YORK
UNIVERSITÉ
UNIVERSITY

redefine THE POSSIBLE.

Integrating attentional modulation with attentional selection

Albert L. Rothenstein and John K. Tsotsos

Technical Report CSE-2013-07

August 16 2013

Department of Computer Science and Engineering
4700 Keele Street, Toronto, Ontario M3J 1P3 Canada

Integrating attentional modulation with attentional selection

Albert L. Rothenstein and John K. Tsotsos

Dept. of Electrical Engineering and Computer Science, and

Centre for Vision Research,

York University, Toronto, Canada

albertlr@cse.yorku.ca, tsotsos@cse.yorku.ca

Abstract

Various models of the neural mechanisms of attentional modulation in the visual cortex have been proposed. In general, these models assume that an 'attention' parameter is provided separately. Its value as well as the selection of neuron(s) to which it applies are assumed, but its source and the selection mechanism are unspecified. Here we show how the Selective Tuning model of visual attention can account for the modulation of the firing rate at the single neuron level, and for the temporal pattern of attentional modulations in the visual cortex, in a self-contained formulation that simultaneously determines the stimulus elements to be attended while modulating the relevant neural processes.

Introduction

While visual scenes typically contain multiple objects, the capacity of the visual system to process all these objects at the same time is limited (Broadbent 1958, Neisser 1967, Schneider and Shiffrin 1977, Tsotsos 1990). In experimental settings, when presented with multiple objects, subjects' performance decreases and typical errors are present (Treisman 1969, Treisman and Schmidt 1982, Duncan 1980, Duncan 1984). Stimuli are said to compete for neural representation, and the mechanisms of this competition and their modulatory effect on neural responses have been a subject of intense investigation and modeling.

There have been several models of visual attentive neural modulations, and all assume that an external 'attention' parameter is provided by some other neural process. Its value as well as the selection of neuron(s) to which it applies are assumed, but its source and how all of this is determined are unspecified. Our theory, Selective Tuning (ST), presents a novel formulation that solves these problems (Tsotsos 2011). In ST, neuron response is the result of attentive modulation of its inputs across time from the whole network involving feedforward, recurrent, and lateral interactions. We show that not only does ST provide very good comparisons to single neuron firing rates in attentive tasks but also shows how, given visual stimuli, the stimulus locus of attention is computed and used

throughout the network. The goal of this paper is to show how this model captures the essence of attentive modulation as well as its competitors, while additionally adding the critically missing element of attentional computation. Importantly, ST goes beyond other models in defining a modulatory mechanism at a finer level of abstraction than previously accomplished; in a very real sense, it subsumes the other models as a result. Other aspects of ST are described elsewhere (Tsotsos, 2011).

The paper starts with a brief overview of ST and a review of computational models of attentional modulation. The microcircuitry of ST and the equations that govern its behaviour are described next, followed by two computational modeling experiments. We describe the experiments that form the foundation for the biased competition theory (Desimone and Duncan 1995, Reynolds and Desimone 1999), followed by a simulation of these experiments using ST. Additionally, the temporal latency of selective attention modulation across areas in ST is presented and compared to that of the macaque visual system (Mehta et al. 2000). These experiments demonstrate the fact that ST can account for the modulation of the firing rate at the single neuron level, and for the temporal progression of attentional modulations in the visual cortex in a self-contained formulation without the external attentional parameter. The paper concludes with a comparison of the specifics of the models considered, analyzing their relative strengths and weaknesses.

Background

A wide variety of models have appeared that exhibit interesting performance with respect to neural modulations due to attention. These will be overviewed first, followed by a description of the Selective Tuning model, the focus of this paper.

Models of Attentive Neural Modulation

A wide variety of models have been proposed to explain the attentional modulation of neural activations, in this section we briefly review a representative subset. Each model will be assigned an acronym for easier reference. The descriptions and equations included below are meant mainly to illustrate the wide variety of solutions proposed, and to highlight the way attentional modulation is implemented, rather than being exhaustive descriptions of the models. For full details, complete sets of equations and biological justification, the reader is referred to the original sources. For each of the models, there is no evaluation of their actual results presented because each in its own way shows good matches to data and/or behavior. As a result, the point of this comparison is to clarify commonalities, differences, gaps, and strengths. Here we will focus on how the different models present attentional neural modulation. For general reviews of theories of attention see (Pashler 1998a, Pashler 1998b, Itti et al. 2005, Tsotsos 2011), and specifically for computational modeling, see (Itti and Koch 2001, Rothenstein and Tsotsos 2008, Tsotsos and Rothenstein 2011).

The biased competition theory (Desimone and Duncan 1995, Reynolds and Desimone 1999) proposes that neurons representing different features compete and that attention

biases this competition in favor of neurons that encode the attended stimulus. The Biased Competition model (BCM) (Reynolds et al. 1999) has been proposed as a demonstration of the biased competition theory. The temporal evolution of a neuron's firing rate is given by:

$$\frac{dy}{dt} = (B - y)E - yI - Ay$$

where $E = x_1 w_1^+ + x_2 w_2^+$ is the excitatory input for two stimuli within the same receptive field, and $I = x_1 w_1^- + x_2 w_2^-$ the inhibitory input for two stimuli within the same receptive field. The *w*'s are positive and negative synaptic weights. The equilibrium response is described by $\lim_{t\to\infty} y = BE/(E + I + A)$, where *B* is the maximum response and *A* is a decay constant. Attention is assumed to increase the strength of the signal coming from the inputs activated by the attended stimulus, implemented by increasing the associated synaptic weights. The model makes no claims of biological plausibility – the equations are not good fits for neural responses, no actual competition is implemented, and there is no known mechanism for the multiplicative modulation of synaptic weights (Spratling and Johnson 2004).

A significant number of other models of biased competition have tried to build upon BCM, by filling in the missing biologically-plausible mechanism, by including additional neuron types, and by extending the model to full networks.

The Neurodynamical model (ND) (Rolls & Deco 2002) is a large-scale implementation of biased competition that consists of several interconnected network modules simulating different areas of the dorsal and ventral path of the visual cortex. Each module consists of a population of cortical neurons arranged in excitatory and inhibitory pools. The inhibitory pool receives excitatory input from all the excitatory pools and provides uniform inhibitory feedback to each of the excitatory pools, thus mediating competition between them. The temporal evolution of the system is described within the framework of a mean-field approximation, i.e. an ensemble average of the neural population is calculated in order to obtain the corresponding activity. For example, the current activity of the excitatory pools in the posterior parietal (PP) module is given by:

$$\tau \frac{\partial}{\partial t} I_{ij}^{PP}(t) = -I_{ij}^{PP}(t) + aF\left(I_{ij}^{PP}(t)\right) - bF\left(I^{PP,I}(t)\right) + I_{ij}^{PP-V1}(t) + I_{ij}^{PP,A} + I_0 + \nu$$

where $I_{ij}^{PP,A}$ is an external attentional spatial-specific top-down bias, I_0 is a diffuse spontaneous background input, υ is Gaussian noise. The intermodular attentional biasing I_{ij}^{PP-V1} through the connections with the pools in the module V1 is:

$$I_{ij}^{PP-V1}(t) = \sum_{k,p,q,l} W_{pqij} F\left(I_{kpql}^{V1}(t)\right)$$

and the activity current of the common PP inhibitory pool evolves according to:

$$\tau \frac{\partial}{\partial t} I^{PP,I}(t) = -I^{PP,I}(t) + c' \sum_{i,j} F\left(I_{ij}^{PP}(t)\right) - dF\left(I^{PP,I}(t)\right)$$

Similar equations govern the dynamics of the other models. The model asserts that feature attention biases intermodular competition along the ventral pathway (simulated areas V4 and IT), and spatial attention biases intermodular competition along the dorsal pathway (simulated areas V1, V4 and PP).

The Feedback Model of Visual Attention (FMVA) (Spratling & Johnson 2004) improves on BCM by providing a biologically-justified mechanism and microcircuitry for input modulation. Compared to most other implementations of biased competition, in which neurons compete by inhibiting each other's output, in FMVA neurons compete by laterally inhibiting other neurons' inputs. The key observation that drives the model is that feedforward connections seem to be primarily made in basal dendrites, while feedback connections preferentially target apical dendrites, thus appearing to have functionally different roles. The activations of the apical and basal dendrites are given by:

$$y^t_{jk,apical} = \sum_{i=1}^{m_a} v_{ijk} x^{t-1}_{ijk} \quad \text{and} \quad y^t_{jk,basal} = \sum_{i=1}^{m_b} w_{ijk} X^{t-1}_{ijk}$$

where j and k are indices used to localize the neuron in the network, m_a and m_b are the total number of apical and basal synapses, respectively, v_{ijk} and w_{ijk} are synaptic weights associated with the input from neuron i (apical and basal, respectively), x is the output of neuron i. The apical inputs x^t_{ijk} originate from higher cortical regions or are top-down signals from outside the model. The basal inputs X^t_{ijk} are the outputs of neurons in lower cortical regions, after pre-integration lateral inhibition:

$$X^t_{ijk} = x^t_{ijk} \left\{ 1 - \alpha^t \max_{\substack{p=1 \\ (p \neq j)}}^{n} \left[\frac{w_{ipk}}{\max_{q=1}^{m}\left(w_{qpk}\right)} \frac{y^{t-1}_{pk}}{\max_{q=1}^{n}\left(y^{t-1}_{qk}\right)} \right] \right\}^+$$

where α^t scales lateral inhibition, y^{t-1}_{pk} is the output of neuron p in region k at the previous time step, and $\{z\}^+$ the positive half-rectified value of $\{z\}$. FMVA proposes that feedback activations multiplicatively modulate the total feedforward activation:

$$y^t_{jk} = y^t_{jk,basal}\left(1 + y^t_{jk,apical}\right) .$$

The reentry hypothesis (RH) (Hamker 2005) models top-down modulation as a gain control mechanism on the input feedforward signal, increasing activations that match top-down predictions. Considering two interconnected areas I and II, the generic formulation of the modulation of the input signal $I^{II}_{d,k,x}$ to area II combines the filtered feedforward input $F\left(r^I_{d,i,x'}\right)$ from area I with the summed top-down signal $\hat{r}^{\gamma}_{d,i,x'}$ where $\gamma \in \{L,F\}$ is the origin of the attentional signal (L for location, F for feature):

$$r^{II}_{d,k,x} = w^{\uparrow} \cdot f\left(F\left(r^{I}_{d,i,x'}\right)\right) + \sum_{\gamma \in \{L,F\}} \left(A - r^{II}_{d,k,x}\right)_{+} \cdot f\left(F\left(r^{I}_{d,i,x'}\right) \cdot \tilde{r}^{\gamma}_{d,i,x'}\right)$$

The nonlinear pooling function *f* defines the influence of the filtered afferents *F* on cell *k*. This generic equation is adapted to the specific connectivity of each area. The multiplicative gain allows for multiple forms of attentional selection, by combining signals originating from different areas (e.g. memory for stimulus-specific features, motor maps for location specific feedback) that allow the system to simulate a variety of experimental tasks.

An alternative account for attentional modulation comes in the form of the "feature similarity gain" theory (FSG) (Treue & Martinez-Trujillo 1999, Martinez-Trujillo & Treue 2004), according to which attention can both enhance and reduce neural activations in proportion to the similarity between the attended stimulus and the preferred stimulus of the neuron. The attentional gain effect on neuronal responses is a graded function of the difference between the attended feature and the preferred feature of the neuron, independent of the stimulus. In the computational model of FSG proposed by Boynton (2005), the neural response is described as a divisive contrast normalization process between the sum of squared linear responses to each stimulus component and the sum of squared contrasts plus a semi-saturation constant σ by:

$$H(x_i, c_i) = \frac{\sum_i (c_i F(x_i))^2}{\sum_i c_i^2 + \sigma^2}$$

where each component that contributes to the activation has its own feature x_i and contrast c_i. The normalized firing rate *H* is modulated by a gain factor *G(y)* that has a tuning function similar to the stimulus-driven tuning function of that neuron, where *y* is the attended feature, resulting in a modulated response given by :

$$R(x_i, c_i, y) = G(y)\left[H(x_i, c_i) + \delta\right].$$

The parameter δ is the inherent baseline-firing rate of the neuron. The gain factor *G(y)* is greater than one for a preferred feature, and lower than one otherwise. The gain factor is a purely feature-based effect, and is independent of the spatial focus of attention and the properties of the visual stimulus.

Yet another class of models start from the observation that the neural representation of multiple concurrent stimuli is equivalent to a normalization operation (Busse et al. 2009, Carandini & Heeger 2012). As normalization is viewed as a fundamental, canonical neural computation, the authors hypothesize that attention has an impact on neural activations by influencing the normalization process.

The Normalization Model of Attention (NMA) (Reynolds and Heeger 2009) is an attempt to unify under a single computational model disparate results that are consistent with attention as a multiplicative gain factor, as a change in contrast gain, a sharpening of neural tuning, and various forms of attenuation and enhancement. The proposed model

combines neural selectivity (termed "stimulus drive") with an external "attention field" and a "suppressive field", that pools activations corresponding to non-preferred stimulus and unattended locations, which is used as in normalization. The resulting firing rates are defined as:

$$R(x,\theta) = \left| \frac{A(x,\theta)E(x,\theta)}{S(x,\theta)+\sigma} \right|_T$$

where

$$S(x,\theta) = s(x,\theta) * A(x,\theta)E(x,\theta).$$

$E(x,\theta)$ is the stimulus drive at location x for orientation θ. $A(x,\theta)$ is the attentional field, $S(x,\theta)$ is the suppressive drive. σ is a constant that determines the neuron's contrast gain. $|...|_T$ specifies rectification with respect to threshold T. $*$ denotes convolution, $s(x,\theta)$ gives the extent of pooling. Stimulus contrast is also included in the equations and is not shown here; see (Reynolds and Heeger 2009) for further details. $A(x,q) = 1$ everywhere except for the positions and features that are to be attended, where $A(x,q) > 1$. The fact that the attentional gain is applied before normalization accounts for the wide range of behaviours exhibited by the model.

The Normalization Model of Attentional Modulation (NMAM) (Lee & Maunsell 2009) proposes that the primary effect of attention is to modulate the strength of normalization mechanisms. With two stimuli present in the RF, the neuron's firing rate is given by:

$$R_{1,2} = \left[\frac{N_1 \cdot (I_1)^u + N_2 \cdot (I_2)^u}{N_1 + N_2} \right]^{1/u}$$

where N_i is the normalization term for each stimulus, I_i is the direct input driven by each stimulus, and u is a power term that enables the modeling of different nonlinear summation regimens. Each normalization term depends on the contrast of the associated stimulus as:

$$N_{attended} = (1-s)(1-e^{-\beta\alpha c}) + s$$

where β=1 for unattended item, β>1 for attended items, α is the slope of normalization, c is stimulus contrast; s is the baseline of normalization.

The attentional modulation of firing rate and synchrony in a biophysical network of spiking neurons is the subject of the Cortical Microcircuit for Attention model (CMA) (Buia & Tiesinga 2008). Attention was modeled as a change in the driving current to the network neurons. In addition to excitatory (E) neurons, the investigation focuses on the role of interneurons, and suggests that both feedforward (FFI) and top-down (TDI) interneurons play a role. These are differentially modulated by attention: the firing rate of the FFI increases with spatial attention and decreases with feature-based attention, whereas the TDI increase their firing rate with feature-based attention and shift the

network synchrony from the beta to the gamma frequency range. The neurons are arranged in columns, each containing neurons of the three types. The synaptic connectivity was designed to ensure inter-stimulus competition and the generation of cortical rhythms. The dynamics of each type of neuron are given by:

$$I_{E1} = I_{0,E1} + A_E(\beta_E c_1 + (1-\beta_E)c_2)$$
$$I_{FFI1} = I_{0,FFI1} + A_{FFI}(\beta_{FFI} c_1 + (1-\beta_{FFI})c_2)$$
$$I_{E2} = I_{0,E2} + A_E(\beta_E c_1 + (1-\beta_E)c_2)$$
$$I_{FFI2} = I_{0,FFI2} + A_{FFI}(\beta_{FFI} c_1 + (1-\beta_{FFI})c_2)$$

where A is a scaling factor; c is the stimulus contrast; β is stimulus selectivity, $0.5 \leq \beta < 1$; I_0 is a constant offset current. Based on the model, the authors propose a canonical circuit for attention, and present a number of concrete and testable predictions.

An example of a modeling effort that tries to reconcile different approaches is the integrated microcircuit model of attentional processing (IMM) (Ardid et al. 2007). IMM is a biophysically based network model of spiking neurons composed of a reciprocally connected loop of two (sensory and working memory) networks. A wide variety of physiological phenomena induced by selective attention are shown to naturally arise in such a system. The proposed neural circuit is an instantiation of "feature-similarity gain modulation." The channel kinetics are modeled by:

$$\frac{ds}{dt} = -\frac{1}{\tau_s} s + \alpha_s x(1-s)$$
$$\frac{dx}{dt} = -\frac{1}{\tau_x} x + \sum_i \delta(t - t_i)$$

where s is the gating variable, x is a synaptic variable proportional to the neurotransmitter concentration in the synapse, t_i are the presynaptic spike times, τ_s ms is the decay time of NMDA currents, τ_x controls the rise time of NMDAR channels, and α_s controls the saturation properties of NMDAR channels at high presynaptic firing frequencies. Attention is modeled as a top-down signal originating in a working memory area, and primed by a cue at the start of the simulation.

Predictive coding (Rao & Ballard 1999) is reformulated as a form of biased competition by Spratling (2008a, 2008b) (PC-BC). Every processing stage in the proposed model consists of prediction and error detection nodes, with firing rate dynamics defined by:

$$e^{S_i} = y^{S_i - 1} - \left(\mathbf{W}^{S_i}\right)^T y^{S_i}$$
$$y^{S_i} \leftarrow (1 - \eta - \vartheta) y^{S_i} + \varsigma \mathbf{W}^{S_i} e^{S_i}$$
$$y^{S_i} \leftarrow y^{S_i} + \eta\left[\left(\mathbf{W}^{S_{i+1}}\right)^T y^{S_{i+1}} + \left(\mathbf{W}^{A_i}\right)^T y^{A_i}\right]$$

where S_i represents the current processing stage, y^{Si} and e^{Si} the vectors of predictive and error node activations, respectively, $\boldsymbol{W}$ are matrices of weight values, and η, ς, and ϑ are

constant scale values. The external attentional modulation signal is y^{Ai}, modulating activations through a set of weights $\boldsymbol{W}^{Ai}$. A nonlinear version of these equations, that obtains better fits to experimental data, is also presented.

Although all the models have good results, the source of the 'attention' parameter takes various forms, e.g.: ventral and dorsal prefrontal areas provide the external top-down bias that specifies the task (Rolls and Deco 2002), top-down signals corresponding to stimulus expectations or feedback processes generated by recurrent connectivity (Spratling & Johnson 2004), prefrontal memory circuits (Hamker 2005, Ardid et al. 2007), and unspecified in some models.

The value of the attentional signal as well as the selection of neuron(s) to which it applies are generally assumed, and its source and the selection mechanism are unspecified. As a result, none of these models can address issues related to temporal relationships between stimuli and attentional modulation, and hierarchical communication. Also, because each model exhibits good performance, there is little that would allow one to decide which is correct. By demonstrating ST performance on a larger experimental set, we hope to resolve this problem.

Selective Tuning

The Selective Tuning (ST) model of visual attention (Culhane and Tsotsos 1992, Tsotsos 1993, Tsotsos et al. 1995, Tsotsos 2011) starts from `first principles' and features a theoretical foundation of provable properties based on the theory of computational complexity (Tsotsos 1987, Tsotsos 1989, Tsotsos 1990, Tsotsos 1992). The `first principles' arise because vision is formulated as a search problem (given a specific input, what is the subset of neurons that best represent the content of the image?) and complexity theory is concerned with the cost of achieving solutions to such problems. This foundation suggests a specific biologically plausible architecture as well as its processing stages. Research on ST has been driven by the desire to create a theory with strong neurobiological predictive power as well as utility in practice. The model has been implemented and tested in several labs applying it to guide computer vision and robotics tasks.

ST is characterized by the integration of feedforward and feedback pathways into a network that is able to take high level decisions, and, through a series of response-based decision processes, identify the neurons that have participated in that decision. The ST feedback process does not rely on a spatial spotlight, 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 visual processing architecture is pyramidal in structure[1], as in other models (e.g. (Fukushima 1986, Riesenhuber & Poggio 1999)) with units within this network receiving

[1] A pyramidal representation is a layered representation characterized by successively coarser spatial representations

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

Selection relies on a hierarchy of Branch-and-Bound decision processes. Branch-and-Bound is a classic mechanism that is used in optimization problems (Lawler & Wood, 1966) and recursive pruning within the branch-and-bound strategy is especially useful for a hierarchical system, such as ours. Our decision processes are implemented as θ-WTA, a unique form of the common winner-take-all algorithm, a parallel algorithm for finding the maximum value in a set. There is no single winner; rather response values are partitioned into ordered groups where partition bins have width θ. All neurons that have responses within the first bin (ie., largest responses within θ of each other in value) are selected as the winners. In the first step of the algorithm, a θ-WTA process operates across the entire visual field at the top layer where it computes the global winner, i.e., the set of units with largest response. The θ-WTA can accept guidance to favor areas or stimulus qualities if that guidance is available but operates independently otherwise. The search process then proceeds to the lower levels by activating a hierarchy of θ-WTA processes. The global winner activates a θ-WTA that operates only over its direct inputs to select the strongest responding region within its receptive field. Next, all of the connections in the visual pyramid that do not contribute to the winner are pruned (inhibited). The top layer is not inhibited by this mechanism. However, as a result, the input to the higher-level unit changes and thus its output changes. This refinement of unit responses is an important consequence because one of the important goals of attention is to reduce or eliminate signal interference (Tsotsos 1990). By the end of this refinement process, the output of the attended units at the top layer will be the same as if the attended stimulus appeared on a blank field. This strategy of finding the winners within successively smaller receptive fields, layer by layer, in the pyramid and then pruning away irrelevant connections through inhibition is applied recursively through the pyramid. The end result is that from a globally strongest response, the cause of that largest response is localized in the sensory field at the earliest levels. The paths remaining may be considered the pass zone of the attended stimulus while the pruned paths form the inhibitory zone of an attentional beam. The θ-WTA does not violate biological connectivity or relative timing constraints. This algorithm is hinted at by Fuster (1990): "[I]f the relevance of a stimulus feature depends on its context, any influences that attention may have on cells that respond to that feature will arrive to those cells after analysis of the context that signals the relevance of the feature. The time taken by that analysis will be reflected by a relatively long latency of attention-modulated cell responses to the relevant feature."

In more neural terms, ST uses recurrent tracing of connections to achieve localization. The idea of tracing back connections in a top-down fashion was present in the NeoCognitron model of Fukushima (1986) and suggested even earlier by Milner (1974).

Within the Selective Tuning model, whose earliest descriptions are found in (Tsotsos 1991, Culhane and Tsotsos 1992 Tsotsos 1993), with accompanying details and proofs in (Tsotsos et al. 1995). It also appeared later in the Reverse Hierarchy Model of Ahissar & Hochstein (1997), Hochstein & Ahissar (2002). Only NeoCognitron and Selective Tuning provide realizations; otherwise, the two differ in all details. Fukushima's model included a maximum detector at the top layer to select the highest responding cell, and all other cells were set to their rest state. Only afferent paths to this cell are facilitated by action from efferent signals from this cell. In contrast, neural inhibition is the only action of ST, with no facilitation. The NeoCognitron competitive mechanism is lateral inhibition at the highest and intermediate levels. This lateral inhibition enhances the strongest single neurons thus assuming all spatial scales are represented explicitly, whereas ST finds regions of neurons, removing this unrealistic assumption. For ST, units losing the competition at the top are left alone and not affected at all — the non-attended visual world does not disappear as in NeoCognitron. ST's inhibition is only within afferent sets to winning units. This prediction of a space-limited suppressive surround firmly distinguishes the two approaches.

ST circuit

Several types of neurons are required for ST to function. The connectivity among four classes of neurons — interpretive, bias, gating and gating control — is presented in Figure 1 (adapted from Figure 5.6 in (Tsotsos 2011)). The figure shows a single assembly that computes a single visual quantity (feature, object, etc.) at a single tuning profile. All elements of this single assembly are at the same location. At the same location, however, there are many such competing assemblies spanning the tuning ranges of all visual qualtities.

Interpretive neurons are the classical feature-detecting neurons. They will be represented by *E*, and their activation by *e*. They receive feed-forward input from other areas that arrives in lamina 4 and provide an output to other areas from laminae 5 and 6.

Task information can be provided to the network by bias neurons. These provide top-down guidance for visual processing, whether the selection is for locations or regions in space, sub-ranges of visual features, objects, events, or whole scenes to attend to.

The gating sub-network, composed of gating and gating control neurons, is the major mechanism by which selection of attended neurons is accomplished and by which those neural activations are traced back down to their source, forming the path of the attentional beam. Their specific roles are described below.

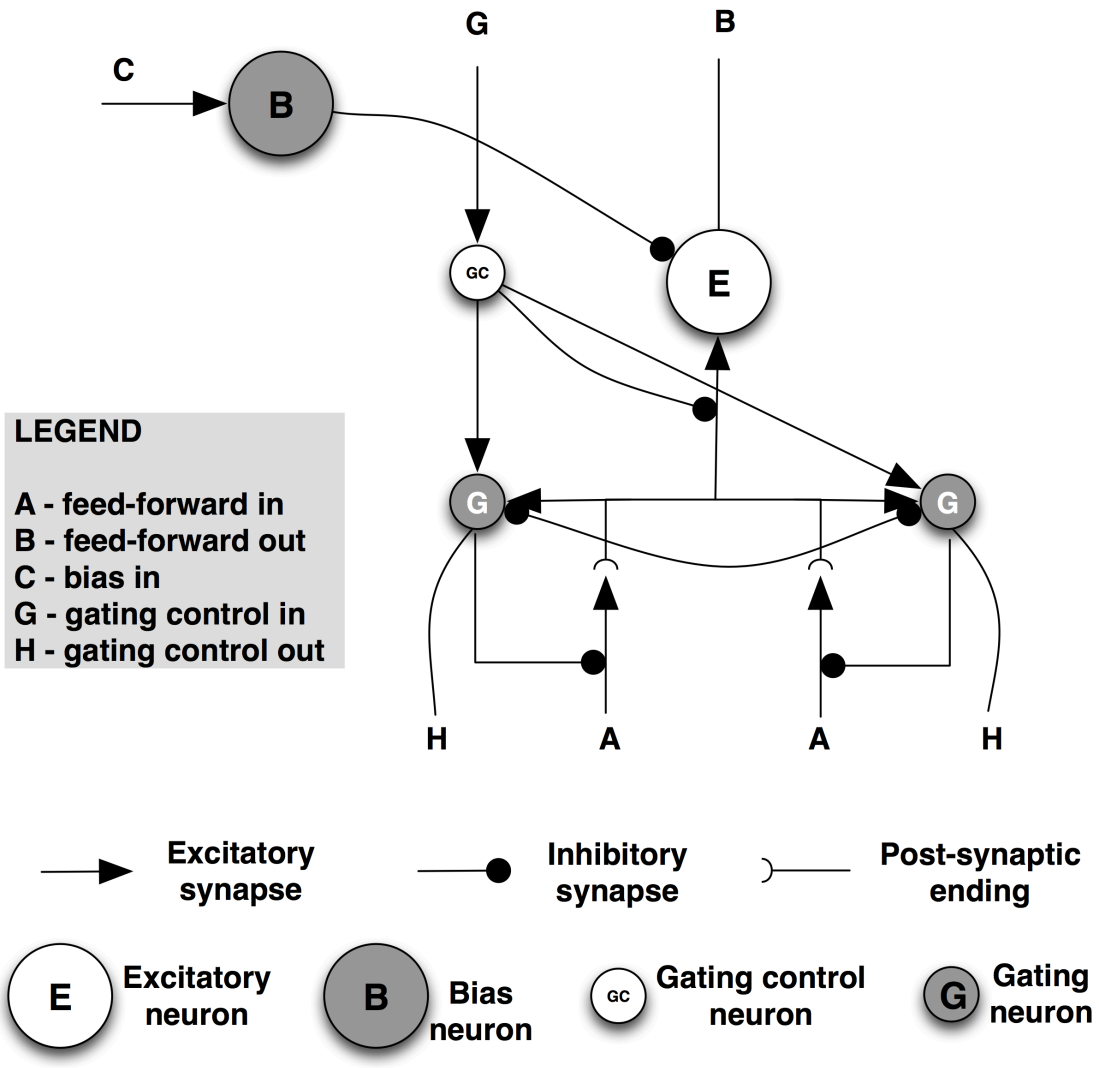

Figure 1 Selective Tuning microcircuit

We will use the following notation to represent connections between neurons:

$\overline{\wedge}$ represents the set of feed-forward connections to a neuron from all sources.

$\underline{\vee}$ represents the set of recurrent connections to a neuron from all sources.

Ξ represents the set of local connections to a neuron; that is, the neurons to which it is connected within the same visual area.

All these will be specialized by means of superscripts to distinguish the different kinds of connections. In order to keep the equations simple, we will assume that all activations and parameters correspond to a given assembly, and forgo indices that localize the neurons within the full network (but see (Tsotsos 2011) for the full formulation).

Given that our goal here is to understand how neural inputs and activations are modulated by attention, we will use a simple weighted sum of inputs formulation for the activation of neurons:

$$P(t) = \sum_{k \in \bar{\wedge}} g_k e_k(t)$$

where g_k are weights specifying the strength of contribution from neuron k to the current neuron E and e_k is the activation of neuron k.

The neuron's firing rate S is defined by:

$$S\left(P(t)\right) = \frac{Z P^{+}(t)^{\xi}}{\sigma^{\xi} + P^{+}(t)}$$

where Z is the maximum firing rate, P^{+} the positive half-rectified value of P, the exponent ξ determines the maximum slope of the function (i.e., how sharp the transition is between threshold and saturation), and σ, the semi-saturation constant, determines the point at which S reaches half of its maximum. The value σ is determined by the base semi-saturation constant σ_0 plus fast and slow after-hyperpolarizing potentials:

$$\sigma = \sigma_0 + F_{fast} H_{fast} + F_{slow} H_{slow} \ .$$

The fast (H_{fast}) and slow (H_{slow}) after-hyperpolarizing potentials are defined by:

$$\frac{dH_{fast}}{dt} = \frac{1}{\tau_{fast}}\left(-H_{fast} + e\right) \qquad \text{and} \qquad \frac{dH_{slow}}{dt} = \frac{1}{\tau_{slow}}\left(-H_{slow} + e\right)$$

respectively. The effect of these variables is to slowly decrease the value of the neuron's activation *e* when the neuron is active.

The temporal variation of a neuron's response is governed by:

$$\frac{de}{dt} = \frac{1}{\tau}\left[-e + S\left(P(t)\right)\right]$$

where τ is a decay time constant.

Bias inputs act by suppressing the input to task-irrelevant interpretive neurons. For any given neuron, the bias input is determined as the minimum value of all converging bias signals:

$$B(t) = \min_{\beta \in \underline{\vee}^{b}} B_{\beta}(t), \quad 0.0 \le B(t) \le 1.0$$

where $\underline{\vee}^{b}$ is the set of bias units making feedback connections to *E*. The default value of each bias unit is 1.0. Adding this bias to the neural response equation yields:

$$\frac{de}{dt} = \frac{1}{\tau}\left[-e + B(t) \cdot S(P(t))\right]$$

In ST, the input signals reaching a neuron are modulated by gating control signals that are a result of the winner-take-all competitions between activations. The gating sub-network is charged with determining the winning subset of the inputs to the pyramidal neuron *E*, suppressing feed-forward connections to *E* that correspond with the losers of the competition and transmitting the values of the gating neurons down to the next layer, to become the gating control signals for the next layer neurons.

The winner-take-all process creates an implicit partial ordering of the set of neural responses. The ordering arises because inhibition between units is not based on the value of a single neuron but rather on the difference between pairs of neural responses, where the difference must be at least as great as a task-specific parameter θ, $\theta \geq 0.0$. This process is not restricted to converging to single values as it is in all other formulations; rather regions of neurons are found as winners. Competition depends linearly on the difference between neuron response strengths: neuron *A* will inhibit *B* in the competition if $e_A(t) - e_B(t) > \theta$. Otherwise, e_A will not inhibit e_B. Each input to the competition must be weighted by its role for the interpretive units that it feeds to reflect the importance of each input to the interpretive computation in the competition, Thus, the inputs to the gating network must be postsynaptic as shown in Figure 1. The θ-WTA process is defined by the recurrence relation:

$$e'_f\left(t'+1\right) = e'_f\left(t'\right) - \varsigma \sum_{\lambda \in \Xi^S} \Delta\left(f,\lambda\right)$$

where *e'* and *t'* represent activation and time during the competition (i.e. competition starts at $t'=0$), and

$$\Delta\left(f,\lambda\right) = g_\lambda e'_\lambda\left(t'\right) - g_f e'_f\left(t'\right) \quad \text{if} \quad 0 < \theta < \left[g_\lambda e'_\lambda\left(t'\right) - g_f e'_f\left(t'\right)\right] \text{ and otherwise } 0.0.$$

The gating control signals ς are defined as:

$$\varsigma = \begin{cases} 1 & \sum_{a \in \underline{v}^c} \varsigma_a > 0 \\ 0 & otherwise \end{cases}$$

where $\underline{v}^c$ is the set of gating control signals converging onto *E*. There is also one gating neuron, γ_f, $f \in \overline{\wedge}$, for each of the feed-forward inputs to *E*, $0 \leq \gamma_f \leq 1.0$. This results in gating signals:

$$\gamma_f\left(t+t_{fc}\right)=\frac{e'_f\left(t_{fc}\right)}{e_f\left(t+t_{fc}\right)} \ .$$

Integrating the gating signals into the ST equation, and dropping the time parameter '(t)' for convenience, we obtain:

$$\frac{de}{dt}=\frac{1}{\tau}\left[-e+B\cdot S\left(\sum_{k\in\bar{\wedge}}\gamma_k\cdot g_k\cdot e_k\right)\right] .$$

The complete ST equation also includes lateral cooperation signals, leading to the complete equation:

$$\frac{de}{dt}=\frac{1}{\tau}\left[-e+B\cdot S\left(\sum_{k\in\bar{\wedge}}\gamma_k\cdot g_k\cdot e_k+\sum_{h\in\Xi^a} g_h\cdot e_h\right)\right]$$

where -1.0 ≤ g_h ≤ 1.0 is the weight of the connection from neuron E_h to E, and Ξ^a represent the connections horizontally across columns for neuron E. As these lateral signals are not relevant to the results presented here, the reader is referred to (Tsotsos 2011) for a full description. The performance of this model will be presented after the experimental setup is described.

Attentional modulation of neural responses

Single-neuron attentional modulation in the macaque

The basic attentional modulation effects, a necessary starting point for any model, are presented by Reynolds et al. (1999) and summarized in Figure 2. The experiment consists of the presentation of one or two stimuli within a neuron's receptive field (RF), with attention directed to the area covered by the RF or away from it. When presented alone, one of the stimuli (the reference stimulus) elicits a strong response from the neuron – black line in Figure 2, while the other (the probe stimulus) elicits a weak response – blue line. When both stimuli are shown, and in the absence of attention, the presence of the probe results in a reduction of the neuron's response relative to the response to the reference stimulus alone – green line. With attention engaged and directed towards the reference stimulus, the response recovers, being similar to the response to the reference stimulus presented alone – red line.

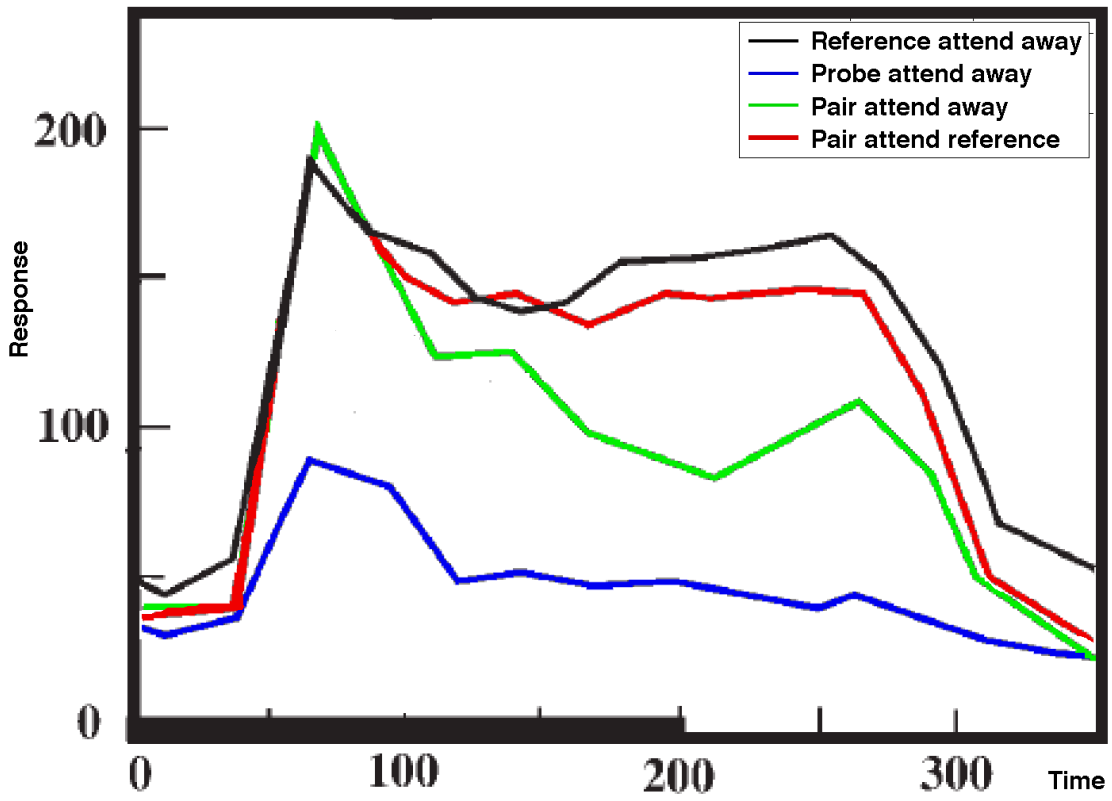

Figure 2 A summary of the Reynolds et al. experimental results. See text for details. Combined from Figure 6a and 6b in (Reynolds & Desimone 1999).

Single-neuron attentional modulation in ST

To investigate the modulation of neural activations due to ST attentional selection we used the simple circuit illustrated in Figure 3. The responses of the model neurons are defined by the ST equations presented above, the same equations are used for all 4 excitatory neurons, while the inhibitory interneurons are described by the same equations, omitting the gating components. Simulating the (Reynolds & Desimone 1999) experiments, the two neurons at the bottom of the figure correspond to the reference (left, labeled $E_{1,1}$) and probe (right, labeled $E_{2,1}$) stimuli. The two neurons at the top of the figure represent neurons that have the reference (left, labeled $E_{1,2}$) and the probe (right, labeled $E_{2,2}$) as preferred stimulus. Their inputs and outputs correspond to the connections labeled A and B, respectively, in Figure 1. The excitatory input to the output

units, represented by arrows is the sum of the activations of the input units multiplied by their respective weights (larger for the preferred stimulus, smaller for the non-preferred). Similarly, the inhibitory input is the weighted sum of the activations of two inhibitory units, one corresponding to each input. The θ-WTA competition between the two output units is represented as mutual inhibition. The model network also includes ST bias, gating and gating control units (with the associated connections, labeled C, D, G and H in Figure 1), for simplicity these are not represented in Figure 3. Bias units are only used in one experiment, as indicated.

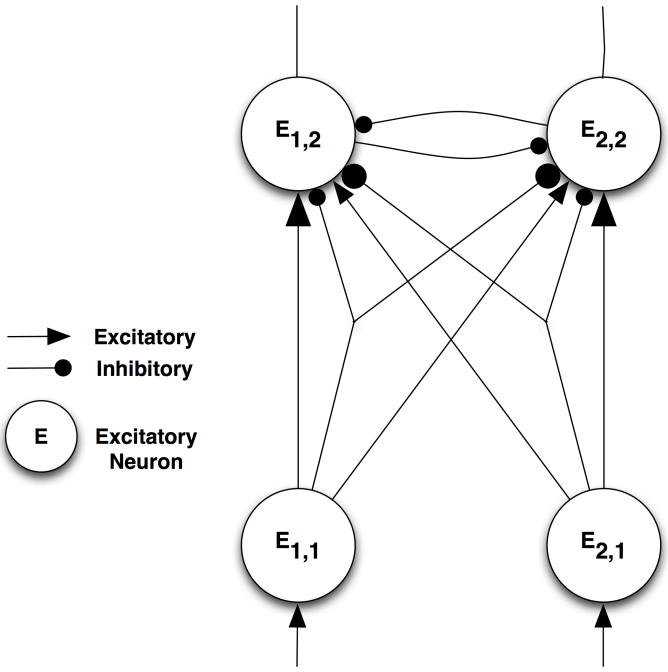

Figure 3 The network structure. Input neurons are at the bottom. Similar to the Reynolds & Desimone (1999) model, we include both excitatory and inhibitory inputs in all combinations. Excitatory and inhibitory connections are represented by arrows and circles, respectively. Connection size correlates with connection weight, i.e. $E_{1,2}$ receives large inputs from $E_{1,1}$ (excitatory) and $E_{2,1}$ (inhibitory), and small inputs from $E_{1,1}$ (inhibitory) and $E_{2,1}$ (excitatory).

The model network is tested in the same four conditions as the Reynolds et al. experiment, and the results are presented using the same color-coding as that used in Figure 2. Figure 4 represents the output of neuron $E_{1,2}$ in these four experimental conditions. The "Pair attend reference" (red) line represents the condition when the reference stimulus is attended (i.e. neuron $E_{1,2}$ wins the top-level θ-WTA). The attentional selection process is triggered 100ms after the presentation of the stimulus, indicated by a vertical line. It can be observed that the response for the unattended pair of stimuli is lower than the response for the reference stimulus alone, and that attending to the reference stimulus in the pair enhances the neuron's response. The relative responses in

the different conditions can be changed by manipulating the weights of the different connections, as illustrated in Figure 11 in the Appendix.

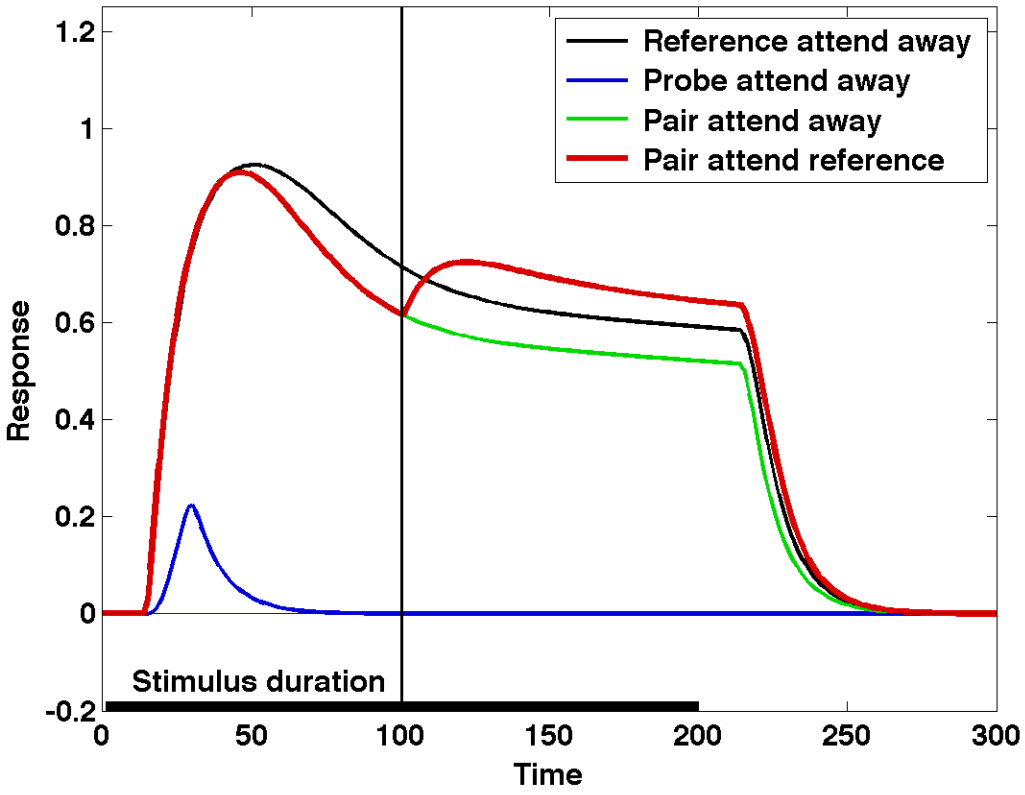

Figure 4 The output of neuron E1,2 in the four experimental conditions. The "Pair attend reference" (red) line represents the condition when the reference stimulus is attended (i.e. neuron E1,2 wins the top-level θ-WTA)

Certain characteristics of the response, such as the amount of attentional modulation and the timing of the effect, can be manipulated to provide further insight into the various modes of operation possible in ST.

Experiments show that the amount of attentional modulation depends on a number of factors, including area studied and target-distractor similarity (see (Walther & Koch 2006) for a summary). **Figure 4** has been obtained by restricting the gating signal γ_k to the values of 0 (unattended) and 1 (attended), resulting in maximum attentional modulation, but by reducing the range of the gating signal, the ST equations can provide control over the modulation. For example, in Figure 5, the amount of inhibition for the unattended stimulus is only 0.5, resulting in an attended response that more closely matches the reference alone condition.

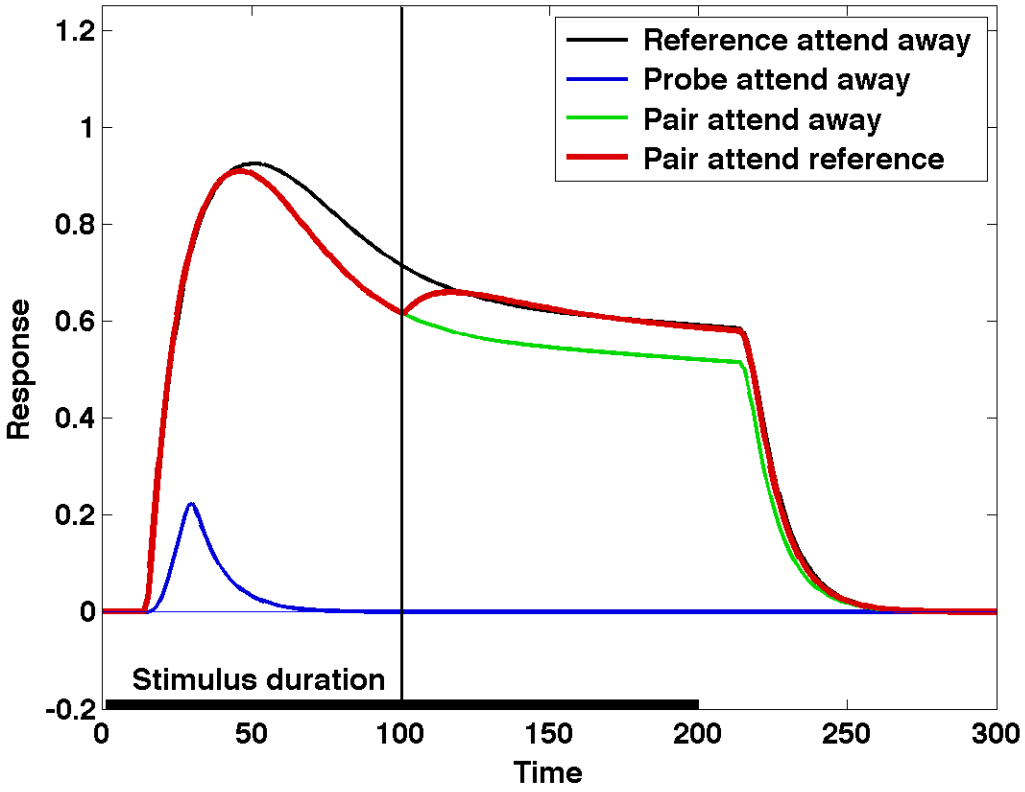

Figure 5 Gating is reduced to 50%

The control of this variable gating effect is not included in the equations, but since gating is controlled by the θ-WTA, and the θ-WTA depends of target-distractor similarity, it is not implausible to hypothesize that the θ-WTA process also controls the magnitude of the gating, but the mechanism is unknown.

One significant difference between the simulation results presented above and the Reynolds et al. experiment is that the location of the target has been cued in the experiments, but not in the model. Including a spatial bias towards the reference stimulus in the ST model, which is the equivalent of spatial cueing prior to stimulus presentation, shows another mode of operation made possible by the ST equations. The response of the model in the four conditions with pre-cueing of the reference stimulus is shown in Figure 6. The effect of cueing in ST is described in detail and experimentally investigated by Cutzu & Tsotsos (2003).

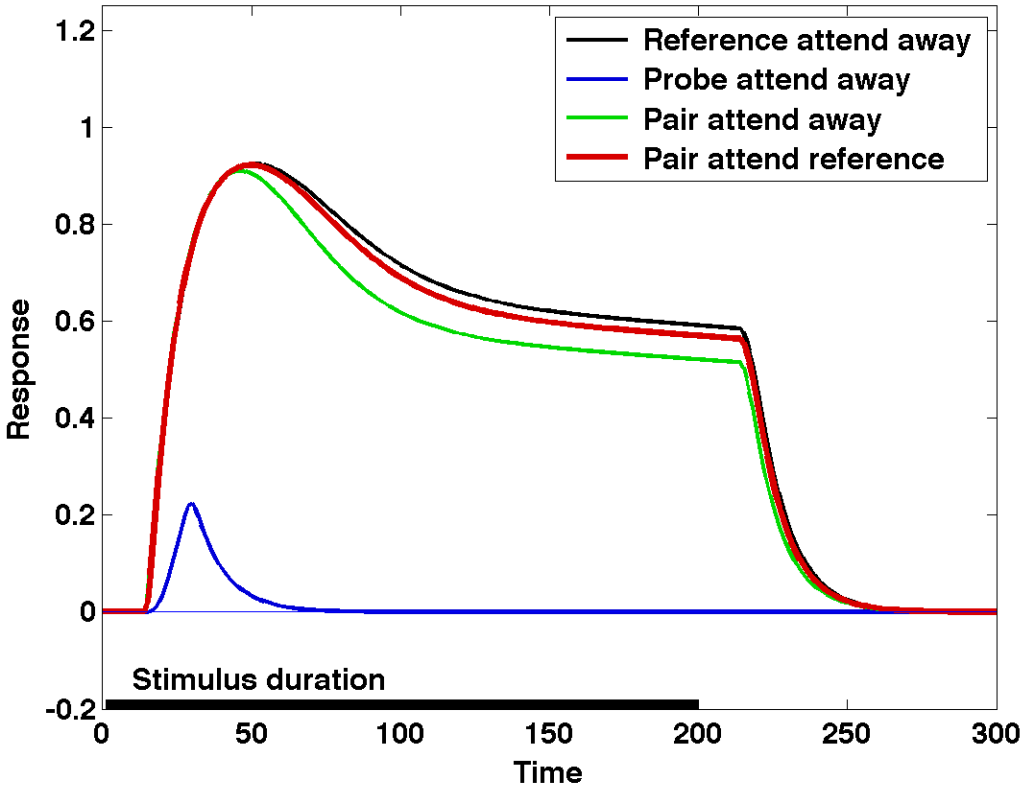

Figure 6 The effect of spatial cueing

An interesting characteristic of the ST response in **Figure 4** and Figure 5 is the strong rebound of the neural response to the pair of stimuli when the selection process is triggered. The rebound is (at least in part) due to the fact that in the model gating is applied instantaneously, resulting in a transient in the input signal to the neuron. This transient produces a strong change in the response, similar to the one corresponding to the stimulus onset. It is possible to control the transient, and thus the rebound, by applying gating gradually, which is realistic, as it is the result of the activation of the gating neurons.

The delay in triggering the attentional selection also determines the location and amplitude of the attentional modulation. Early distractor gating produces the rebound while the neuron is closer to saturation, resulting in a relatively lower effect, as seen in Figure 7.

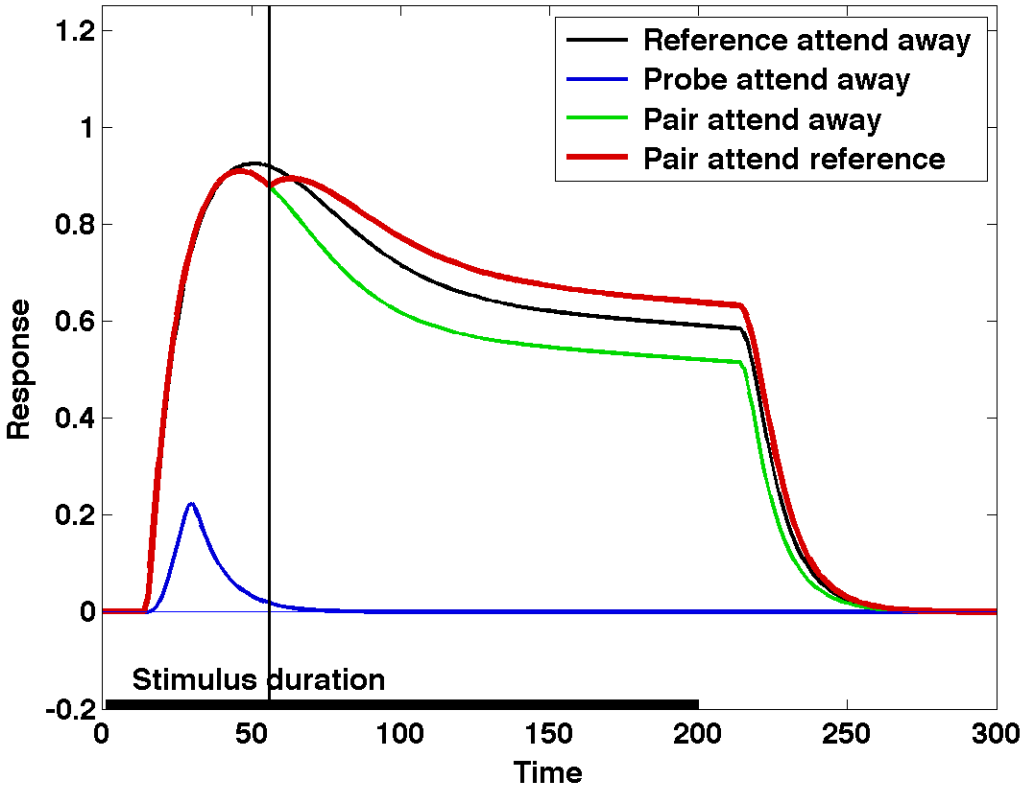

Figure 7 Earlier attentional modulation. Same parameters as in Figure 4, but the θ-WTA takes place sooner.

Before dismissing the rebound as a timing-dependent modeling artifact, we need to investigate what could account for different delays in the attentional modulation. Assuming that the rebound might correspond to a real phenomenon, can we predict the kinds of experiments that will show a similar effect? The Reynolds et al. experiment does not show any obvious equivalent, possibly due to the spatial cueing, as discussed above. The network discussed so far contains a single layer of neurons. In more realistic multi-layer networks, the results of top-level competitions take time to propagate back through the network, resulting in the attentional modulation being applied at different times in different layers. This means that under the right conditions, not only will attentional modulation show the rebound, but since at each different layer the corresponding rebound will be produced at a different time, the modulation at higher levels of the network will show a pattern of attentional modulation consistent with the accumulation of lower level rebounds.

Latency of attentional modulation in a processing hierarchy

One of the early predictions of ST is that of a temporal ordering and thus a time course of hierarchical modulation effects. Specifically, modulation will be seen at the highest levels first and at the lowest levels last, the opposite of what other models would suggest. Selection of the strongest response at the top of a hierarchical network triggers a recurrent localization process. At each successive layer of recurrence, part of the input to the selected neurons is suppressed, leading to a change in response for all neurons upstream. The changes in response thus occur in time steps defined by the number of layers in the pathway downward, plus the time it would take for the effect of suppression at one layer to ripple up to the neuron being examined. Significant experimental evidence for this prediction has been presented - e.g. (Mehta et al. 2000, O'Connor et al. 2002, Roelfsema et al. 2007, Lauritzen et al. 2009, Boehler et al. 2009, Buffalo et al. 2010)

Reference experiment

One of the earliest experiments to study the timing of selective attention modulation across areas of the macaque visual system by Mehta et al. (2000), was performed by simultaneous recordings from different areas, thus allowing direct comparison of the magnitude and timing of the responses and modulation. Recordings of laminar event-related potential and current source density response profiles were sampled with linear array multicontact electrodes. The subjects were required to perform alternative discrimination tasks on auditory and visual stimuli, while ignoring the stimuli in the other modality. The visual stimuli were diffuse light flashes differing in intensity or color presented at the fixation point, and the effect of attention was evaluated by comparing responses to the visual stimuli when attended vs. when ignored.

Responses were summed over all contacts at each time point to obtain a sum average rectified current flow (sAVREC), while the difference between the ignored and attended conditions was summed to obtain the difference average rectified current flow (dAVREC). The temporal evolution of the responses and of the attentional modulation was determined by comparing sAVREC and dAVREC in each of the investigated areas. Figure 9(a) shows in sAVREC in black, and dAVREC in gray. Of interest in this context is the finding that the earliest attentional modulation (i.e. the earliest significant dAVREC) was observed in the highest areas, and progressively later towards the lower areas.

Latency of attentional modulation in ST

To investigate the timing of the effects in a ST hierarchy, the circuit described in Figure 3 was replicated to form four layers, as shown in Figure 8, with the output of one processing layer driving the input of the next. To illustrate the top-down nature of the ST process, gating control units are shown on the left side of the hierarchy. The circuit is symmetrical, and gating control units exist for each connection, but are omitted for clarity (same for the inhibitory interneurons). The neurons are characterized by the same

differential equations introduced above. Stimuli are presented at the bottom and the activations propagate through the network – the black curves in Figure 9. The top-level θ-WTA process determines a winner, and the corresponding gating signals are propagated down the network, triggering local θ-WTA processes within each winning neuron's afferents. This results in the modulation of the neural responses, as described above.

Figure 9 compares the model spiking rates on the right with average transmembrane currents recorded from neurons in different visual areas in the attention experiment described in the previous section on the left, for the neurons corresponding to the attended stimulus. The figures compare the relative timing of the initial response (in black) and the attentional modulation (in gray) across visual areas. The attentional modulation for the experiment is dAVREC, and similarly, for the model it is difference between the responses of the interpretive units in the ignored and attended conditions.

A detailed representation of the relevant 130-200 ms time interval is presented in Figure 10. The neural activation (in black) shows the responses being generated progressively later in more superior areas, while the attentional modulation (grey) appears earlier in superior areas and later in early areas. In the model, the propagation time between visual areas has been set to 15 ms, for both the feedforward and the feedback stage. Mehta et al. do not provide a quantitative evaluation of the delays, but this could easily be integrated into the model. Note that the modeling is qualitative, meant to show only the general timings and shape of the response and modulation, as the details of the real network are unknown, however, the similarity to (Mehta et al. 2000) is striking, and a key characteristic of attention not found in other models.

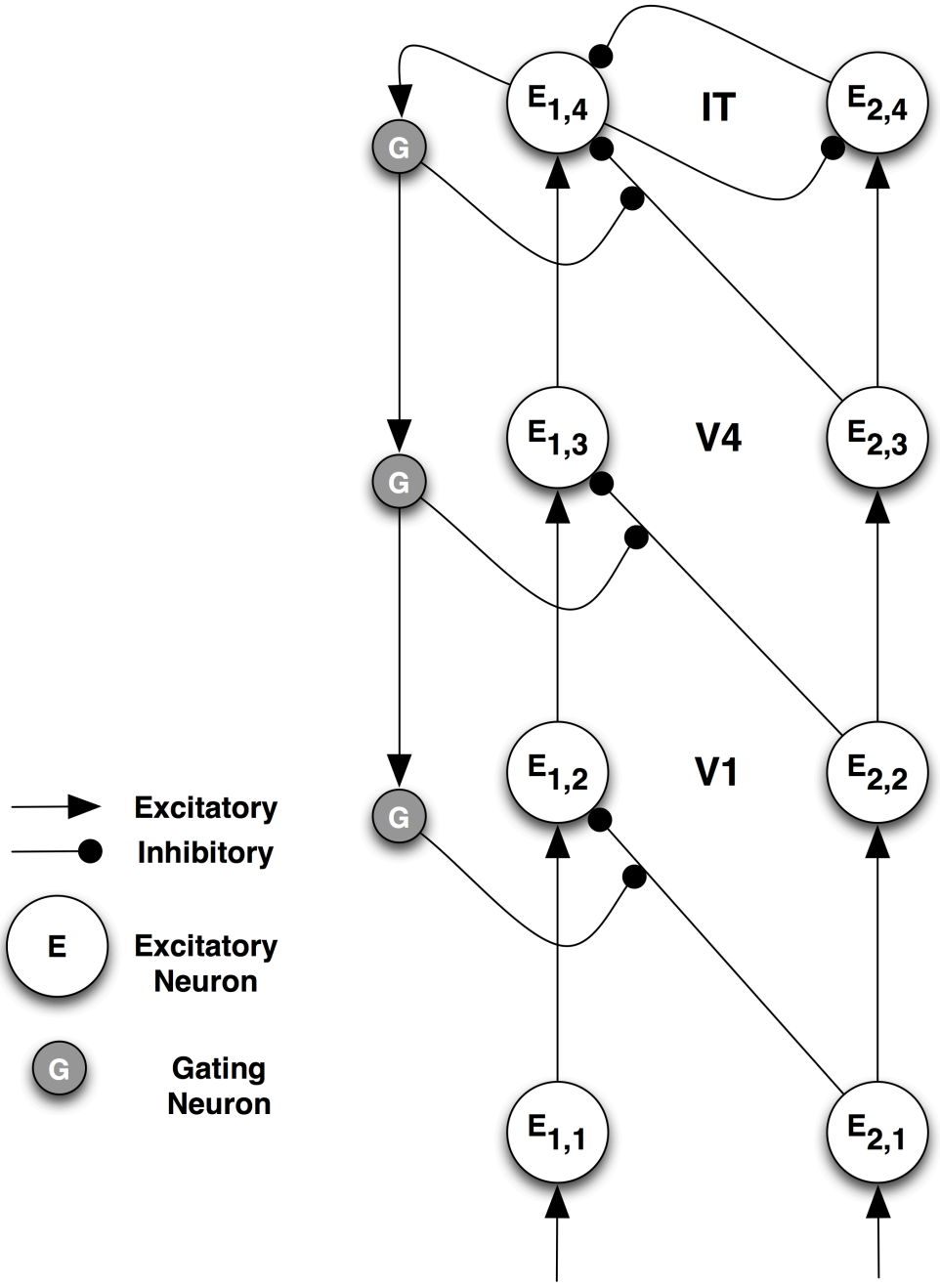

Figure 8 Full network for detailed timing analysis. The θ-WTA competition is indicated by the mutually inhibitory connections at the top level of the network.

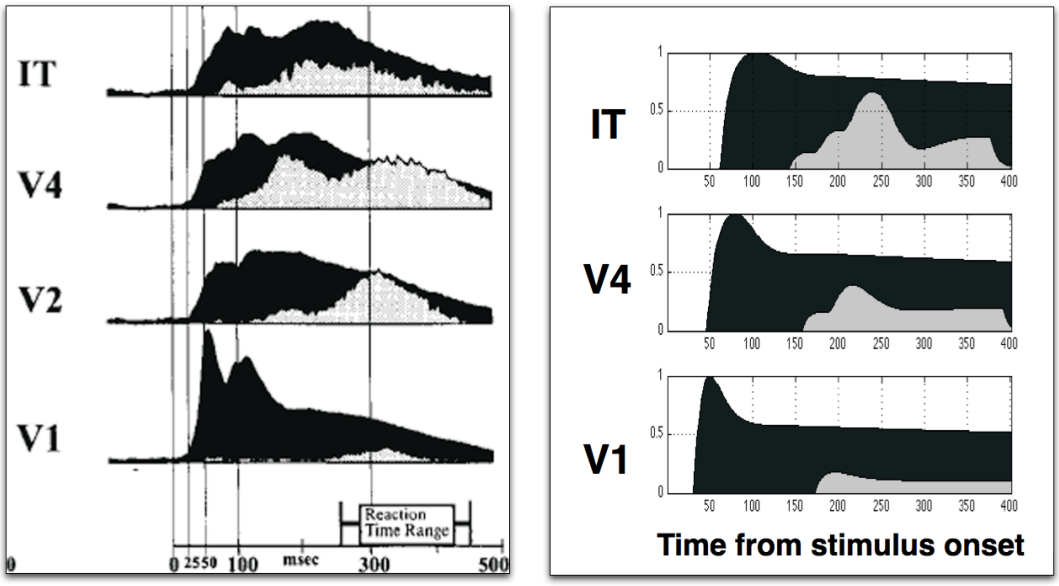

Figure 9 Attentional modulation of responses. Black indicates the neural response to the stimulus, while the attentional modulation is represented in grey. (a) Temporal pattern of activations and attentional modulation in single-unit recordings in primates performing attentional tasks. Adapted from Fig. 9b in (Mehta et al. 2000). The neural activation (in black) shows the responses being generated progressively later in more superior areas, while the attentional modulation (grey) appears earlier in superior areas and later in early areas. (b) Model results showing a similar activation and modulation temporal pattern.

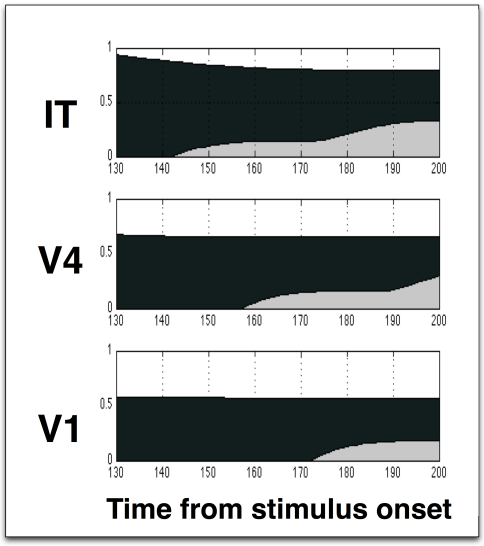

Figure 10 Model results - detail on the 130-200 ms interval, showing the temporal pattern of attentional modulation.

Discussion

We have shown through computational modeling that ST produces qualitatively equivalent modulatory effects for single neurons, similar to other models, but in addition qualitatively correct results for a hierarchy of neurons in contrast to other models, without the need for external attentional or bias inputs. Other models encode attention by modifying contrast, bias or gain parameters whose value changes from attended to unattended values. These models are all silent on how this value is set or how selections may occur whereas ST has an integrated selection mechanism. All except for ST are data-fitting models and unable to accept image input and produce the required behavior as ST can. The most immediate first impression of comparison is how different all the formulations appear. ST, BCM, RH and PC-BC are based on the firing rate neuron formulation; FSG, NMA, NMAM are divisive contrast normalization models; CMA and IMM are spiking neuron models; ND employs mean-field approximation. BCM, FSG, NMA, and NMAM are single-neuron models. NMA goes beyond a single neuron in that it takes larger visual fields into account. CMA, ND, RH, PC-BC and ST employ networks of several types neurons. ND, RH, PC-BC and ST operate over complex network architectures. But this seems to be more of a feature of the model scope and starting assumptions than of substance.

As a first point of comparison, all of the models except for ST are data-fitting models. Each would take existing data and determine parameter values of a set of equations that provide the closest fit to the data. As such, equations with high degrees of freedom (most variables) and nonlinearities have the greatest potential to capture the data presented. They also are the least specific or have the least scientific value because a high-enough number of variables and nonlinearities may capture just about any data set. ST takes input images and determines responses to that input, a completely different approach because the data and/or behavior must be produced for specific input. Again, a computer program may behave in any manner its programmer sees fit; it too may have suspect scientific value unless it has been developed on a sound and principled theoretical foundation. The development of ST has been conducted on such a sound theoretical foundation, and all aspects of its realization have been guided by it (Tsotsos 2011).

The next dimension along which these models may be compared is the manner in which attention is incorporated. CMA encodes attention by modifying a linear contrast parameter in the equations for the E and FFI neurons. Similarly, BCM, FSG, NMA, and NMAM all provide a single parameter that controls attention; this is a bias or gain whose value changes from ‘attended’ to ‘unattended’ values. For example, in BCM, attention is implemented by increasing by a factor of 5 both excitatory and inhibitory synaptic weights projecting from the input neuron population responding to the attended stimulus. In NMAM there is a parameter that takes values equal to 1 for unattended stimuli and larger for attended ones. Its effect is multiplicative; it multiplies the product of slope of normalization and contrast in the exponent of the response function. These models are all silent on how this value is set or how selections may occur. CMA is also silent in this regard. In IMM the self-sustained activity of an additive gating signal is triggered by the presentation of the stimulus to be attended during a cueing interval. This allowed the investigation of the effect of attention on the baseline activity of neurons. In ND,

processing can be controlled by external task signals that select either neuron pools associated with an object to be searched, or a location, in order to identify the object at that location. In RH, object recognition neurons with RFs covering the visual area of interest have their sensitivity and gain increased by reentrant signals from movement areas, thus having an advantage in the competition process. PC-BC relies on feedback signals that can originate from node activations calculated at higher-levels in the hierarchy and/or external inputs, and attentional feedback is treated exactly the same as feedback from higher stages in the hierarchy.

Different forms of competition have been employed by different authors. Most models implement competition through the mutual inhibition between neural outputs, e.g. BCM, ND, RH. In a few cases the competition involves the inputs, in the form of output neurons suppressing the input of other neurons, e.g. PC-BC, or a direct competition between the inputs in ST.

Conclusion

In this paper we have shown that ST generates patterns of attentional modulation and its temporal progression. The selection mechanism employed is completely integrated within the basic equations, without the need for external attentional signals. The selection can be aided by feature and location task biases, but these are not necessary. Further, it is consistent with the conclusions of Khayat et al. (2010) who concluded that attention can be best considered as modulating inputs to neurons, both spatial and feature: ST's equations do exactly this, manipulating neural inputs to achieve the required attentive effect. In a very real sense, ST subsumes other models; however, this is at a qualitative level of description only. Most of the other models described are quantitative, that is, they can be quantitatively compared to actual neural recordings in terms of time and firing rates, whereas ST cannot. On the other hand, they cannot explain the top down latency of attentional modulation just like they cannot explain how attentional focus is determined. The transition of ST into a quantitative model is not an intellectual challenge; parameters derived from real data can be easily obtained and ST's basic equations modified appropriately. For the other models, easy transitions to deal with determination of focus and timing are not possible without whole-scale changes to the model. Further, ST makes unique predictions: 1) that circuits responsible for computing the θ-WTA function should exist; 2) that feedback connections should be considered to not have uniform function but may entail separable functionalities as in the model (bias signals, gating control signals at least). Although we have shown strong performance of the model, we are by no means convinced that this is the final word on the topic. Far from it, the need to tighter relationship between theory and experiment are critical as only close collaborations between theory and experiment will reveal the ways in which to further develop our models.

Appendix

The following values are used for the various parameters in the ST equations.

The exponent $\xi = 3.0$ represents a typical mean exponent for visual neurons (Sclar et al. 1990).

The maximum firing rate is normalized to 1 (Z=1).

The base semi-saturation constant $\sigma_0 = 0.8$. The semi-saturation constant is modulated by fast and slow afterhyperpolarization currents, with $F_{fast} = 1.3$, $F_{slow} = 2$.

Time constant values (τ = 10ms, τ_{slow} = 900ms , τ_{fast} = 50ms) are consistent with physiological data (McCormick & Williamson 1989, Sanchez-Vives et al. 2000).

As task bias is not used in these experiments, so B =1, except for the spatial cueing experiment presented in Figure 7, where a value of B =0.5 was used.

While all the constants used in this model are in the general range found in the visual cortex, they are not crucial to the results presented here. The same equations are used for all excitatory neurons.

The top level θ-WTA starts 100 ms after stimulus presentation, and θ is 20% of the maximum firing rate Z, i.e. 0.2.

The parameters that determine the attentional modulation in the first experiment are the four weights that feed into each top-level neuron, two excitatory and two inhibitory. The weights, given here with the default values used to produce Figures 4-7 are:

- $pref_+ = 1$ - excitatory input for preferred stimulus. The lower the preferred excitation, the more effect the inhibitory inputs have, so the modulation is stronger.
- $pref_- = -0.1$ - inhibitory input for preferred stimulus. The stronger the preferred inhibition, the lower the initial response and the higher the rebound after ST.
- $nonpref_+ = 0.2$ - excitatory input for non-preferred stimulus. The main effect is on the poor, attend away condition (green line). Higher excitation offsets distractor inhibition, so it reduces the effect of ST.
- $nonpref_- = -0.35$ - inhibitory input for non-preferred stimulus. The higher the inhibition, the stronger the modulatory effect of ST.

The circuit being symmetrical, the same four weight values are used for both top-level neurons. The effect of changing each weight independently is illustrated in Figure 11.

- Left: $pref_+$ (1.0…0.4) - excitatory input for preferred stimulus. The lower the preferred excitation, the more effect the inhibitory inputs have, so the modulation is stronger.
- Center-left: $pref_-$ (0.1…-0.7) - inhibitory input for preferred stimulus. The stronger the preferred inhibition, the lower the initial response and the higher the rebound after ST.

- Center-right: $nonpref_{+}$ (0.0...0.5) - excitatory input for non-preferred stimulus. The main effect is on the poor, attend away condition (green line). Higher excitation offsets distractor inhibition, so it reduces the effect of ST.
- Right: $nonpref_{-}$ (0.0...-0.6) - the higher the inhibition, the stronger the modulatory effect of ST.

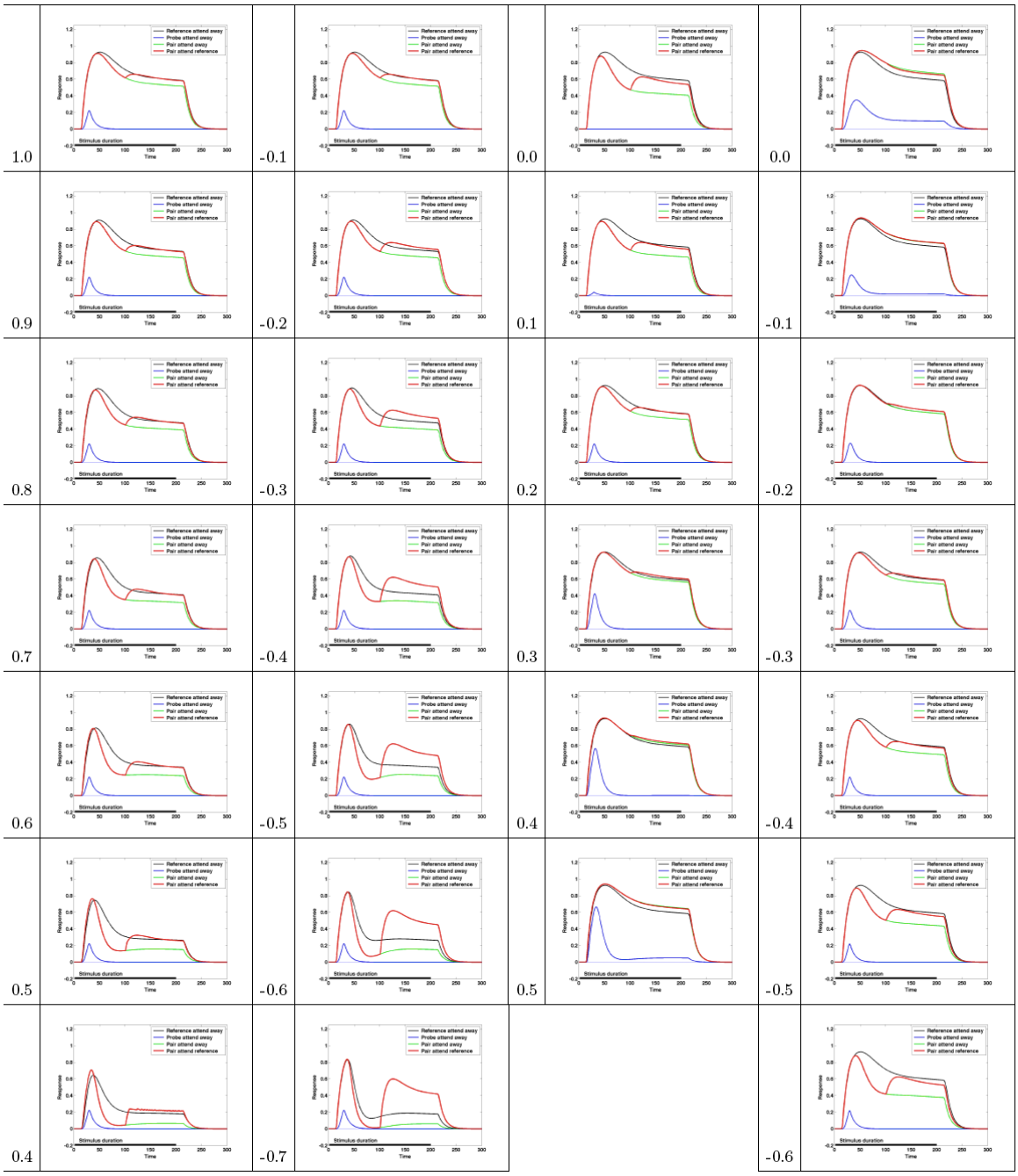

Figure 11 The effect of changing the weight values. Each column represents the modulation when one weight is changed, while the others are fixed at the default value. See text for details.

Acknowledgments

Funding gratefully received from the Canada Research Chairs Program and the Natural Sciences and Engineering Research Council of Canada.

References

Ahissar, M. & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature,* **387**(6631), 401–406.

Ardid, S., Wang, X.-J. & Compte, A. (2007). An integrated microcircuit model of attentional processing in the neocortex. *Journal of Neuroscience,* **27**(32), 8486–8495.

Boehler, C.N., Tsotsos, J.K., Schoenfeld, M., Heinze, H.-J., Hopf, J.-M. (2009). The center-surround profile of the focus of attention arises from recurrent processing in visual cortex, *Cerebral Cortex,* 19:982-991.

Boynton, G. (2005). Attention and visual perception. *Current Opinion in Neurobiology,* **15**(4), 465–469.

Broadbent, D. E. (1958). *Perception and communication.* New York, Pergamon Press.

Buffalo, E. A., Fries, P., Landman, R., Liang, H. & Desimone, R. (2010). A backward progression of attentional effects in the ventral stream. *Proceedings of the National Academy of Sciences,* **107**(1), 361–365.

Buia, C. I. & Tiesinga, P. H. (2008). Role of interneuron diversity in the cortical microcircuit for attention. *Journal of Neurophysiology,* **99**(5), 2158–2182.

Busse, L., Wade, A. R. & Carandini, M. (2009). Representation of concurrent stimuli by population activity in visual cortex. *Neuron,* **64**(6), 931–942.

Carandini, M. & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience,* **13**, 51–62.

Culhane, S. and Tsotsos, J. (1992). An Attentional Prototype for Early Vision, Proceedings of the Second European Conference on Computer Vision, Santa Margherita Ligure, Italy, G. Sandini (Ed.), LNCS-Series Vol. 588, Springer Verlag, pages 551-560.

Cutzu, F. & Tsotsos, J. K. (2003). The selective tuning model of attention: psychophysical evidence for a suppressive annulus around an attended item. *Vision Research,* **43**(2), 205–219.

Desimone, R. & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience,* **18**(1), 193–222.

Duncan, J. (1980). The locus of interference in the perception of simultaneous stimuli. *Psychological Review,* **87**(3), 272–300.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General,* **113**(4), 501–517.

Fukushima, K. (1986). A neural network model for selective attention in visual pattern recognition. *Biological Cybernetics,* **55**(1), 5–16.

Fuster, J. M. (1990). Inferotemporal units in selective visual attention and short-term memory. *Journal of Neurophysiology,* **64**(3), 681–697.

Hamker, F. H. (2005). The reentry hypothesis: the putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement. *Cerebral Cortex,* **15**(4), 431–447.

Hochstein, S. & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron,* **36**(5), 791–804.

Itti, L. & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience,* **2**(3), 194–203.

Itti, L., Rees, G. & Tsotsos, J. K. (eds.) (2005). *Neurobiology of Attention.* Elsevier Press.

Khayat, P. S., Niebergall, R. & Martinez-Trujillo, J. C. (2010). Attention differentially modulates similar neuronal responses evoked by varying contrast and direction stimuli in area MT. *Journal of Neuroscience,* **30**(6), 2188–2197.

Lauritzen, T. Z., D'Esposito, M., Heeger, D. J. & Silver, M. A. (2009). Top–down flow of visual spatial attention signals from parietal to occipital cortex. *Journal of Vision,* **9**(13).

Lawler, E. L., & Wood, D. E. (1966). Branch-and-bound methods: A survey. *Operations research*, *14*(4), 699-719.

Lee, J. & Maunsell, J. H. R. (2009). A normalization model of attentional modulation of single unit responses. *PLoS ONE,* **4**(2), e4651.

Martinez-Trujillo, J. C. & Treue, S. (2004). Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current Biology,* **14**(9), 744–51.

McCormick, D. A. & Williamson, A. (1989). Convergence and divergence of neurotransmitter action in human cerebral cortex. *Proceedings of the National Academy of Sciences of the United States of America,* **86**(20), 8098–8102.

Mehta, A. D., Ulbert, I. & Schroeder, C. E. (2000). Intermodal selective attention in monkeys. I: distribution and timing of effects across visual areas. *Cerebral Cortex,* **10**(4), 343–358.

Milner, P. M. (1974). A model for visual shape recognition. *Psychological Review,* **81**(6), 521–535.

Neisser, U. (1967). *Cognitive psychology.* New York, Appleton-Century-Crofts.

O'Connor, D., Fukui, M., Pinsk, M. & Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience,* **5**(11), 1203–1209.

Pashler, H. E. (ed.) (1998a). *Attention.* Philadelphia: Taylor & Francis.

Pashler, H. E. (1998b). *The psychology of attention.* Cambridge, Mass.: MIT Press.

Rao, R. P. N. & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience,* **2**(1), 79–87.

Reynolds, J. and Desimone, R. (1999). The role of neural mechanisms of attention in solving the binding problem. *Neuron,* **24**(1), 19–29.

Reynolds, J. & Heeger, D. (2009). The normalization model of attention. *Neuron,* **61**(2), 168–185.

Reynolds, J. H., Chelazzi, L. & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience,* **19**(5), 1736–1753.

Riesenhuber, M. & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience,* **2**(11), 1019–1025.

Roelfsema, P. R., Tolboom, M. & Khayat, P. S. (2007). Different processing phases for features, figures, and selective attention in the primary visual cortex. *Neuron,* **56**(5), 785–792.

Rolls, E. T. & Deco, G. (2002). *Computational neuroscience of vision.* Oxford University Press.

Rothenstein, A. L. & Tsotsos, J. K. (2008). Attention links sensing to recognition. *Image and Vision Computing,* **26**(1), 114–126.

Sanchez-Vives, M. V., Nowak, L. G. & McCormick, D. A. (2000). Cellular mechanisms of long-lasting adaptation in visual cortical neurons in vitro. *Journal of Neuroscience,* **20**(11), 4286–4299.

Schneider, W. & Shiffrin, R. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review,* **84**(1), 1–66

Sclar, G., Maunsell, J. H. R. & Lennie, P. (1990). Coding of image contrast in central visual pathways of the macaque monkey. *Vision Research,* **30**(1), 1–10.

Spratling, M. & Johnson, M. (2004). A feedback model of visual attention. *Journal of Cognitive Neuroscience,* **16**(2), 219–237.

Spratling, M. W. (2008a). Predictive coding as a model of biased competition in visual attention. *Vision Research,* **48**(12), 1391–1408.

Spratling, M. W. (2008b). Reconciling predictive coding and biased competition models of cortical function. *Frontiers in Computational Neuroscience,* **2**(4), 1–8.

Treisman, A. M. (1969). Strategies and models of selective attention. *Psychological Review,* **76**(3), 282–299.

Treisman, A. M. & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology,* **14**(1), 107–41.

Treue, S. and Martinez-Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature,* **399**(6736), 575–579.

Tsotsos, J. K. (1987). A 'complexity level' analysis of vision. In International Conference on Computer Vision: Human and Machine Vision Workshop. London, England.

Tsotsos, J. K. (1989). The complexity of perceptual search tasks. In International Joint Conference on Artificial Intelligence, pp. 1571–1577. Detroit.

Tsotsos, J. K. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Sciences,* **13**(3), 423–445.

Tsotsos, J. K. (1991). Tsotsos, J.K., Localizing Stimuli in a Sensory Field Using an Inhibitory Attentional Beam, October 1991, RBCV-TR-91-37, Dept. of Computer Science, University of Toronto.

Tsotsos, J. K. (1992). On the relative complexity of active vs. passive visual-search. *International Journal of Computer Vision,* **7**(2), 127–141.

Tsotsos, J. K. (1993). An inhibitory beam for attentional selection. In L. Harris and M. Jenkin (eds.), *Spatial Vision in Humans and Robots,* pp. 313 – 331. Cambridge University Press.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y. H., Davis, N. & Nuflo, F. (1995). Modeling visual-attention via selective tuning. *Artificial Intelligence,* **78**(1-2), 507–545.

Tsotsos, J. K. (2011). *A Computational Perspective on Visual Attention.* MIT Press

Tsotsos, J. K. & Rothenstein, A. L. (2011). Computational models of visual attention. *Scholarpedia,* **6**(1), 6201.

Walther, D. & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks,* **19**(9), 1395–1407.