning of connections proceeds downwards, new results of interpretive unit computations become available as their inputs are restricted. Time is allowed for the complete upward propagation of the new results. After this upward propagation, a period of time is provided where the same path through the pyramid cannot be active. This inhibition of the selected region and pathway is a concept extended from Koch & Ullman (1985) where it was referred to as inhibition of the input.

The winner-take-all algorithm and its properties are described in (Tsotsos et al. 1995). The algorithm is provably convergent, permits multiple winners, and appears optimal when compared to the provably optimal parallel maximum-finding algorithm of (Valiant 1975).

The bias network has two functions. First, bias units provide top-down, task-related selection via multiplicative inhibition if prior information about the scene is present. The bias is communicated in a top-down fashion through the bias network. Second, the network has feedforward pathways that are used to communicate positional information regarding the winners of the WTA competition. Therefore, the bias network can play a large role in object recognition by computing and conveying the position of the winning elements. Each WTA determines its local winner and codes its position as a particular element within that receptive field. The next level WTA then finds its winner and can do the same. The position can be passed upward one level and thus provides an index into the space coded by the top-level winner. It is directly analogous to each layer

of WTA computing a higher order bit of a binary address before activating the WTA process directly below. A full address is computed when the top-down traversal is complete. As a result it is conceivable that a code can be constructed that represents position to a relatively high degree of accuracy. This may be conveyed upwards into the object recognition portion of the network using the bias network connections. This information allows the object recognition process to verify whether the configuration of the strongest detected features corresponds to the object to be recognized. The precision of location information associated with an attended stimulus is directly dependent on the amount of time provided for the attentional process to complete its feedback traversal of the hierarchy.

2.5 A Neural Correlate to the WTA Circuit?

Is there a neural correlate to such a localized WTA network? The proposal of such a network, first made in (Tsotsos 1993), makes a strong prediction. There must be single neurons (or collections of neurons) in each visual area that make contact with neurons of similar, competitive, selectivity (so they can compete with one another in order to represent the same area of visual space; there would be such a competition for each feature type in each part of visual space[2]). The spatial extent of the connections of such a neuron must correspond to the extent of the receptive field of a neuron in the next higher visual area (the extent of the gray units in area B of Figure 5a as they relate to their target in area A). The nature of contact is inhibitory as required by local WTA competition. The appropriate kind of feedforward/feedback divergence of connections is found between areas 17 and 18 in the cat (Salin & Bullier 1995; shown in Figure 5a) and our proposal for their function is not inconsistent with findings described in that paper. Area V1 in macaque monkey contains laterally spreading axons that span several millimeters. The connections from pyramidal and spiny stellate neurons form periodic clusters, preferentially linking columns of neurons with similar response properties. In cats, ferrets and monkeys, columns with similar orientation preference are linked (for review see Callaway 1998; it should be pointed out that the presentation here is simplified). The neural architecture thus seems to have suitable components for the WTA gating networks proposed in the selective tuning model. It remains, of course, to determine exactly what information processing is occurring within this cluster architecture and to see if it matches the model's time course of selection of strongest inputs while inhibiting the transmission of the remaining inputs to the next higher processing area.

In Figure 4, the WTA network is shown as a collection of units each connected to one another. In fact, in order to have a fully parallel, distributed processing implementation of such a network, this connectivity is necessary. It seems though that this connectivity may be difficult to implement in neural hardware. Although it is a useful principle to assume that the brain uses parallel processing whenever possible, this might be one instance where a central processing mechanism is preferred. Figure 5b and c show the two possibilities and it is important to note they are functionally equivalent. The central processing strategy seems the easier one to map onto the neurons described by Callaway. The central node of the figure may correspond to the neuron's soma while the

2. For example, at any point in the visual field the set of oriented line detectors would yield the best orientation, the set of color-opponent units would yield the best color contrast, the set of directionally selective motion cells would yield the best direction, and so on.

neuron's dendrites connect with the feedforward connections of the neurons that provide input to neurons that perform feature analysis (the interpretive units of the model). The neuron would determine the winners in the WTA in a central fashion[3], and provide inhibitory signals to the connections of the losers.

********Figure 5 ******

This algorithm has been applied to digitized real scene images and an example in shown in Figure 6. Several other examples may be found in (Tsotsos et al. 1995).

*******Figure 6 ********

2.6 Is Attentional Control Centralized or Distributed?

The issue of attentional control is critical. How might it occur? Where is its source? How does the source make decisions and where in the visual processing hierarchy are those decisions applied? The selective tuning model makes very strong statements on these issues. The original arguments for distributed, local control and decision-making are found in (Tsotsos 1990) and are rooted in the analysis of the space complexity of the task (number of units, number of connections, number of inputs and outputs for each unit, lengths of connections). It is however important to consider the various alternatives to this strategy. In particular, would a centralized, external attentional decisional process[4] not also be feasible? The following paragraphs address this potential alternate explanation.

Consider the set of visual areas where attentive modulation has been observed, and their connections as shown in Figure 7[5]. It is important to mention that one of the original predictions of the selective tuning model was that attention must be applied throughout the visual processing hierarchy[6]. Assume that the signals that define the attentional modulation are determined in a central structure (an attentional control center, AC) and they are communicated to the areas of the processing hierarchy where they then affect individual neurons.

*******Figure 7 ******

In order to properly address this issue, several key questions must be answered and these

---

3. Little is to be gained by speculation at this point on the exact mechanism. However, it is easy to think of a simple circuit that can compute differences between pairs of inputs as the WTA algorithm requires (Tsotsos et al. 1995) and with this information determine the largest of its inputs within a few iterations.

4. Personal communication, David Van Essen, John Maunsell 1999. Crick (1984) proposed that attention is controlled by the thalamic reticular complex; Koch & Ullman (1985) propose that the LGN as the locus of the central saliency map and control; Olshausen et al. (1993) propose that the pulvinar as the controlling neural structure; He et al. (1996) suggest that the dorsal parietal areas play this role.

5. I thank Dan Felleman for sharing the software used to create the original figures for Felleman and Van Essen (1991).

6. This appeared in (Tsotsos 1990). Desimone's commentary on that article included some disbelief that this would be the case (Desimone 1990).

will be addressed in turn.

What are the goals that AC must satisfy ? At least three tasks seem important: selection of stimuli to attend, routing of stimuli through the processing hierarchy in order to remove interfering contextual stimuli, and coordination in space and time so neurons throughout the hierarchy attend to the same items. There is no claim that these are the only functions of AC, however, for this argument these are sufficient.

What information does the AC require in order to accomplish these tasks? In order to select the items to attend, it seems necessary to have a global view of the visual field; no algorithm could determine the most salient item otherwise (a view first appearing in Milner, 1974). Task instructions modify the selection of most salient item so that the selection reflects current goals and therefore AC must have access to those instructions. In order to localize the attended stimulus in retinotopic space as well as in feature space the abstraction accomplished by the feedforward paths must be reversed. In other words, in feedforward processing, precise location information associated with features is discarded (many-to-one neural mapping). The reversal may occur by using a local search process in the feedback direction to determine what is the most important or salient item in a neuron's input. Such a search process must occur for every neuron that exhibits a feedforward many-to-one mapping and must be done in such a manner so that all relevant neurons attend to the same stimulus. The information required for this is local to each neuron.

Where is the source of the required information? The global view of the visual field can be satisfied in one of two ways. Either neural representations where the largest receptive fields are present are examined because each includes most of the visual field (and it seems that for the visual hierarchy the neurons in the anterior layers of inferotemporal cortex do have this characteristic), or all of the neurons in any other representation are examined because together they cover the entire visual field. On the other hand, localization of stimuli requires the fine scale representations available in the earliest layers of the visual hierarchy. It can be concluded that AC may require access to both the highest levels and lowest levels of representation. Task instructions seem to have a separate representation (see Corbetta in this volume) and AC would require links to this area.

What is the connectivity required for AC to receive this information? Since it cannot be determined in advance where salient stimuli appear in the visual field, full connectivity from each neuron in each layer of the processing hierarchy seems necessary in order to provide a path for the information to the AC.

What processing must AC perform in order to satisfy its goals? In order to detect the most salient item in the visual field, it has been shown that a winner-take-all model operating on a representation of saliency will suffice (Koch & Ullman 1985, Tsotsos et al. 1995). Although the question of what form this representation takes is not yet fully answered, nor is it known whether a biological correlate to WTA exists, it does appear that a centralized saliency representation would require far greater connectivity than is observed in any visual area while a distributed representation of saliency does not exhibit this problem. A centralized AC faces the same problems that a single centralized representation of saliency does because it subsumes such a representation.

What is the connectivity required for AC to communicate its decisions to the visual processing hierarchy? Due to the dual requirements of modulation of the pathway

that an attending stimulus takes through the hierarchy and ensuring that all neurons along the path are attending to the same stimulus, AC must have feedback connections to each neuron in each layer of the processing hierarchy.

In the selective tuning model, additional circuitry is inserted into the network of interpretive units taking advantage of the locality of information and minimizing both connection length and transit time for moving information from its source to attentional control. This circuitry corresponds to the red and green portions of the circuit in Figure 4. Of course, temporal synchronization signals are required and this is the role played by the gating network as shown in that figure. Note that the temporal coordination signals have an oscillatory nature. Spatial coordination is accomplished due to the properties of the top-down selection algorithm. In the AC model, the same functionality is possible but the resulting architecture is far more expensive. First, there is an additional neural area, whose location is unknown at this point. There are additional long connections (actually connections to and from AC and each neuron in the hierarchy). There is additional delay required to move information to and from AC, and enormous convergence of connections in some representation in AC. There is one prediction of the selective tuning model that can provide some evidence in favor of the localized top-down model as opposed to the central model. The AC model has the potential of offering a single synchronized attentional modulation signal for the entire hierarchy. In other words, attentional modulation could be applied simultaneously to all the layers. Thus, modulation would be seen in V1 neurons at the same time as for IT neurons. The selective tuning approach would yield a different pattern, with modulation appearing first in the more abstract layers (IT) and later in earlier layers, the latency being dependent number of synapses to the layer. Evidence, however, seems unclear on this point. Luck et al. (1997) report that neurons in V4 show attentional modulation at about 75ms after stimulus onset while V2 neurons in the same condition show modulation after 100ms. Roelfsema et al. (1998) observed an attentional latency in V1 of 200ms for a curve-tracing and saccade task. These seem to support the top-down approach. In contrast, Chelazzi et al. (1993) report IT attentional modulation has a latency of about 200ms. In the first experiment, monkeys were tested with multiple stimuli within a single receptive field and no eye movements were involved while in the second and third the stimuli were not within one receptive field and the monkey responded by making a saccade. The results are not directly comparable; the time course of attentional effects derived from the selective tuning model is relevant for the former situation only. If it can be established that the onset of attentional modulation for all visual areas does not follow a top-down order then this would rule out the localized, distributed control of the selective tuning model. It should be noted that if Figure 7 is followed, a path exists from AIT to V1 with only one intermediate area, V4. Thus, the timing differences may not be large.

Although the above argument does not 'prove' that attentional control must be internal, local and distributed, it does demonstrate that such a control strategy is much more efficient. In fact, it is common in design to assume that the optimization principle of minimum cost (Tsotsos 1990) is imposed and if it can be assumed that evolution expresses such a preference, then it is clear that centralized external control is not the preferred strategy. In addition, the localized, top-down approach makes specific predictions that may help in discriminating between the opposing viewpoints.

3. Psychophysical Investigations

In this section an experimental investigation of the spatial variation of attention across the visual field is presented . The term *attentional field* denotes the dependency of the intensity of visual attention on two-dimensional visual field location. This representation is deliberately simplified; there is no interest here in the dependence of attention on time or stereoscopic depth, for example.

The goal was to map the variation of the attentional field around a target and discriminate between the predictions of the traditional models and of the selective tuning model. Recent psychophysical and neurophysiological studies report evidence for an inhibitory zone surrounding the attended target, in seeming agreement with the model. In the experiments reported in (Bahcall & Kowler 1999) subjects were required to identify two target letters in a circular display of distractor letters. Contrary to the prediction of the traditional models of selective attention, it was observed that recognition performance actually improved with increasing spatial separation between the targets.

In (Caputo & Guerra 1998) the target, the distractor and the non-target elements were arranged in a circular display. Both the target and the distractor stood out from the rest of the display: the target was the form pop-out and the distractor was the color pop-out. Subjects had to discriminate the length of a longer line segment included in the target. The gist of the results was that discrimination performance decreased with decreasing distractor-target distance.

Evidence for this lateral inhibition type of effect comes also from neuroscience. In (Schall & Hanes 1994) neurons in an area involved in generating intentional eye movements (the frontal eye field) of rhesus monkeys performing a visual search task were recorded. It was found that these neurons initially respond equally to both targets and distractors located in their receptive fields. However, while the neuronal response to the target continued until the saccade to the target, the response to the distractor was suppressed, and more so when the target was closer to the receptive field of the neuron. Furthermore, it was recently demonstrated by means of functional MRI (Kastner et al. 1998) that the cortical representations of stimuli appearing simultaneously in the visual field interact suppressively.

3.1 Experiments

The principle of the experimental method was the following: direct the subjects' attention to a *reference* location in the visual field and concomitantly measure their ability to process visual information -- i.e., the intensity of the attentional field -- at different, probe locations of equal retinal resolution. By systematically varying the reference - probe distance, one can determine the dependence of the attentional field on distance to the focus of attention.

The experimental requirements were threefold. (1) Engaging visual attention: the classical L - T discrimination task was used. Discrimination accuracy was employed as performance measure. (2) Directing the attention of the subject to one, pre-specified, reference target location: we resorted to pre-cueing. (3) Ensuring equal retinal resolution for all stimuli: we used a circular array display with fixation point in the center.

A typical experimental sequence, shown in Figure 8 from left to right, consisted of cue image, test image, and mask. The cue, a light gray disk, anticipated the position of

the reference target in the following test image. This will be referred to as the peripheral cue condition. It was shown for 180 msec, which is within the time range of effective cueing.

The stimulus set in the test image consisted of six randomly oriented Ls and six randomly oriented Ts, arranged in random order on a ring. The characters were evenly spaced, and were overlaid on light gray disks as shown in Figure 8, middle panel. Two of the characters, the reference target and the probe target, were red (shown in the figures in bold) and the rest, the distractors, were black. The orientation of the imaginary line joining the two targets was randomly changed from trial to trial. The radius of the ring was 4˚ and character size was 0.6˚ visual angle.

The task of the subject was to decide whether the two red characters were identical or different by pressing one of two keys on the computer keyboard. After 200 msec the test image was replaced by a mask consisting of red disks positioned at the locations of the disks enclosing the L and T characters in the preceding test image. The role of the mask was to erase the iconic memory of the target letters in the test display. It was during the mask that the subjects made their response. To ensure that all characters in the ring were perceived at same resolution, the subjects were instructed to always fixate the cross-hair in the center of the ring.

The main variable of interest in this experiment was inter-target separation, taking on six distinct values, from one, when the two target characters were next neighbors, to six, when the targets were diametrically opposite. Each of the six inter-target separations was tested four times in the same condition and four times in the different condition. Thus, an experimental session included 24 same and 24 different trials.

A control condition was employed to demonstrate the importance of the peripheral cue. The control consisted in cueing the fixation cross rather than the reference target. This was achieved by changing the color of the fixation cross from black to red 180 msec before the test screen. This central cue, being equidistant to all characters, could act only as a non-specific, temporal cue. This will be referred to as the central cue condition. The test and mask images were identical to the ones in the peripheral cue condition.

*********Figure 8******

3.2 Results

Eight subjects were employed, in both peripheral cue and central cue (control) experiments. The data of all subjects were pooled. Subject performance depended on two variables: inter-target separation and the orientation of the (imaginary) inter-target line segment. The random orientation values ranged from -180˚ to 180˚ and were binned in intervals 45˚ wide, resulting in eight levels for the orientation variable. For a given inter-target separation, all trials (from all subjects) falling within a given orientation interval were pooled, and processed to derive the mean accuracy rate characterizing the corresponding (separation, orientation) pair. Mean accuracies and standard deviations were determined for each separation level by averaging over the different orientations. Also, mean accuracies and standard deviations were determined for each orientation level by averaging over the different separation values.

Dependence of accuracy on inter-target separation The results of the central cue condition experiment are shown in Figure 9a while those of the peripheral cue condition are presented in Figure 9b. In the peripheral cue condition, the average accuracy in the same trials was 0.60, and in the different trials, 0.66. In the central cue condition average accuracy in the same trials was 0.5, and in the different trials, 0.55. The aspect of the error rate plots suggested that in the peripheral cue condition, but not in the central cue condition, accuracy increases with distance between the targets. Indeed, ANOVA tests confirmed that the differences in mean accuracy are significant ($F=4.5$, $p < 0.01$) for the peripheral cue condition, and insignificant for the central cue condition. In the peripheral cue experiment accuracy gradually increased with the distance between the targets; this pattern of variation is in direct contradiction to the prediction of the spotlight model and in agreement with the results reported in (Bahcall & Kowler 1999). However, in the control experiment, accuracy did not vary systematically with inter-target separation. Since the test images were identical in the peripheral and central cue conditions, the difference in performance must be attributed to the differences in cueing, and represents, therefore, a genuine attentional effect.

******* Figure 9 configured as a above c, b above d and a beside b, c beside d *****

Dependence of accuracy on orientation of the line joining the reference and probe targets The orientation of the imaginary line joining target and probe was considered as a potential confounding element and was also tested for both control and peripheral cue conditions; ANOVA tests indicated that in neither condition were the accuracy differences significant. The absence of any effect for the orientation of the imaginary line joining the two targets is reassuring. In preliminary experiments in which we tested subjects with upright letters, we obtained a strong effect of orientation: the best performance occurred when the two characters were at the same height on the disk, suggesting that the subjects were actually reading the letters, and not simply processing them as shapes without any particular meaning.

Dependence of response time on inter-target separation Figure 9c and d show the response time profiles versus inter-target separation showing that poor accuracy in the peripheral cue condition is not a result of reduced processing time on the part of the subject. In fact, subjects take longer to respond in order to achieve the low accuracy observed.

3.3 Discussion

These results can be interpreted as follows. In the main experiment, the cue centers attention on the reference target and thus in contrast to the Bahcall & Kowler and Caputo & Guerra experiments, the focus of attention for each trial is known. As a consequence of the inhibitory-surround structure of the attentional field, the ability of the subject to discriminate a probe target increases with distance from the cued position. When the center of the display is cued all characters, being at equal distance from the center of the attentional field, are equally inhibited. Consequently, the distance between targets does not affect accuracy of discrimination. Future experiments will test the effect of cue size on the attentional field, and will try to provide a more accurate measure of the size of the inhibitory surround with finer spacing of probe locations.

Conclusions

There are at least two strategies for modeling biological information processing. The most common approach is to develop a mathematical framework (in its simplest form, fitting curves to sets of data) that can account for experimental data (the 'data fitting approach'). Such models may even provide some predictive power for experiments of the same type. The second strategy is to develop a model from first principles of information processing, without direct incorporation of any particular data sets (the 'first principles approach'). If the model is defined appropriately not only is it possible to have the same explanatory power as the data fitting approach, but also there are at least two other major benefits. For vision, the data-fitting approach does not directly lead to an algorithm that can take images as input and produce the measurements being modeled whereas the first principles approach does. Thus, in a very real sense, data fitting solves only a part of the problem of understanding the nature of information processing that leads to the data. Secondly, the first principles approach has much broader predictive power because it makes no early commitment to a particular experimental paradigm, a necessary ingredient of the data fitting approach.

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. The predictive power of the model seems broad:

- An early prediction (Tsotsos 1990) was that attention seems necessary at any level of processing where a many-to-one mapping of neurons was found. 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 (Moran & Desimone 1985). 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 (Kastner et al. 1998). Evidence cited by Britten (1996) who reached the conclusion that 'attention is everywhere', save for the Moran and Desimone work, was all post-1990. Vanduffel et al. (in press) have shown that attentional modulation appears at early as the LGN.
- The notions of competition and of attentional inhibition were also early components of the model (Tsotsos 1990) and this too has gained evidence over the years (Desimone & Duncan 1995; Kastner et al. 1998; Reynolds et al. 1999).
- The model has always included an inhibitory surround component (Tsotsos 1990). This implies that perception may be negatively affected in the vicinity of the attended stimulus. This too has recently gained support (Caputo & Guerra 1998; Bahcall & Kowler 1999; Vanduffel et al. in press, and the experiments described in Part 2 of this paper).
- The Tsotsos 1990 paper also explained how so-called pre-attentive vision was only a special case of attentive processes; no separate pre-attentive process operates independently of attention, a view Joseph et al.(1997) seem to be suggesting too.

- A prediction that is still unconfirmed is that attentional control is not centralized in some particular brain structure outside of the visual hierarchy that provides all control signals, but rather is local, distributed and internal to the processing hierarchy (as discussed in section 2.2). This is not to say that there is no input to those control circuits from outside the visual processing hierarchy. Indeed, at least information about the task must be communicated from another source.

Additional predictions are made throughout the present paper:

- Figure 3 summarizes a pattern of sub-regions within visual areas where neural increases or decreases due to attentional modulation can be expected. The pattern has a spatial as well as a temporal component.
- In Figure 5, a direct link is made between the WTA circuit and the pyramidal and spiny stellate cells of V1 (described by Callaway, 1998, as participants in periodic clusters) connecting columns of similarly tuned neurons. The suggestion here is that the functionality of these cells (or perhaps of others with similar connectivity patterns) is the gating function of the models' WTA networks.

Based on theoretical analyses, computer simulations and human psychophysics we conclude that the particular circuit we propose does indeed capture many of the functional characteristics of attentive selection. We are actively working on extending the model in several directions. It would be interesting to see how the circuit details might be mapped onto actual neural circuitry.

References

Bahcall, D., Kowler, E., (1999). Attentional Interference at Small Spatial Separations, *Vision Research 39(1)*, p 71 - 86.

Britten, K., (1996). Attention is Everywhere, *Nature 382*, p. 497 - 498.

Callaway, E., (1998). Local Circuits in Primary Visual Cortex of the Macaque Monkey, *Annual Review of Neuroscience 21*, p47 - 74.

Caputo, G., Guerra, S., (1998). Attentional Selection by Distractor Suppression, *Vision Research 38(5)*, p. 669 - 689.

Chelazzi, L., Miller, E., Duncan, J., Desimone, R., (1993). A Neural Basis for Visual Search in Inferior Temporal Cortex, *Nature* , Vol. 363, p 345 - 347.

Corbetta, M., (this volume).

Crick, F., (1984). Function of the Thalamic Reticular Complex, *Proc. National Academy of Science USA 81*, p4586 - 4590.

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

Desimone, R., Duncan, J., (1995). Neural Mechanisms of Selective Attention, *Annual Review of Neuroscience 18*, p193 - 222.

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

He, S., Cavanaugh, P., Intrilligator, J., (1996). Attentional Resolution and the Locus of Attentional Awareness, *Nature 283*, p 334 - 337.

Joseph, J., Chun, M., Nakayama, K., (1997). Attentional Requirements in a 'Preattentive' Feature Search Task, *Nature 387*, p. 805 - 807.

Kastner, S., De Weerd, P., Desimone, R., Ungerleider, L., (1998). Mechanisms of

Directed Attention in the Human Extrastriate Cortex as Revealed by Functional MRI, *Science 282*, p108 - 111.

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

Luck, S. Chelazzi, L., Hillyard, S., Desimone, R., (1997). Neural Mechanisms of Spatial Selective Attention in Areas V1 V2 and V4 of Macaque Visual Cortex, *J. Neurophysiology 77*, p 24 - 42.

Maunsell, J., Ferrera, V. (1995). Attentional Mechanisms in Visual Cortex, in **The Cognitive Neurosciences**, ed. by M. Gazzaniga, p. 451 - 461, MIT Press.

Milner, P. (1974), A Model for Visual Shape Recognition, *Psychol. Rev. 81*, p521-535.

Moran, J., Desimone, R. (1985). Selective Attention Gates Visual Processing in the Extrastriate Cortex, *Science 229,* p 782 - 784.

Motter, B., (1993). 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.

Olshausen, B., Anderson, C., Van Essen, D (1993). A Neurobiological Model of Visual Attention and Invariant Pattern Recognition based on Dynamic Routing of Information, *J. Neuroscience 13*, p 4700 - 4719.

Reynolds, J., Chelazzi, L., Desimone, R., (1999). Competitive Mechanisms Subserve Attention in Macaque Areas V2 and V4, *The Journal of Neuroscience, 19(5)*, p1736–1753

Roelfsema, P., Lamme, V., Spekreijse, H., (1998). Object-based Attention in the Primary Visual Cortex of the Macaque Monkey, *Nature 395*, p 376 - 380.

Salin, P.-A., Bullier, J., (1995). Corticocortical Connections in the Visual System: Structure and Function, *Physiological Reviews, 75(1)*, p107 - 154.

Schall, J., Hanes, D. (1993). Neural Basis of Saccade Target Selection in Frontal Eye Field during Visual Search, *Nature 366*, p 467 - 469.

Treue, S., Maunsell, J., (1996). Attentional Modulation of Visual Motion Processing in Cortical Areas MT and MST, *Nature 382*, p539 - 541.

Tsotsos, J. K. (1990). A Complexity Level Analysis of Vision. *Behavioral and Brain Sciences*, 13(3), p423 - 455.

Tsotsos, J.K., (1993) An Inhibitory Beam for Attentional Selection, in **Spatial Vision in Humans and Robots**, ed. by Harris & Jenkin, p 313-331, Cambridge University Press.

Tsotsos, J. K., Culhane, S., Wai, W., Lai, Y., Davis, N., Nuflo, F. (1995). Modeling visual attention via selective tuning, *Artificial Intelligence 78*, p507 - 545.

Valiant, L., (1975). Parallelism in Comparison Problems, *SIAM J. Comput. 4(3)*, p. 348 - 355.

Vanduffel, W., Tootell, R., Orban, G. (in press). "Attention-dependent suppression of metabolic activity in the early stages of the macaque visual system, *Cerebral Cortex.*

Captions

Figure 1. (a) Feedforward activation of the visual processing pyramid and (b) its modulation after attentional selection has been applied. Red connections are those affected by the stimulus, gray connections are those which play no role, while black connections are those inhibited by the WTA selection process. Note how the top layer is not inhibited by the top-layer WTA and thus the feedforward divergence of the stimulus is seen to the output layer. If it were inhibited, then no other stimulus could reach the output layer making the system effectively blind to all non-attended stimuli. The model predicts that non-attended stimuli do reach the output layer of the system but their representation may be incomplete or corrupted by interfering signals (Tsotsos 1997). The shading of the units' colour reflects the assumption that unit weighting profiles are Gaussian in nature.

Figure 2. A four-step sequence showing attentional modulation when there are two stimuli in the input and the system attends to one. (a) 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. (b) The location selection is applied and two units in the output layer are identified (location cues can be placed anywhere in the visual field prior to a test stimulus). The first WTA stage then takes place and the largest responses within the next layer receptive fields of the selected units are found. The connections not corresponding to those largest response units are inhibited. (c) The results after the second stage of WTA. (d) The results after the third and final stage of WTA. Due to the complexity of the figure, the variations in unit strength due to the Gaussian weighting profile are not shown.

Figure 3. Modulation predictions. Following the changes of a particular unit of the pyramid through the 5-step sequence of Figure 3 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 on which side of the attended stimulus they fall) 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 Figure 3 is to select a specific unit in the pyramid in Figure 3a and track its changes over time as depicted in the sequence from Figure 3a through to Figure 3d.

Figure 4. The circuit that implements the hierarchical selection described in the text is shown; this is a conceptual view and is not intended to correspond to specific neurons and their connectivities. A more detailed explanation of this circuit can be found in (Tsotsos et al. 1995).

Figure 5. WTA circuit proposals. (a) Two layers of visual processing are shown and the feedforward and feedback divergence of connections highlighted. The patterns of connectivity overlap exactly. The expanded section of the feedforward convergence in the left-hand side of the figure shows two possible implementations of the WTA circuit.

If the WTA is to be implemented in strictly parallel, distributed processing fashion, each unit must be connected to each other in the competition (c). A central processing implementation is functionally equivalent (b).

Figure 6. Example of computer simulation. In this example, an image of several coloured blocks is the test image. The algorithm is instructed to search for blue regions, and it attempts to do this by searching for the largest, bluest region first. This test image is shown on the right half of each image with the regions selected outlined in yellow with blue lines between them showing the systems scanpath. The left side of each image shows a 4-level visual processing pyramid. The instruction is applied to the pyramid to tune its feature computations and the result is that the regions within each layer of the pyramid that remain are those that are blue. The left side of (a) shows the set of blue of objects found. Then the WTA algorithm selects the largest, bluest one first, selected in part (b), inhibits that region (note it does not appear in part (c)), and then repeats the process 6 more times. Note that the system does not know about objects, only rectangular regions; thus, although it appears to select whole blocks sometimes, this is only due to fortuitous camera viewpoints.

Figure 7. The network of visual areas of the macaque where attentional modulation of the kind addressed by the selective tuning model has been observed are shown (V1 - Motter 1993; V2 - Motter 1993; V4 - Moran & Desimone 1985; IT - Chelazzi et al. 1993; MT - Treue & Maunsell 1996; MST - Treue & Maunsell 1996; FEF - Schall & Hanes 1993; LGN - Vanduffel et al. in press). This diagram is based on the Felleman and Van Essen (1991) diagrams and were created using their software. These are the minimal set of areas to which an attentional control center must connect and provide control signals.

Figure 8. Peripheral cue condition, experimental trial sequence. (a) The cue, a light gray disk indicated the position of the reference target character in the following test screen. Shown for 180 msec. (b) Test screen, shown for 200 msec. The target characters were red (drawn in this figure with thick lines), the distractors were black. The task is to decide whether the two target characters are same or different. (c) Mask shown until the subject responded.

Figure 9. Dependence of accuracy on inter-target separation. Abscissa: distance between the two targets. Ordinate: mean accuracy; standard deviations correspond to different orientation values. (b) Peripheral cue condition; there is a significant improvement in accuracy beyond an inter-target separation of 2 for this set of targets and cues. (a) Central cue (control) condition; subjects are cued to the fixation point. There are no significant performance differences with changes in separation. Response time vs. inter-target separation. (c) The control condition; there is no significant differences of response time as a function of separation when the subjects are cued to the fixation point. (d) The peripheral cue condition. Note that subjects spend more time on smaller inter-target separations and still are not as accurate as for large separations.

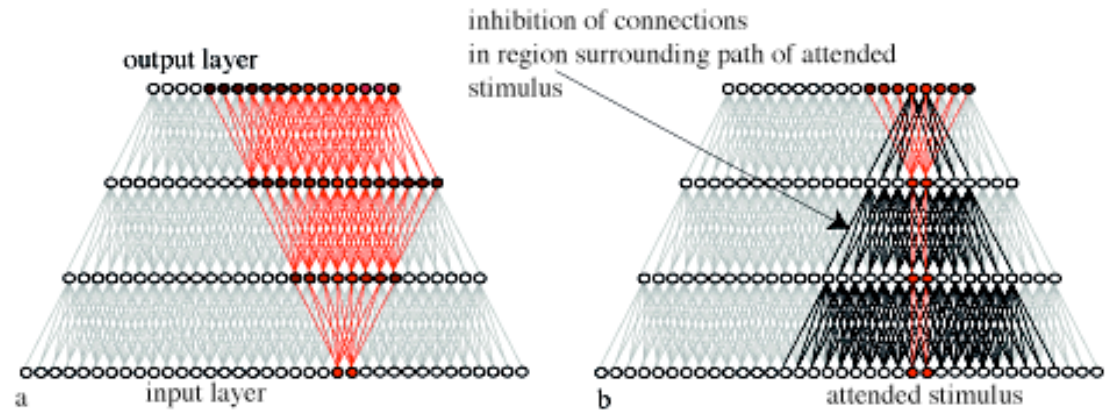
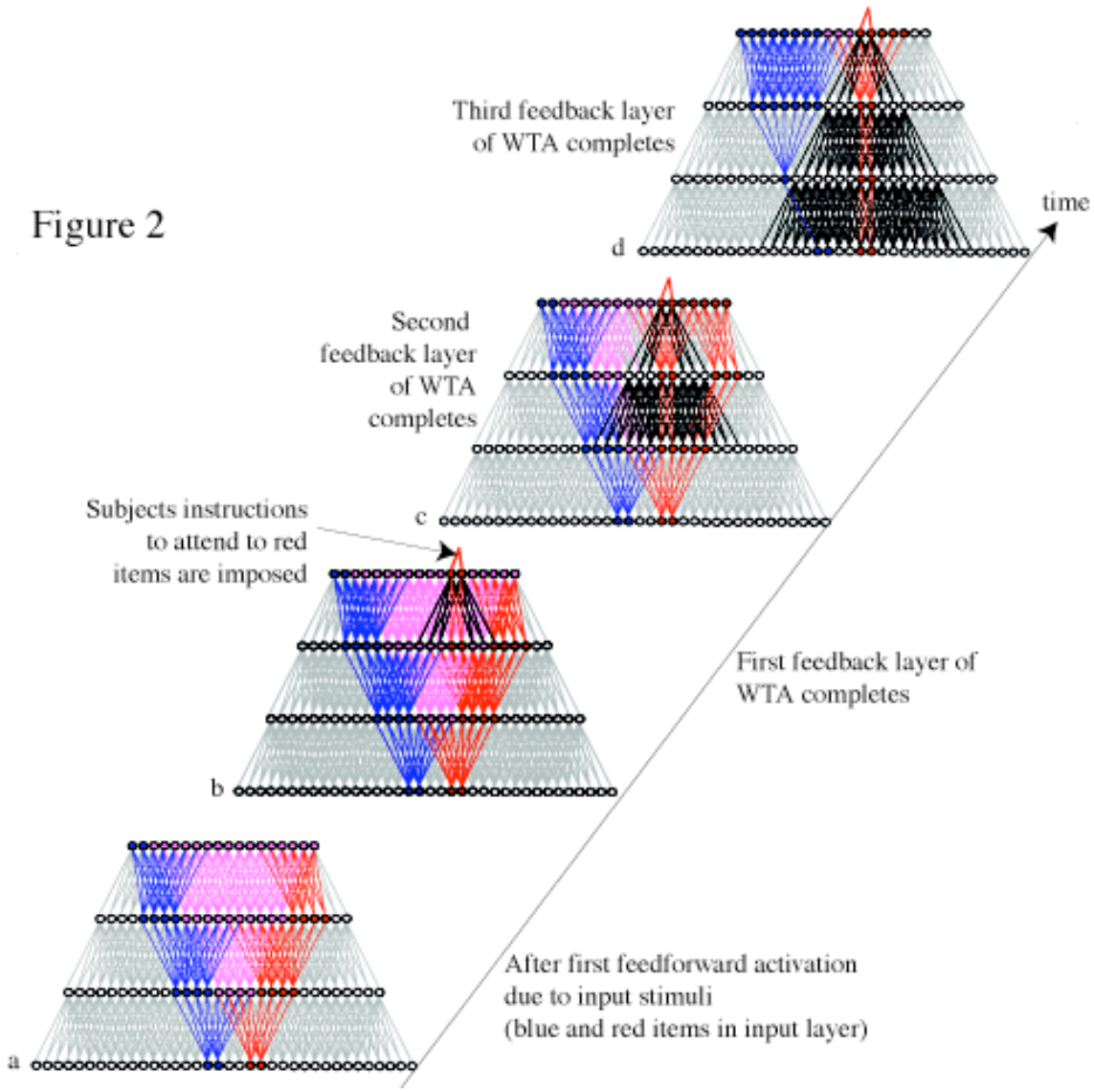


Figure 1

Figure 2



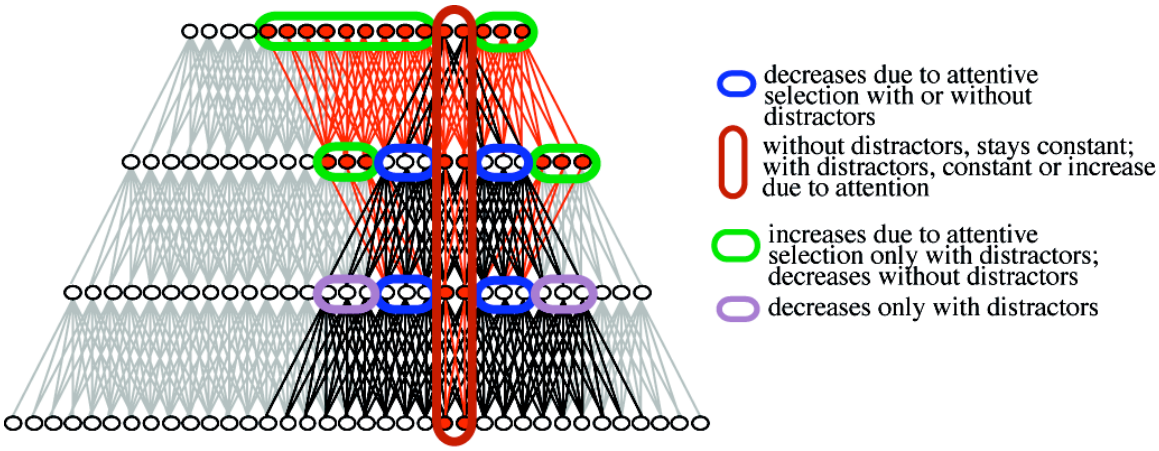


Figure 3

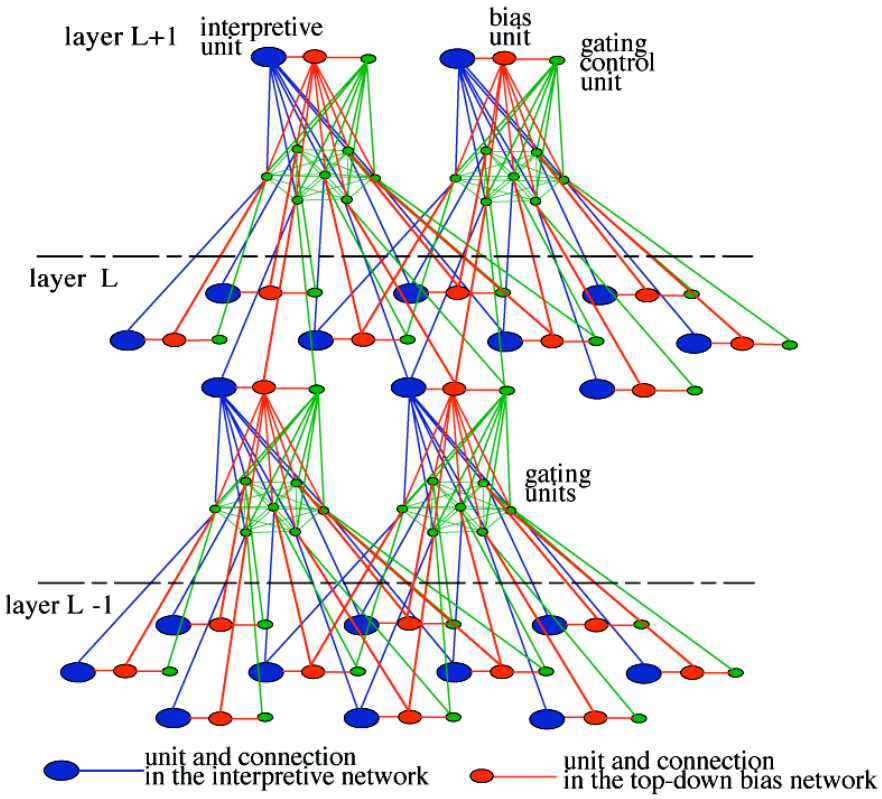


Figure 4

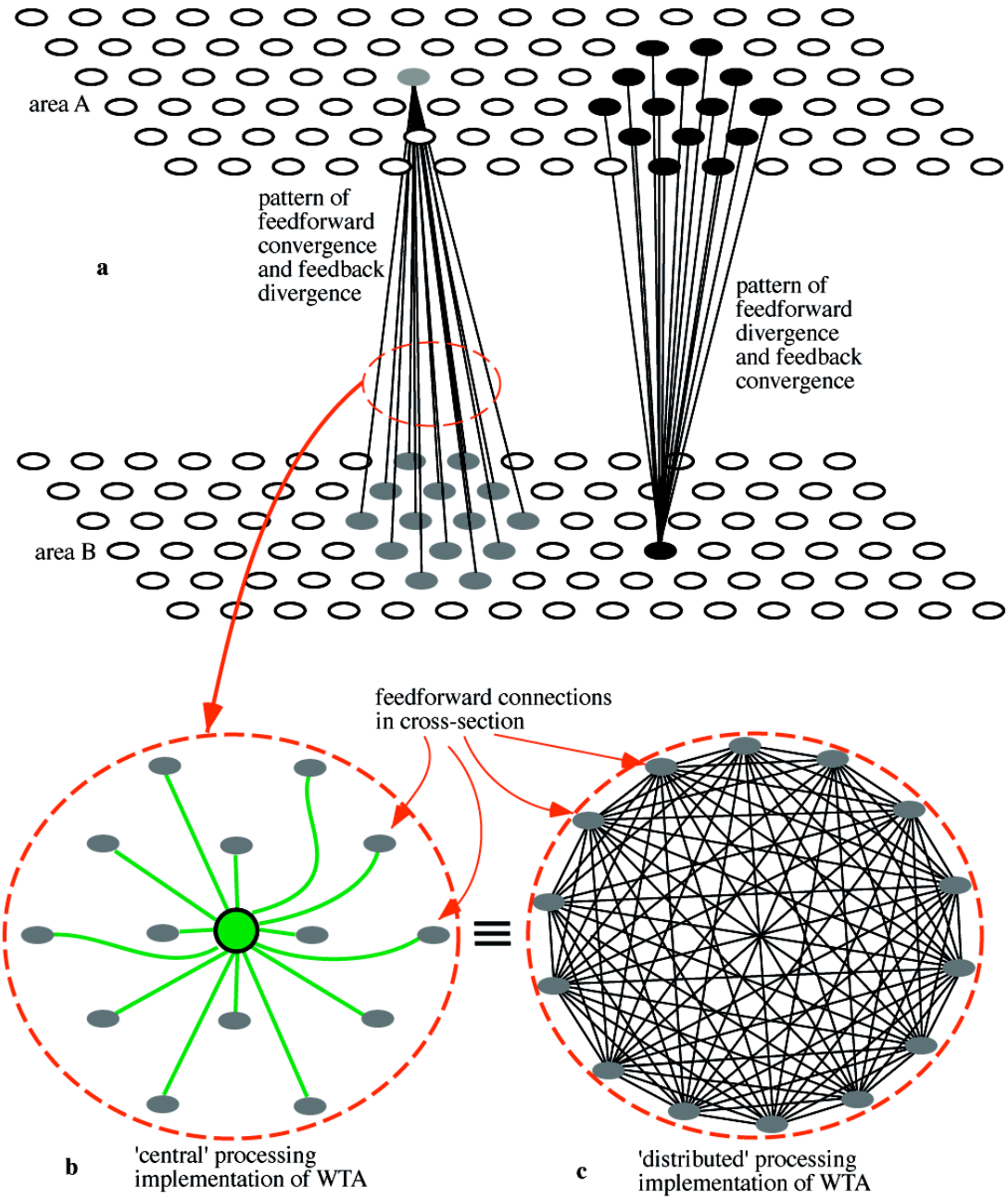

area A
pattern of
feedforward
convergence
and feedback
divergence
a
pattern of
feedforward
divergence
and feedback
convergence
area B
feedforward connections
in cross-section
b
'central' processing
implementation of WTA
c
'distributed' processing
implementation of WTA

Figure 5

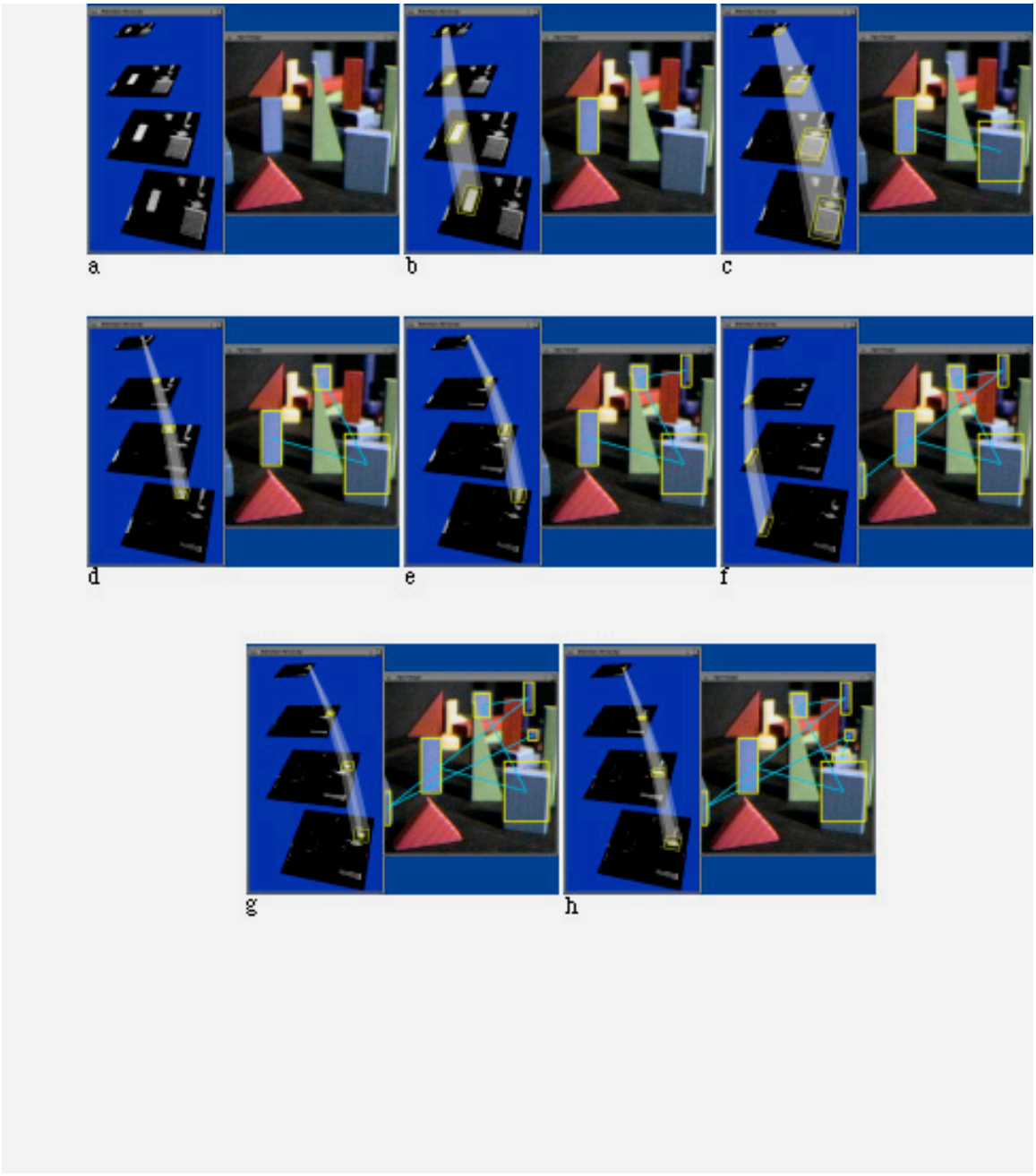


Figure 6

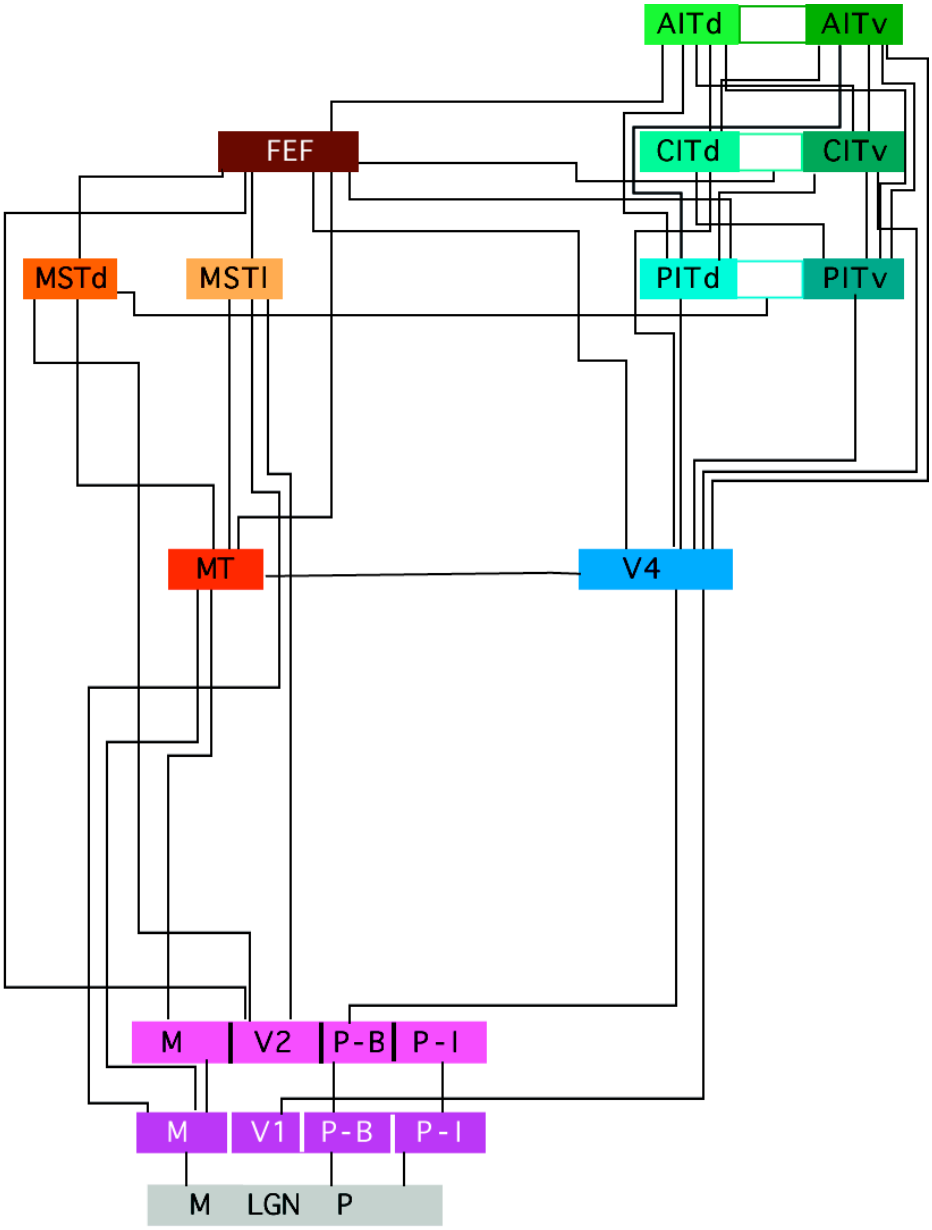

AITd
AITv
FEF
CITd
CITv
MSTd
MSTl
PITd
PITv
MT
V4
M
V2
P-B
P-I
M
V1
P-B
P-I
M
LGN
P

Figure 7

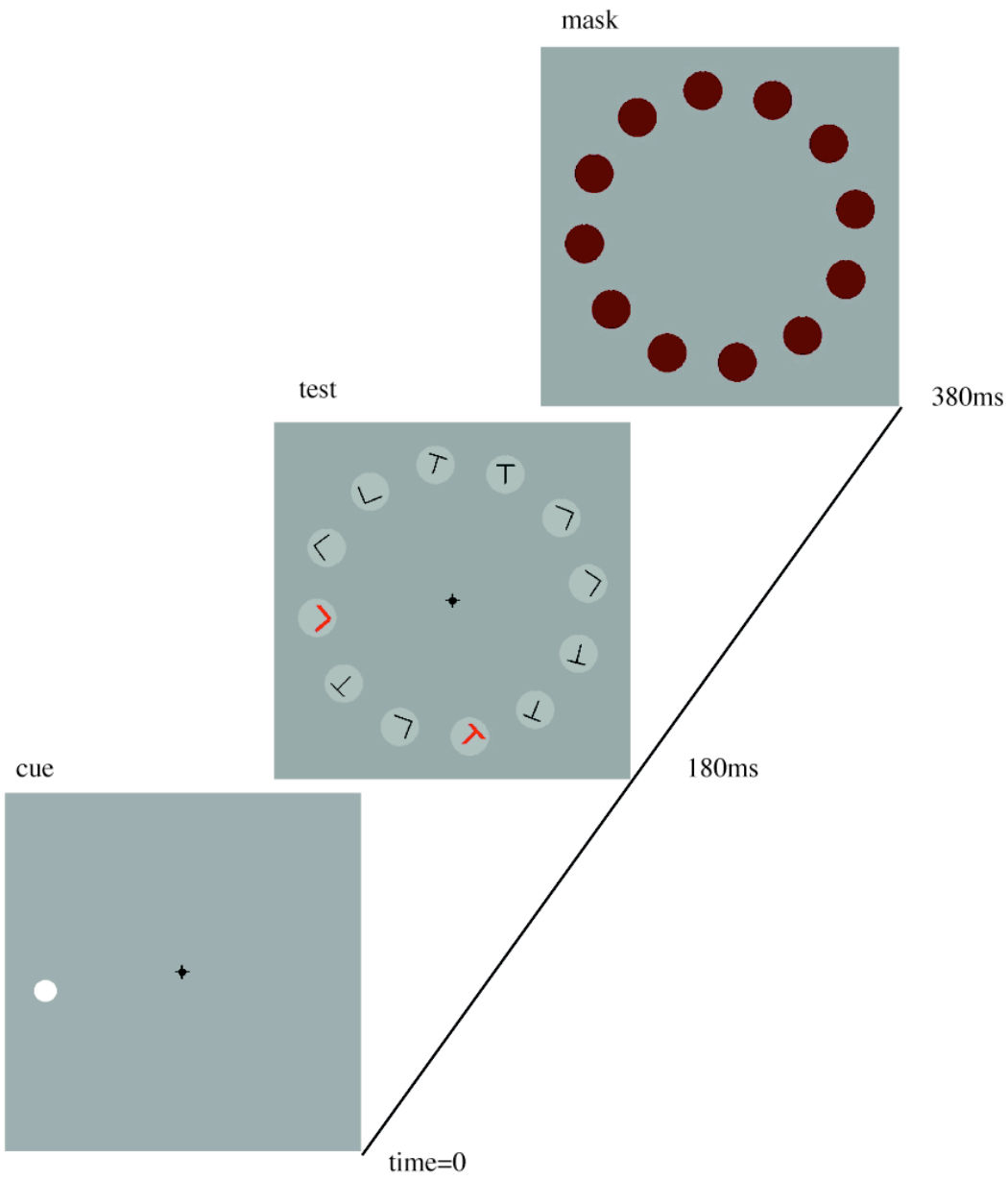


Figure 8

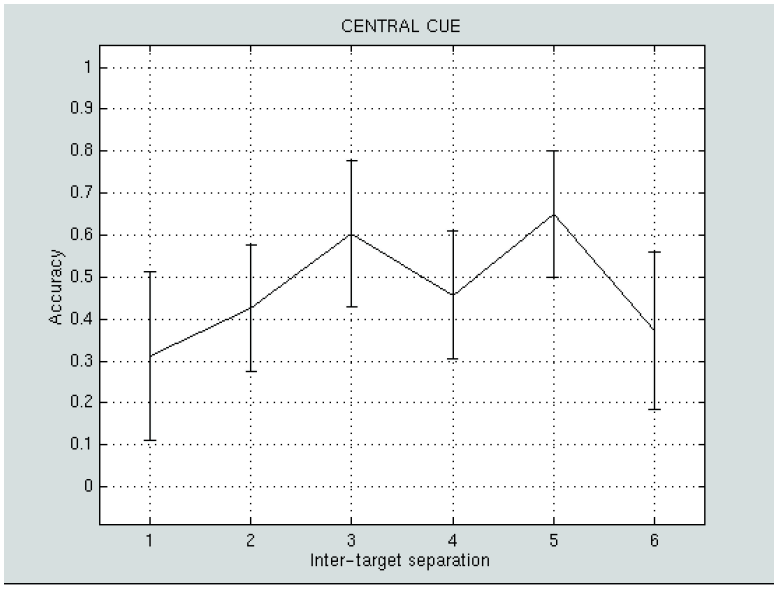


Figure 9a

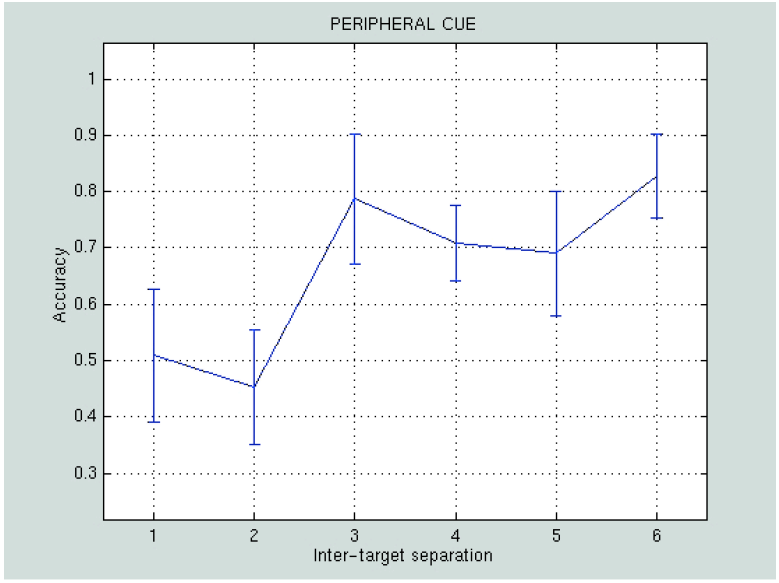


Figure 9b

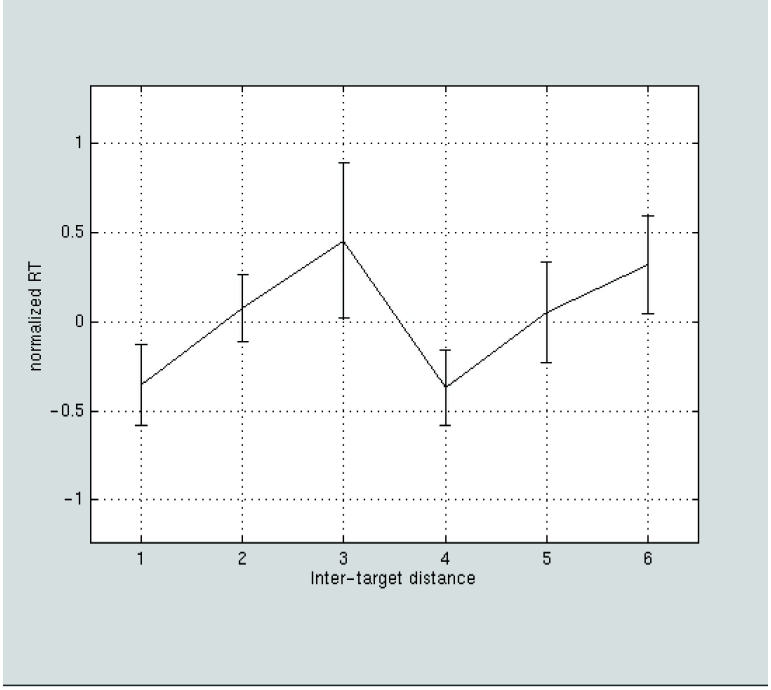


Figure 9c

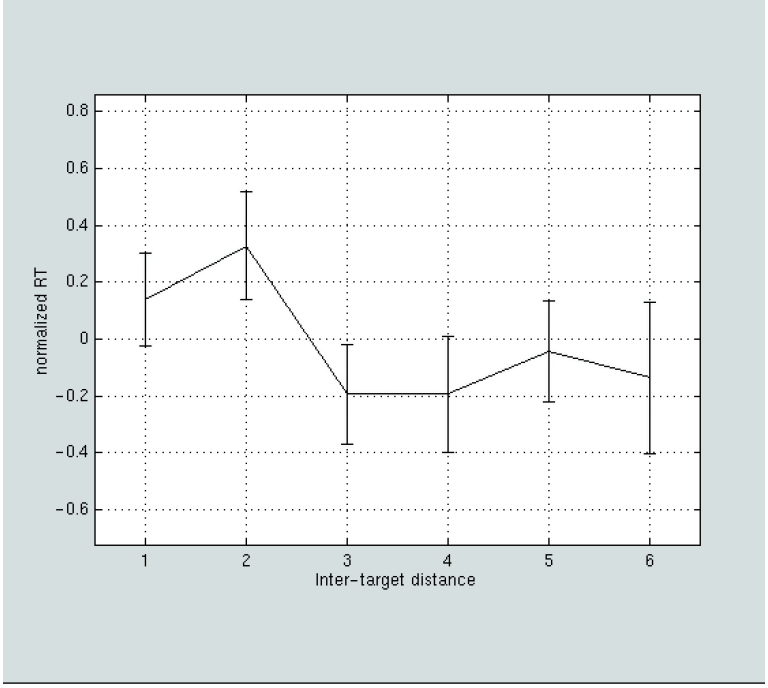


Figure 9d