Towards a neural network model of the visual short-term memory

Anders Petersen[1] (ap@imm.dtu.dk)
Søren Kyllingsbæk[1,2] (sk@imm.dtu.dk)
Lars Kai Hansen[1] (lkh@imm.dtu.dk)

[1]Department of Informatics, Technical University of Denmark,
Richard Petersens Plads, B. 321,
2800 Kgs. Lyngby, Denmark

[2]Department of Psychology, Copenhagen University
Linnésgade 22,
1361 Copenhagen, K, Denmark

Abstract

In this paper a neural network model of visual short-term memory (VSTM) is presented. The model aims at integrating a winners-take-all type of neural network (Usher & Cohen, 1999) with Bundesen's (1990) well-established mathematical theory of visual attention. We evaluate the model's ability to fit experimental data from a classical whole and partial report study. Previous statistic models have successfully assessed the spatial distribution of visual attention; our neural network meets this standard and offers a neural interpretation of how objects are consolidated in VSTM at the same time. We hope that in the future, the model will be developed to fit temporally dependent phenomena like the attentional blink effect, lag-1 sparing, and attentional dwell-time.

Keywords: visual attention, visual short-term memory, the magical number 4, winners-take-all network

Introduction

For everyday life, it is important for us to be able to perceive, comprehend, and react to events in our environment. Often, our rate of success is heavily dependent upon how efficient and how fast we can process, interpret and react to sensory stimuli, e.g. like when we are driving a car.

In the following we shall refer to *visual attention* as the process that enables us to focus our processing resources to certain important objects in the visual scene. Following the theory of visual attention (TVA, Bundesen, 1990) we assume that features have already been extracted and objects successfully segregated on the basis of their individual feature spaces. Our model deals with the important question of how only a limited sub span of all objects are actually selected and further encoded into VSTM.

Cattell already in the late 19th century demonstrated a surprising limit in how many objects that can be perceived at the same time – a limit only about 4 objects which may be held in the VSTM at the same time (Cattell, 1886; Cowan, 2000). This finding is independent of the number of objects visually presented at the same time (Sperling, 1960). Evidence further exist that the "magical number" of 3-to-4 objects is largely independent of how many features are encoded for each object, i.e. the complexity of the visual object, does not hold an influence on the memorial capacity of the VSTM; see (Luck & Vogel, 1997), but see also (Alvarez & Cavanagh, 2004).

Modelling the function of the VSTM, it is essential that the inherent capacity limitation is properly mimicked, since it seems a fundamental limit of the system. Most likely the VSTM would be heavily overloaded, should the system lack the ability to represent only the most salient of the visually appearing objects

The model

The model that we are presenting in this paper can actually be understood as three consecutive processes (See Figure 1).

The first process is simply extraction of visual features, we speak of this process as *'object matching'*, since we find it relevant to think that objects in the visual field are to some extent 'matched' against objects representations in Visual Long-Term Memory (VLTM). In this paper we do not consider the problem of which feature extraction techniques are biologically most plausible or perhaps technically most appropriate to use.

The second process that we shall consider in more detail is '*the attentional race'*. According to Shibuya & Bundesen (1988), all objects in the visual scene take a place in what one could think of as a race to become encoded. In Shibuya & Bundesen's race model, the 'odds' that a given object is selected as a winner in the race is directly related to the rate value with which the object participates. It is worth noting that the race is a stochastic, rather than a deterministic process, meaning that no one can beforehand predict readily which objects will win the race.

The third and last process that we shall consider is that of '*storage'* of object representation in VSTM. Inspired by (Usher & Cohen, 1999) we propose a competitive neural network model of VSTM, directly linking with several important assumptions expressed in Bundesen's Theory of Visual Attention (Bundesen, 1990).

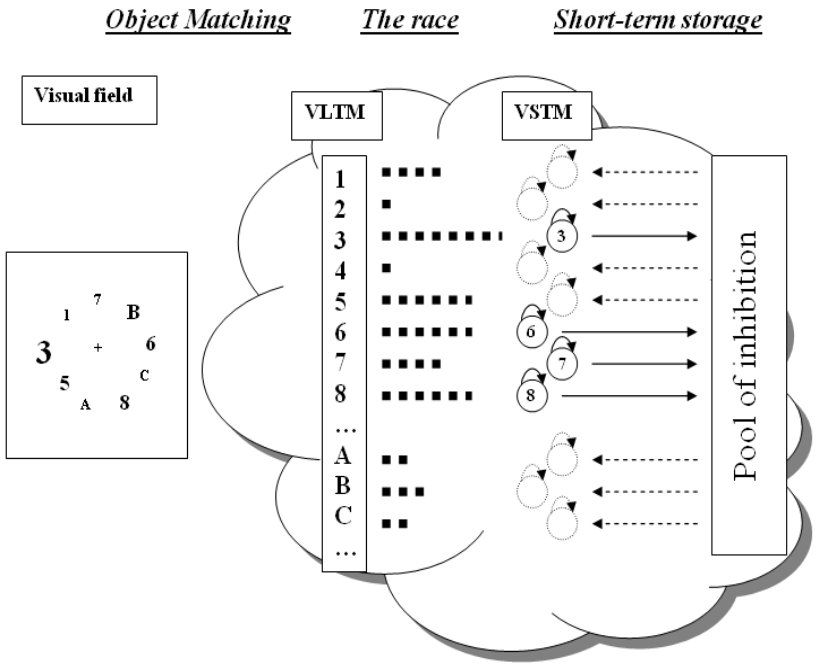

Figure 1: The Model Scheme – a partial report example. The task is to report the targets, i.e. digits and ignore the distractors, i.e. letters. The model predicts how visual elements participate in a race, where the winners become selected to be encoded in visual-short-term memory. Generally targets are processed faster than distractors, however we also see that in the example *homogeneity* is not assured, i.e. the targets (and distractors) are not of equal size (could also be contrast, letter type etc.) and therefore in the example they are illustrated as being processed with slightly different rates.

The neural theory of visual attention

The theory of visual attention (TVA) proposed by Bundesen (1990) is a unified theory of visual recognition and attentional selection. TVA provides a mathematical framework describing how the visual system is able to select individual objects in the visual field *S*, based on the visual evidence, η and the setting of two different types of visual preference parameters (pertinence, π and bias, β), representing the influence from higher cortical areas, including VLTM.

The output of the TVA-model is a set of rate parameters *v* that are directly related to the probability that a given characterization, *object x belongs to category i,* is encoded into the VSTM. The rate parameters are given by:

$$v(x,i) = \eta(x,i)\beta_i \frac{w_x}{\sum_{z \in S} w_z} \qquad (1)$$

Where the attentional w_x weight of object *x* is:

$$w_x = \sum_{j \in R} \eta(x,j)\pi_j \qquad (2)$$

Here $\eta(x,i)$ is defined as the strength of the sensory evidence that object *x* belongs to the visual category *i*. The pertinence of the visual category *j* is denoted by π_j and setting of these values effectively implements the so-called filtering mechanism. The perceptual decision bias of a visual category *i* is denoted by β_i and setting of these values conversely implements a complementary mechanism called pigeonholing.

The filtering mechanism increases the likelihood that elements belonging to a target category are perceived, without biasing perception in favor of perceiving the elements as belonging to any particular category.

Pigeonholing, conversely changes the probability that a particular category *i* is selected without affecting the conditional probability that element *x* is selected given that category *i* is selected.

A neural interpretation of TVA is given in (NTVA, Bundesen, Habekost, & Kyllingsbæk, 2005). Basically here pigeonholing (selection of features) is considered an increase in the rate of firing of neurons while filtering (selection of objects) is considered an increased mobilization of neurons.

Corresponding to the interpretation in NTVA the fraction $w_x/\sum w_z$ in equation (1), which is the relative attentional weight of object *x* compared to the weight of all objects *z* in the visual field *S*, can be directly interpreted as the relative fraction of neurons allocated to process a given object *x,*

compared to the total number of neurons processing just any object z belonging to the visual field *S*.

Each and every encoding generally takes the form *object x belongs to category i*.

Denoting the set of all features as *R* the total processing capacity, can be considered a constant *C*, which equals the sum of all encoding rates *v*; see (Bundesen, 1990).

$$C = \sum_{x \in S} \sum_{i \in R} v(x,i) \tag{3}$$

Shibuya and Bundesen (1988) assume target as well as distractor homogeneity in their whole and partial report paradigm. This means that processing capacity is distributed equally among targets as well as among distracotors. When this is the case the rates of encoding for targets, v_T and for distractors, v_D can be calculated according to the formulas:

$$v_T = \frac{C}{T + \alpha D} \qquad v_D = \frac{\alpha C}{T + \alpha D} = \alpha v_T \tag{4}$$

Where *T* and *D* denote the number of targets and distractors presented, respectively. The ratio of discrimination between distractors and targets is denoted α.

The effective exposure duration τ is smaller than the actual exposure duration *t* by an amount t_0 corresponding to the temporal threshold before conscious processing begins. However the effective exposure duration can not be negative so computationally it is set to:

$$\tau = \max(0, t - t_0) \tag{5}$$

In the neural network model that we shall now describe we adopt the parameters *C*, α and t_0 and further, following Bundesen, we make use of equation (4) and equation (5).

The neural network model of VSTM

In TVA object features are encoded independently, and further there is the assumption that only one feature needs to be encoded for the object to be stored in VSTM. On the other hand; and in agreement with (Luck & Vogel, 1997), several features of the same object can be in the encoded state, and still it will only count as if one object is stored in VSTM. For this reason, and because here we are concerned about objects rather than features encoded, we simply sum over the entire number of object features, and in this way we obtain the total encoding rate v_x for object *x*:

$$v_x = \sum_{i \in R} v(x,i) \tag{6}$$

An object *x* can enter VSTM once it receives external excitation, *G* taking the shape of Poisson distributed spike trains, arriving with the rate parameter v_x. (See Figure 2).

A neural assembly that has obtained a positive level of activation will automatically seek to re-excite itself, so that it can stay in VSTM, at the same time trying to inhibit activation in other neuron assemblies representing other objects, i.e. working to suppress other object from co-temporally being stored in VSTM.

The initial condition for the simulations is that all neuron assemblies start with an activation of zero, i.e. no objects are initially stored in VSTM. As a consequence neither re-excitation nor lateral inhibition exists, before the assemblies are externally activated.

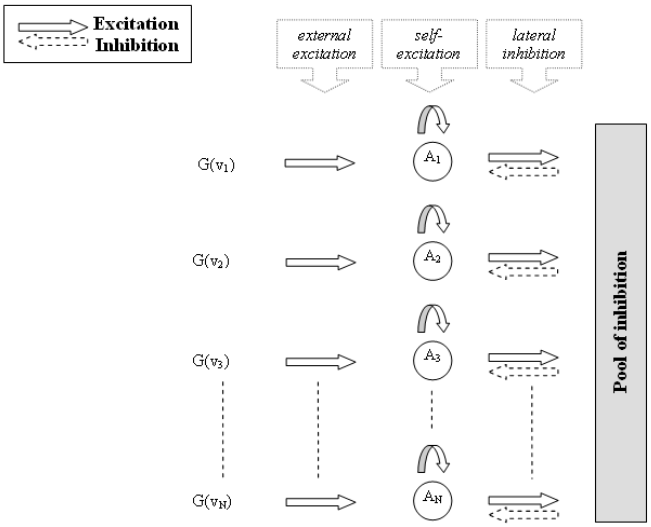

Figure 2: The neural network model of VSTM. The total number of neuron assemblies is *N* and each assembly is represented by a level of activation *A*

Implementation

The activation A_x of neuron assembly *x* (representing object *x*) is given by the first order differential equation:

$$\frac{dA_x}{dt} = -A_x + \alpha^* F(A_x) - \beta^* \sum_{z \neq x} F(A_z) + \gamma^* G(v_x) \tag{7}$$

The above equation characterizes a leaky accumulator model. There is passive decay of the activation towards the rest level, with a time constant chosen as 1, reflecting the time scale that physiologically is observed with synaptic currents (Usher & Cohen, 1999).

F is a squashing function that keeps the activation within bounds:

$$F(A) = \begin{cases} 0, & for\ A \le 0 \\ \dfrac{A}{1+A}, & for\ A > 0 \end{cases} \quad (8)$$

As a consequence of the squashing function F, the parameter α^* is the limiting value of maximal self-excitation that assemblies can up-hold and the parameter β^* is the limiting maximal value of inhibition that can be sent from one assembly to another.

Also the model assumes we can not have negative self-excitation, i.e. self-inhibition and further the model does not implement any terms that could account for excitation laterally between the assemblies. The latter effect could for instance be included if one wanted to account for semantically related objects and their effect on the number of reported objects.

The attentional significance that object i is present in the visual field R is represented by the encoding rate v_i. In our model we follow the approach from (Bundesen, 1990) and interpret this rate as the firing rate of a Poisson spike generator G. Hence γ^* characterizes the amplitude of the Poisson distributed input spikes arriving to the neuron assembly x.

The model was implemented in Matlab's Simulink toolbox. At least in the operated parameter domain we judge the stiffness of the system to be negligible so for simplicity we numerically apply Euler integration[1].

Model performance

The dataset

The data covers the performance of a single subject, participating in an extensive series of whole and partial report experiments. The subject was instructed to report targets, i.e. digits while ignoring distractors, i.e. letters displayed on an imaginary circle around a small fixation cross at the center of the screen. In practice experimental trials covered twelve whole and partial report conditions. In these the number of targets, T was between 2 and 6 and the number of distractors, D was between 0 and 6. Further, exposure durations t were varied systematically at 10, 20, 30, 40, 50, 70, 100, 150 and 200 ms. Each experimental condition was repeated 60 times but trials were mixed so that the subject had no a-priori knowledge of the experimental condition. Moreover trials were grouped into blocks to minimize the element of fatigue. Each presented character was immediately followed by a mask lasting for 500 ms. Further information can be found in (Shibuya & Bundesen, 1988).

[1] Assuming that only one spike should be allowed in each time step we must keep the integration step size sufficiently small. If the processing capacity C is 60 Hz, and the integration step size is kept at dt = 0.001, then the risk that two or more spikes will be present in a given time step is as low as 0.36 %.

Performance of the neural network model

Figure 3 shows accumulated score distributions. The score is defined as the number of targets reported correctly. The upper most curve represents the accumulated score of j = 1, i.e. the probability of reporting 1 or more targets correctly. Other curves represent accumulated probabilities for reporting at least 2, 3, 4 or even 5 targets.

Shibuya and Bundesen (1988) proposed a mixture model, mixing probabilities obtained with using a statistical model that assumed memorial capacities of either K = 3 or K = 4 respectively.

There is a relatively close fit between the proposed mixture model and the empirical data. We see however that data points obtained with exposure duration around 50 ms are generally under-fitted and more noticeably the model does not account for cases where more than 4 targets are reported, as is actually the case in two out of three of the lower most plots.

What we observe with the previous model can be considered a trade-off between two conflicting demands. The first demand is to fit the initial part of the curves, i.e. the larger the processing capacity C the steeper the curves will rise, on the other hand the second demand, which is to keep the score distribution reasonably low for long exposure durations, require that the processing capacity C is not set too high. Hence the setting of C is set subject to a compromise.

Addressing the performance of our neural network model we think it clearly meets the standard of Shibuya and Bundesen's model. The neural model does however seem to have some trouble predicting 4 recognized items in the situations where no distractors were presented. Possibly this misfit can be diminished by running a more exhaustive optimization of model parameters. The parameters used for producing the figure were: $\alpha^* = 5$, $\beta^* = 0.1$, $\gamma^* = 2$, $C = 61.5$ Hz, $t_0 = 23$ ms and $\alpha = 0.367$. Moreover, and in contrast to Shibuya and Bundesen's model, our new model readily demonstrates its capability of predicting extreme cases, where more than 4 objects are reported.

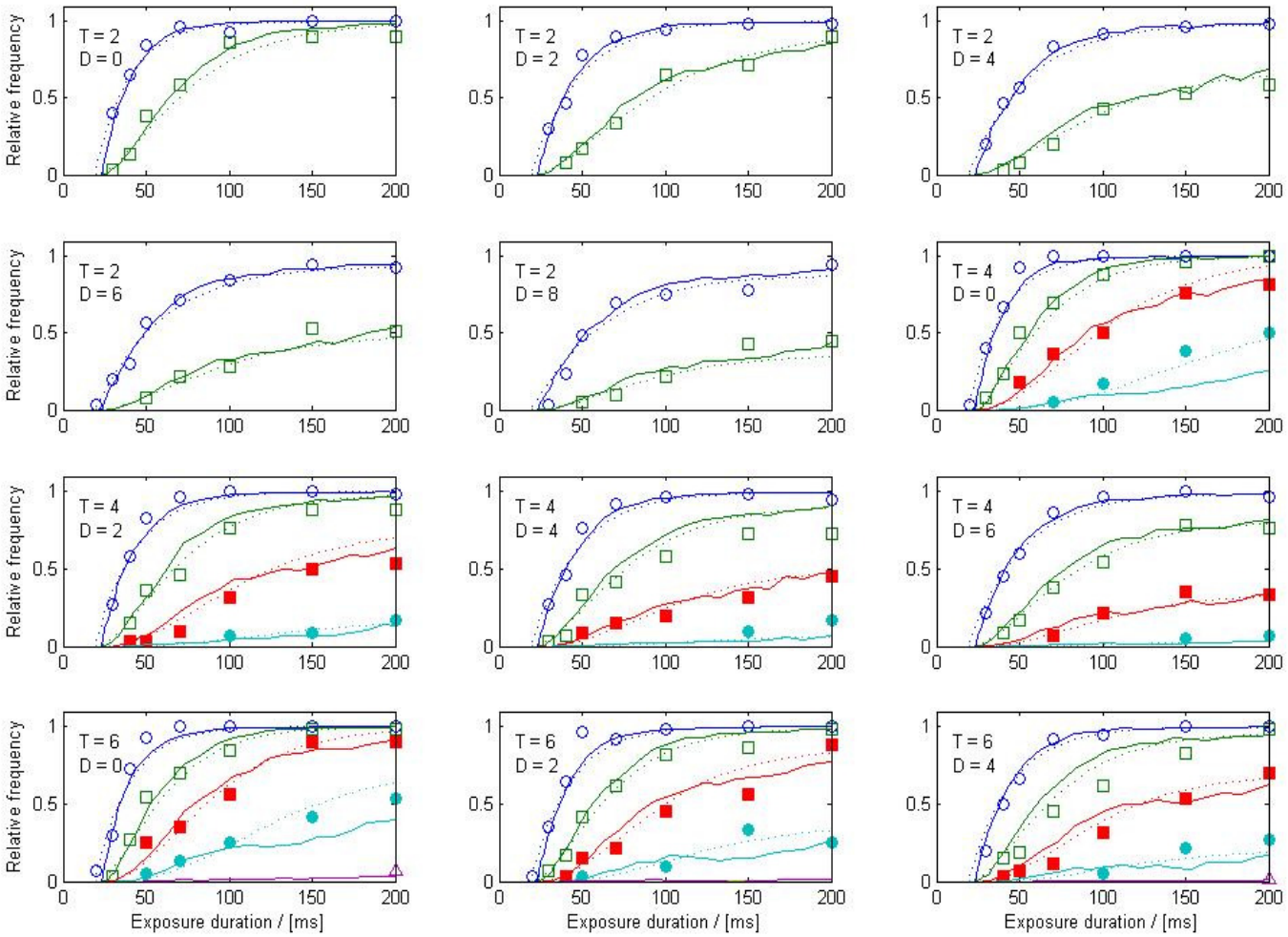

Figure 3: Accumulated score distribution for subject MP in (Shibuya & Bundesen, 1988). Probability of correctly reporting at least 1 target (blue, open circles), 2 targets (green, open squares), 3 targets (red, closed squares), 4 targets (cyan, closed circles) and 5 targets (magenta, open triangles). Empirically found values are plotted with symbols as markers. The dotted lines represent the fit by Shibuya & Bundesen (1988). Solid lines represent the performance of our neural network model. *T* and *D* denote the number of targets and distractors presented, respectively.

Discussion

This work represents an attempt to integrate the Theory of Visual Attention (Bundesen, 1990) with a simple type of winners-take-all type of network (Usher & Cohen, 1999), in the sense that the later implements a limited storage capacity of VSTM. Our new dynamic model of visual attention and VSTM is able to account for the complete set of data from whole and partial report experiments. Where the previous account by Shibuya and Bundesen (1988) treated extreme scores as outliers, the new model encompasses these as natural consequences of the internal dynamics. Further, the model explains VSTM capacity and consolidation as the result of a dynamic process rather than as a static store, which capacity is independent of processing capacity and the attentional set of the subject.

From daily life we know that humans are able to identify a very larger number of different objects. Therefore, we might think that we would have to include a neural assembly for each of these many objects candidates in our model of identification. However, what we shall argue is that our model's predictions are not affected if irrelevant neural assemblies (representing non-stimuli type of objects) are not included in the model, a useful feature which we of course make use of when we simulate with the model. The reason for this is that in the model only activated neural

assemblies affect other assemblies, and so there is no lateral inhibition from inactive neural assemblies (which irrelevant assemblies tend to be) upon any other assembly. This means that adding more irrelevant assemblies generally does not affect our conclusions, except that computationally simulations become slower.

The model described gives no account of identification of individual features of an object; however it would be possible to approach this situation by having one neural assembly in the network per object feature, rather than just one neural assembly per object. In this case assemblies representing features that belonged to the same object might be modeled as having little or no lateral inhibition, ensuring that several features of the same object can be encoded without taking up additional VSTM storage space (Luck & Vogel, 1997).

Speaking of adding more neural assemblies, we ought to touch upon what it is that we think an assembly represents. Does the assembly manifest itself in one or more neurons, and how would this relate to efficient or distributed processing? The way we think about the model is that the assemblies conceptually represent different states of neural activation. As assumed, these states interact and as we have described we suppose that feedback mechanisms play an important role in keeping the activation of the assembly sustained, allowing for visual short-term memories.

A possible confound of the model is that it does not consider internal noise, which is likely to play a key role in many neural systems. A way to deal with this would be to transform the input stage (the Poisson distributed spike trains, arriving with the rate parameter v) to a stochastic diffusion process with wiener noise process included. For this to make sense the activation threshold for consciousness would have to take a higher value than the level of initial activation.

In future studies, we think it would be relevant to explore the implication of transforming the model into a stochastic differential equation as mentioned above. Because the model is temporally dependent it would also be interesting to know if it would be able to address the dynamic consolidation in VSTM found in temporally extended paradigms such as the attentional blink paradigm and studies of attentional dwell time; e.g. (Ward, Duncan, & Shapiro, 1996). Here, consolidation in VSTM is strongly dependent on competition between items already encoded into VSTM and visual items presented at a later point in time. Incorporation of such a competitive process follows naturally from the dynamic architecture of the present model.

Acknowledgments

We would like to thank Hitomi Shibuya & Claus Bundesen for access to the experimental data in (Shibuya & Bundesen, 1988). Also we would like to thank Marius Usher for kindly verifying parameters and settings in (Usher & Cohen, 1999). Further we would like to thank the anonymous reviewers for providing helpful suggestions as well as relevant comments.

References

Alvarez, G., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. Psychological Science, 15(2), 106-111.

Bundesen, C. (1990). A Theory of Visual Attention. Psychological Review, 97(4), 523–547.

Bundesen, C., Habekost, T., & Kyllingsbæk, S. (2005). A Neural Theory of Visual Attention: Bridging Cognition and Neurophysiology. Psychological Review, 112(2), 291–328.

Cattell, J. M. (1886). The inertia of the eye and brain. Brain, (8), 295-312.

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences, 24(1), 87-185.

Luck, S., & Vogel, E. (1997). The capacity of visual working memory for features and conjunctions. Nature, 390(6657), 279-281.

Shibuya, H., & Bundesen, C. (1988). Visual Selection from Multielement Displays: Measuring and Modeling Effects of Exposure Duration. Journal of Experimental Psychology Human Perception Performance, 14(4), 591–600.

Sperling, G. (1960). The Information Available in Brief Visual Presentions. Psychological Monographs, 74(11), 1-29.

Usher, M., & Cohen, J. (1999). Short Term Memory and Selection Processes in a Frontal-Lobe Model. In Connectionist Models in Cognitive Neuroscience (pp. 78-91). Birmingham: Springer-Verlag.

Ward, R., Duncan, J., & Shapiro, K. (1996). The slow time-course of visual attention. Cognitive Psychology, 30(1), 79-109.