A Dynamic Neural Field Theory of Multi-Item Visual Working Memory and Change Detection

Jeffrey S. Johnson (jeffrey-johnson-2@uiowa.edu)
John P. Spencer (john-spencer@uiowa.edu)
Department of Psychology, E11 Seashore Hall
Iowa City, IA 52242 USA

Gregor Schöner (gregor.schoener@neuroinformatik.ruhr-uni-bochum.de)
Institut für Neuroinformatik, Ruhr-University
44780 Bochum, Germany

Abstract

Many visually-guided behaviors rely critically on the ability to maintain visual information in working memory. However, to date, there are few formal models of visual working memory (VWM) that directly interface with the growing empirical literature on this basic cognitive system. In particular, no current theories address both the maintenance of multiple items in VWM *and* the process of change detection within a neurally-plausible framework. In the present study, we describe such an approach, along with initial data from a change detection task that confirm a novel prediction of our model.

Introduction

Visual working memory (VWM) plays a central role in everyday activities ranging from the integration of visual information obtained across successive saccadic eye movements, to maintaining visual representations in an active state in the service of complex cognitive tasks. Moreover, impaired working memory functioning has been implicated in the broad array of cognitive impairments associated with illnesses such as schizophrenia (Keefe, 2000). Given these ties to both basic and applied issues, there has been an explosion of interest in the function of the working memory system in the past decade (see Miyake & Shah, 1999). In addition, there has been a growing push to develop neurally-plausible models of VWM that can both synthesize the extant literature and shed light on the profound impairments seen in atypical populations.

The current state-of-the-art in the study of VWM is to use a canonical change detection task to assess the characteristics of VWM. In this task, observers are shown two visual displays separated by a brief delay interval and are asked to report whether they are the same or different. This task requires observers to maintain multiple stimuli in memory, and compare these memory representations to incoming perceptual representations in order to generate a same or different response at test. Research using this paradigm has revealed that VWM has a very limited capacity and appears to store items in the form of integrated object representations rather than as individual features. Additionally, electrophysiological and fMRI studies of change detection have begun to isolate the neural substrates of these functions.

Although most theoretical accounts of these findings have remained verbal/conceptual in nature, some have moved in the direction of formal theory. For example, Raffone and Wolters (2001) have developed a neural network model where WM for objects is maintained by synchronized firing among neurons representing an object's features. Although models of this type offer exciting links between brain and behavior, they have several well-known limitations, including a relatively low tolerance to noise, and a simplified approach to the representation of space. Perhaps most critically, such models have yet to specify an approach to change detection that addresses how populations of neurons compare incoming perceptual representations to items stored in memory, giving rise to the "same" and "different" responses required by the task.

In the present study, we describe a first step in the development of a new approach to VWM and change detection that builds on a neurally plausible, process-based framework for understanding spatial cognition: the Dynamic Neural Field Theory (DNFT) (Spencer & Schöner, 2003, 2006). This framework allows a tight relationship between theory and experimentation, and has provided important insights into the processes underlying spatial working memory and the development of this cognitive system (Schutte & Spencer, 2003). Here we extend this approach to address multi-item VWM and change detection.

VWM for features and objects: Insights from change detection.

Research investigating working memory for nonspatial object properties—or visual working memory—began to explode in the mid-1990s, due in large part to the widespread use of change detection tasks. For example, in a highly influential study, Luck and Vogel (1997; Vogel, *et al*., 2001) investigated the storage of visual features and feature bindings in VWM in a series of change-detection experiments using visual arrays composed of simple colored shapes. Participants were shown arrays of 1 to 12 items (e.g., colored rectangles) for 100 ms, followed by a 900-ms delay interval and then a test array that remained visible for 2000 ms. When the test array appeared, it was either

identical to the original display, or one item had been changed (e.g., to a different color or orientation). They found that same/different judgment accuracy sharply declined for arrays containing more than four items, prompting the conclusion that VWM has a limited capacity of approximately 3-4 items, in keeping with other findings (Cowan, 2001; Irwin & Andrews, 1996; Sperling, 1960). Additional research using this paradigm has suggested that VWM stores integrated object representations (Luck & Vogel, 1997; Vogel, Woodman, & Luck, 2001), rather than single features, although the exact nature of the representations maintained in VWM remains a contentious issue (see Alvarez & Cavanagh, 2004 and Wheeler & Treisman, 2002, for alternative proposals).

In a related line of research, the change-detection paradigm has begun to be used to investigate the neural substrates of VWM. For example, converging evidence from event-related potential (ERP) (Vogel & Machizawa, 2005) and functional Magnetic Resonance Imaging (fMRI) (Todd & Marois, 2004) studies of change detection have revealed localized neural activity in the posterior parietal cortex that is specifically related to the amount of information being held in VWM. Additionally, Xu & Chun (2006) have provided fMRI evidence suggesting that the maintenance of object properties (e.g., the color or orientation of a stimulus) and locations in WM rely on the activation of distinct cortical networks in the occipital-temporal cortex, and regions of the posterior parietal cortex, respectively.

Additionally, ERP and fMRI studies have begun to focus on neural activity associated with the detection of changes at test. For example, using fMRI, Pessoa and Ungerleider (2004) found that the detection of orientation changes was associated with activation of a network of brain areas (e.g., frontal, parietal, and anterior cingulate cortex) known to be involved in the control of attention (Kastner & Ungerleider, 2000). On the basis of these findings, they proposed that these regions may be involved in orienting the eyes and/or attention to the location of the change, facilitating further processing of the changed item. Consistent with this proposal, a recent series of ERP and eye-movement studies looking at working memory for color (Hyun, 2005) have demonstrated that both attention and the eyes are rapidly shifted to the location of the changed item at test.

In summary, research using the change-detection task has established several important facts about VWM. First, VWM has a small storage capacity of approximately 3-4 items, which appears to be limited by the number of objects that are stored rather than the number of features. Second, the maintenance of information in VWM appears to rely on activation of regions of the occipital-temporal and posterior parietal cortex, which may be differentially involved in the maintenance of object-property and spatial information, respectively. Finally, the detection of change in WM tasks engages neural systems that may play a role in rapidly orienting attention and the eyes to the changed location.

Although studies of change detection have begun to make significant contributions to our understanding of VWM at both the behavioral and neural levels, few theoretical models have been formulated within a neurally-plausible framework that could effectively address both lines of research. Thus, a critical goal in this area is to move in the direction of neurally plausible models that allow us to link behavioral performance in change detection tasks to the underlying neural substrate.

A neural synchrony approach to VWM

An important first step in this direction was suggested by Vogel, et al. (2001), who proposed that VWM representations are stored in cell assemblies in which neurons that code the features of an object are linked by virtue of synchronized firing. Raffone and Wolters (2001) have created a detailed neural network model (hereafter referred to as 'Sync') of this *neural synchrony hypothesis* that captures many of the properties of VWM discussed above. In Sync, assemblies of neurons modeling inferotemporal cortex (IT) are linked to matching assemblies that model prefrontal cortex (PF). Individual neurons in IT code for the presence of individual feature values, while the assemblies represent perceived objects. When stimulated, the IT-PF system enters into sustained oscillation. Synchronized activity within the IT-PF assemblies establishes links among features, temporally "binding" these features into objects. Inhibition among these assemblies effectively chunk input into separate objects, which are separately maintained in WM. The model accounts for the limited capacity of VWM in terms of spurious synchronization, that is, the increasing instability of temporal binding as the number of chunks increases. Because the assemblies code for bound feature sets, the capacity is determined by the number of objects rather than the number of features, consistent with observed data.

Limitations of Sync. Although Sync offers formal ties between neural processes and capacity limits in VWM, there are several limitations of this model. First, a central question with any neurally plausible approach to working memory is how stable sustained activation is in the face of noise. How stable are self-sustaining oscillations in Sync? This is not entirely clear. Raffone and Wolters (2001) demonstrated that Sync can maintain synchronized oscillations for 300 ms in the absence of input. Importantly, though, the amount of noise was dramatically reduced during this rather short memory delay (the noise was 10 times smaller during the delay interval v. when the perceptual input was "on"). Moreover, no data were presented from multiple simulations of the model, so it is not clear whether Sync's memory for features captures a realistic balance between robust WM and variability in performance.

Second, Sync treats spatial dimensions like any other feature dimension; consequently, this model does not address several lines of evidence suggesting that space has a special status in VWM tasks. For example, space plays an important role in the calibration of reference frames and in

linking perceptual and cognitive systems to action systems. Additionally, one of the central insights in the field of visual attention is that attending to a spatial location can bind the features at that location together into an object representation (Treisman & Gelade, 1980). Thus, space may function as the medium or ground that facilitates binding, rather than being just another object property.

Finally, and perhaps most importantly, the framework provided by Sync only addresses the maintenance of information in VWM and does not provide a process model of the primary task used to probe working memory—change detection. That is, the process of comparing incoming perceptual information (e.g. the test array) with VWM representations and generating the "same" and "different" responses required by the task has not been specified to date. Rather, Raffone and Wolters relied solely on single simulations to capture aspects of results from Luck and Vogel (1997). In particular, they simply observed whether oscillations were maintained (or not) as the set size was increased. This provides very limited ties to participants' performance.

In summary, the Sync framework represents an important first step toward a formal model of VWM based on neural principles. However, a number of concrete limitations suggest that it might be fruitful to look at other models of WM to help explain the rapidly accumulating empirical database on VWM.

The Dynamic Neural Field Theory of SWM

Over the last several years, we have developed a neurally-plausible theoretical framework—the Dynamic Neural Field Theory (DNFT)—to capture the processes that underlie spatial working memory (SWM) (Schutte & Spencer, 2003; Spencer & Schöner, 2003, 2006). To describe the theory, consider an activation field defined over a metric spatial dimension, x (e.g., the direction of a target). The continuous evolution of the activation field is described by an activation dynamics, that is, a differential equation which generates the temporal evolution of the field by specifying a rate of change, $dw(x,t)/dt$, for every activation level, $w(x,t)$, at every field location, x, and any moment in time, t. The field achieves stable patterns of activation through time via an inverse relationship between the rate of change and the current level of activation. This means that at high levels of activation, negative rates of change drive activation down, while at low levels, positive rates of change drive activation up.

The activation level that emerges from this basic stabilization mechanism is a function of the balance of different inputs to the field (e.g., from perceptual systems, long-term memory, etc.) and neural interactions within the field. We use a locally excitatory/laterally inhibitory form of interaction captured in Figure 1A (see also Amari, 1989; Amari & Arbib, 1977; Compte, Brunel, Goldman-Rakic, & Wang, 2000 for similar formulations). According to this type of interaction, neurons that "code" for similar values along the spatial dimension, x (e.g., similar directions in space), excite one another while neurons that code for very different values (e.g., different directions in space) inhibit one another. This form of interaction allows self-sustained peaks of activation to be maintained following the withdrawal of the stimulus (i.e., the memory display). In addition, such interaction can cause peaks of activation to "drift" over delays, depending on activation at other field sites and the current noise level.

The basic architecture of the DNF model within which these concepts are implemented can be seen in Figure 1. Figure 1B shows the excitatory layer of a two-layered perceptual field, PF (u), and Figure 1C shows the excitatory layer of a two-layered spatial working memory field, SWM (w). Both of these layers are coupled to a single layer of inhibitory interneurons, Inhib (v) (see reciprocal solid (excitatory) and dashed (inibitory) arrows between PF and Inhib as well as between SWM and Inhib). In addition, the perceptual layer passes excitatory input to the SWM field.

The simulation shown in Figure 1, panels B-D, depicts a single trial in a SWM task. In each panel, the direction of the targets in the task space is shown along x; y captures the elapsed time from the start of the trial; and z shows the activation of each site in the field. At the start of the trial PF (u) builds a small peak of activation at 180°, reflecting perception of a salient reference frame in the environment (e.g., the midline of the task space). Next, the target appears at 220°. This creates a peak of activation centered at that

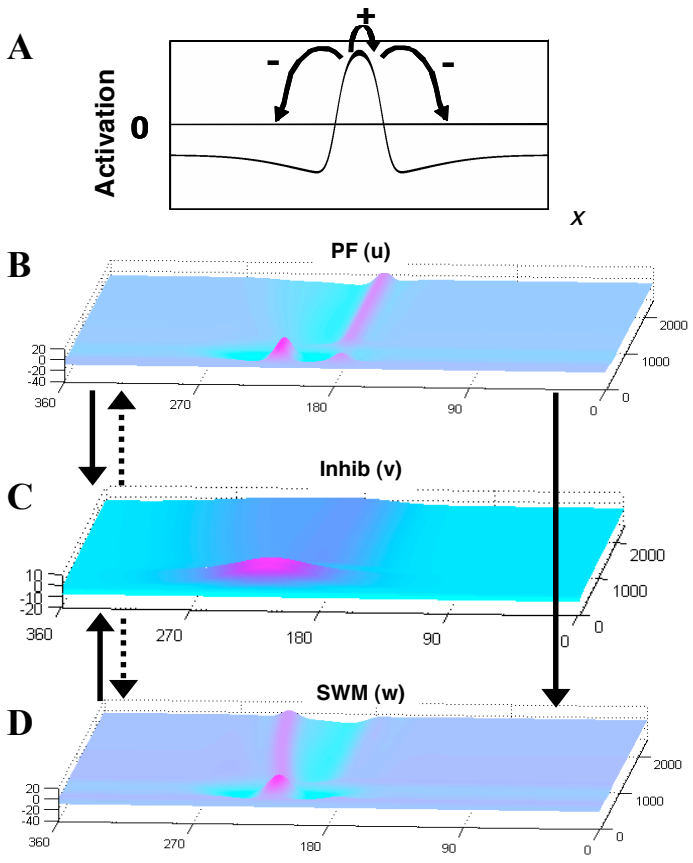

Figure 1. The DNFT of Spatial Working Memory

location. When the target disappears, a peak re-forms in PF (u) at 180° as the system locks onto the reference cues in the task space. Panel D shows the effect of coupling PF (u) to SWM (w). At the start of the trial, SWM (w) receives relatively weak reference input from PF (u). Next, the target is turned on, passing strong target-related input into the working memory field. This event—combined with a boost in the resting level of SWM (w) and Inhib (v)—moves the working memory field into a strongly self-sustaining state. An active memory of the target location is maintained in working memory throughout the delay as a result of coupling between the SWM (w) and Inhib (v) fields, which implements the form of interaction depicted in Figure 1A. Importantly, this occurs even though PF (u) has re-acquired the reference frame. However, weak reference-related input to the inhibitory layer (v) causes the peak of activation in SWM to "drift" away from the midline of the task space, consistent with observed data.

Limitations of the DNFT

As with Sync, the DNFT is limited in several respects. Most critically, the model described above focuses solely on SWM for single items. To capture performance in change detection tasks, we must modify the model in three key ways. First, we must extend this approach to capture WM for non-spatial object properties such as color and orientation. Second, the model must be expanded to address WM for multiple items. Finally, we must specify the processes that underlie change detection decisions (i.e., 'same' or 'different' responses).

A Dynamic Neural Field Theory of Multi-Item VWM and Change Detection

To address these limitations, we have extended the DNFT to address multi-item WM, where the to-be-remembered stimulus can be either featural or spatial in nature. To capture this, we have introduced the concept of a 1D feature WM field (FWM), which has all of the characteristics of a 1D SWM field, but the metric dimension along which activation is defined is featural in nature (e.g. hue, orientation, line length). Note that the FWM field captures more than just a re-labeling of an axis in our model. The claim here is that WM for metric features shares all of the properties captured by dynamic neural fields—WM as stabilized peaks, coupling between perception, WM and LTM, metric interactions leading to "drift", and so on. The "Mexican-Hat" interaction profile depicted in Figure 1A, where inhibition is stronger near the focus of excitation, allows a multi-peak solution of the field dynamics with moderate levels of global inhibition. This allows the locally excitatory interactions associated with each peak to be isolated by lateral inhibition, while keeping the total amount of inhibition in the field low enough that multiple items can be maintained. However, as more items are encoded in working memory, the overall amount of inhibition is also increased, which, together with metric interactions between peaks, provides a natural basis for capacity limits in the model.

To capture performance in change-detection tasks, we have extended a recent dynamic field model of position discrimination (Simmering *et al*., in press). A 1D version of the model is shown in Figure 2. As with the DNF model of SWM described above, this model consists of a 1D feature-selective perceptual field that provides afferent input to a layer of inhibitory neurons and to an excitatory feature WM field (FWM (w)). Excitatory and inhibitory interactions between each of these three fields allow the network as a whole to function as a "difference" detector. Specifically, peaks in WM produce localized regions of inhibition in PF (u) via inhibitory feedback from Inhib (v). Thus, PF (u) is only able to build a new peak when a change occurs at test, which signals the presence of a new input that needs to be attended to.

The simulations shown in Figure 2A and 2B illustrate how "same" and "different" responses emerge in the model. Both simulations show three peaks of activation that are built following the presentation of a memory array (e.g. three colored squares). Note that these peaks are only transiently sustained in PF (u), but are maintained throughout the delay interval in the FWM (w) field (bottom panels). Each of the peaks in WM activates similarly-tuned neurons in the Inhib (v) field through excitatory feedback, which then projects localized inhibition back to PF (u), inhibiting similarly-tuned neurons in that field. At the end of the delay, a single test item is presented to probe WM for color. At the same time, the resting levels of PF (u) and Inhib (v) are boosted, which prepares PF (u) for new inputs and stabilizes the peaks in WM. In panel A, the test item is the same as one of the items being held in working memory. As a result, PF (u) is unable to build a peak due to strong localized inhibition from the Inhib (v) field at that location,

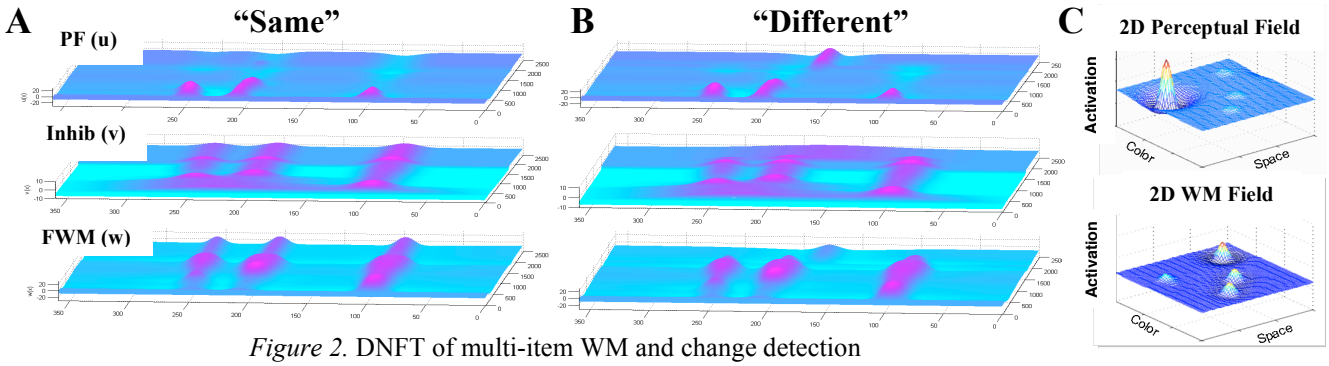

Figure 2. DNFT of multi-item WM and change detection

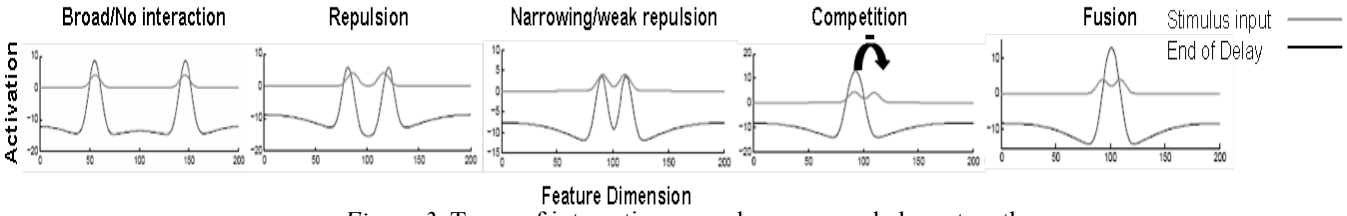

Figure 3. Types of interaction as peaks are moved closer together

and therefore the three peaks remain stable in WM at test and a "same" response is produced. In panel B, however, the test item is a new color that wasn't present in the memory array, and thus the input comes in at a relatively uninhibited region of PF (*u*). In this case, a stable self-sustained peak builds in PF (*u*), which suppresses each of the peaks in WM, producing a "different" response.

Although the model shown in Figure 2 is capable of detecting color changes, it isn't clear how the system could identify where in the world the change occurred. This is an important issue because adaptive, visually-guided behavior depends critically on knowing not only *that* a change has occurred but *where* it has occurred. To address this, we have developed a 2D version of the basic change detection model (see Figure 2C). This model introduces the concept of a two-dimensional feature-space field (FSWM) that captures WM for what-where conjunctions.

The generation of a "change" response in the 2D version of the model is shown in Figure 2C. As with the 1D model, the presence of a peak in the perceptual field at test globally suppresses activation in the WM field. However, in addition to signaling the presence of a feature change, the 2D perceptual field also contains information about where in the task space the change occurred. This signal could serve as the basis for shifting attention and/or the eyes to the location of the change, in keeping with recent evidence (Hyun, 2005; Pessoa & Ungerleider, 2004).

Additionally, metric properties of FSWM fields together with the change detection process described previously leads to a set of novel predictions. In particular, the local excitation/lateral inhibition function underlying sustained activation in the DNFT leads to interactions between peaks when more than one item is being held in WM. The specific form of the interaction depends critically on how similar the items are along a given dimension (e.g., color). For ease of exposition, the full pattern of multi-peak interactions as items are made more similar is illustrated in Figure 3 using the 1D change detection model described above (see Figure 1 B-D). It should be noted, however, that the same metric interactions are also present in the 2D version of the model.

As the far left panel of Figure 3 shows, when stimuli are far apart along a given dimension (e.g., color), WM peaks will be broad and will not interact. At smaller separations, peaks repel one another due to strong lateral inhibition between the peaks (compare the positioning of the large WM peaks at the end of the delay to the position of the stimulus input). At even smaller separations, peaks should be stable and narrow with only slight repulsion, because lateral inhibition from one peak begins to extend to the "far" side of the other peak. At still smaller separations, peaks should compete, leading to the destabilization of one peak in favor of the other. Finally, at the extreme limit, the activation profiles associated with each item will fuse, forming a single WM peak.

This pattern of multi-peak interactions combined with the change detection mechanism described above leads to a set of novel predictions we have begun to test in our lab. First, the model predicts *enhanced change detection when items are highly similar!* This occurs due to the narrowing of close peaks. When WM peaks are narrower, they leave narrower inhibitory traces PF (*u*), allowing this field to detect even small changes in input. For instance, for the simulations shown in Figure 3, a 30-unit change was required to produce a different response for the unique target on the right, compared to a 20-unit change when each of the similar targets on the left were probed.

This highly counterintuitive prediction has been confirmed in a recent study comparing color change detection accuracy for close vs. far colors (Luck, Lin, & Hollingworth, 2005). In this study, target items were presented sequentially to prevent color contrast effects, and a single item was probed at test. As can be seen in Figure 4, color change-detection accuracy was significantly better when target colors were drawn from a close color set vs. a far color set. Importantly, this was the case regardless of the serial position of the probed item (i.e., probing the 1st, 2nd, or 3rd target item presented).

A second prediction of the model is derived from the "repulsion" type of interaction shown in Figure 3. Such interactions predict that there will be asymmetries in change detection for similar items depending on the probed direction. In particular, change detection should be worse when probed in directions consistent with repulsion vs. in the opposite direction. We are currently testing this novel prediction in our lab.

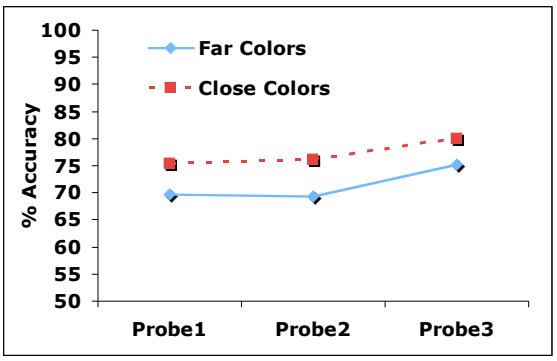

Figure 4. Enhanced change detection with close colors

Conclusion

To our knowledge, the model proposed here provides the first neurally-plausible theory of the processes that underlie change detection. Thus, unlike Sync, our approach makes explicit how incoming perceptual representations are compared to items in WM, and how this process leads to the generation of the same/different responses required by the task. Our approach also retains the characteristics of our previous model of SWM, making the link between feature-based WM and spatially-based action systems (e.g. visual attention and eye movement systems) explicit and relatively straightforward. The proposed model represents a first step in the development of a comprehensive theory of visual working memory based on neural principles. Future efforts will focus on extending the model to more fully explore the integration of what and where in feature-space working memory fields (see Figure 2C), and to address working memory for more complex, multi-feature objects.

Acknowledgements

This work was made possible by grants from the National Science Foundation (BCS 00-91757 and HSD 0527698) and the National Institute of Mental Health (R01 MH62480) awarded to John P. Spencer.

References

Alvarez, G.A., & 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*, 106-111

Amari, S. (1989). Dynamical stability of formation of cortical maps. In M. A. Arbib & S. Amari (Eds.), *Dynamic Interactions in Neural Networks: Models and Data* (pp. 15-34). New York, NY: Springer.

Amari, S., & Arbib, M. A. (1977). Competition and cooperation in neural nets. In J. Metzler (Ed.), *Systems Neuroscience* (pp. 119-165). New York: Academic Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87-185.

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review, 109*, 545-572.

Hyun, J.-S. (2005). How are visual working memory representations compared with perceptual inputs? University of Iowa, Iowa City, Iowa.

Irwin, D. E., & Andrews, R. V. (1996). integration and accumulation of information across saccadic eye movements. In T. Inui & J. L. McClelland (Eds.), *Attention and Performance XVI: Information Integration in Perception and Communication*. Cambridge, MA: MIT Press.

Kastner, S., & Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience, 23*, 315-341.

Keefe, R. S. (2000). Working memory dysfunction and its relevance to schizophrenia. In T. Sharma & P. D. Harvery (Eds.), *Cognition in Schizophrenia: Impairments, Importance, and Treatment Strategies*. Oxford: Oxford University Press.

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

Luck, S.J., Lin, P-H., & Hollingworth, A. (2005). Similarity and interference in visual working memory. Talk presented to the 46th Annual Meeting of the Psychonomic Society, Toronto, Ontario, Canada, November 2005.

Miyake, A., & Shah, P. (Eds.). (1999). *Models of Working Memory*. Cambridge, UK: Cambridge University Press.

Pessoa, L., & Ungerleider, L. G. (2004). Neural correlates of change detection and change blindness in a working memory task. *Cerebral Cortex, 14*(5), 511-520.

Raffone, A., & Wolters, G. (2001). A cortical mechanism for binding in visual working memory. *Journal of Cognitive Neuroscience, 13*, 766-785.

Schutte, A. R., & Spencer, J. P. (2003). Testing the dynamic field theory: Working memory for locations becomes more spatially precise over development. *Child Development, 74*(5), 1393-1417.

Simmering, V. R., Spencer, J. P., & Schöner, G. (in press). Reference-related inhibition produces enhanced position discrimination and fast repulsion near axes of symmetry *Perception and Psychophysics*.

Spencer, J. P., & Schöner, G. (2003). Bridging the representational gap in the dynamical systems approach to development. *Developmental Science, 6*, 392-412.

Spencer, J. P., & Schöner, G. (2006). A dynamic field theory of spatial working memory. *Manuscript in preparation*.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs, 74*, (Whole No. 498).

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature, 428*, 751-754.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*, 97-136.

Vogel, E. K., & Machizawa, M. G. (2005). Neural activity predicts individual differences in visual working memory capacity. *Nature, 428*, 748-751.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 27*, 92-114.

Wheeler, M. & Triesman A.M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General, 131*, 48-64.

Xu, Y. & Chun, M.M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature, 440*, 91-95.