8

Integrating "What" and "Where"

Visual Working Memory for Objects in a Scene

SEBASTIAN SCHNEEGANS, JOHN P. SPENCER, AND GREGOR SCHÖNER

Imagine sitting at a computer late at night. As you close your eyes, you have a compelling sense of the space around you. You can point to the coffee cup on the right and reach for the phone to the left (making sure not to bump the water cup). And this map of the local surroundings is impressively updateable—objects may have come and gone over the past 5 minutes, but it seems trivial to keep track of them. On the other hand, the mental representation of your surroundings also has severe limitations. If you try to bring to mind a complete image of your desk with all the objects on it, you will likely have considerable trouble recalling the fine details of objects and arrangements. And when you open your eyes again, you may realize that you have missed a lot of things, and that some items that you thought you remembered really look quite different from what you had imagined.

Experimental research on visuospatial cognition and memory has elucidated the impressive capabilities of adults to form an internal representation of a visual scene but also the sometimes shocking limitations of human scene perception and memory. Adults can form maps quickly: They can form detailed scene representations of three to four objects in a few hundred milliseconds, and these representations can subserve the detection of changes in those objects a second or two later, even when all the objects have swapped positions (Johnson, Hollingworth, & Luck, 2008; Wheeler & Treisman, 2002). Moreover, give an adult 19 more seconds to scan the richly structured surrounds, and this person can detect often subtle changes in the details of objects in complex scenes after viewing more than 400 total objects—up to 24 hours later (Hollingworth, 2005)! Other experiments, however, show that human scene memory can also be surprisingly limited in certain situations. Studies on change blindness demonstrate that observers frequently overlook even substantial changes in an image they are studying, as long as the change co-occurs with some visual disruption such as an eye movement (Pashler, 1988; Rensink, O'Regan, & Clark, 1997). People can even fail to notice that they are talking to a new person when that person changes from one moment to the next (Simons & Levin, 1998). These studies highlight that humans do not form instant and photographic memories of visual scenes. Instead, forming a scene memory, as well as using that memory for change detection and other tasks, is an active process that takes time and uses neural resources and thus comes with specific limitations in its capacities.

What exactly are the processes that underlie our ability to form a usable internal representation of a visual scene? To construct a scene representation, one must parse complex visual environments, which often involves visual search and object-based attention (Luck & Vecera, 2002; Wolfe, 1998). One must establish a spatial frame (McNamara, Halpin, & Hardy, 1992; Pick, Montello, & Somerville, 1988) and stay aligned with this frame despite continual eye, head, and whole-body movements (Darling & Miller, 1993; Soechting & Flanders, 1989). Moreover, one must establish robust object representations in real time that are localized and updateable in this frame (Huttenlocher, Hedges, & Duncan, 1991; Huttenlocher, Hedges, & Vevea, 2000; Spencer & Hund, 2002). And all of this must be coordinated by complex neural processes, from object-related cells in inferotemporal cortex (Baker & Olson, 2002; Tanaka, 2000), to cells in parietal cortex involved in spatial transformations (Andersen, 1995), to cells in prefrontal cortex involved in the maintenance of spatial and featural information in working memory (Rao, Rainer, & Miller, 1997).

Typically, these different aspects of visuospatial cognition have been studied separately. Although this piecemeal approach has been highly successful, it has created an often overlooked challenge: It may not be so easy to put the pieces back together again. A growing number of examples demonstrate that ignoring integration can lead to major theoretical quandaries (Bridgeman, Gemmer, Forsman, & Huemer, 2000; Jackendoff, 1996). Central to these theoretical challenges is a key question: Can piecemeal accounts of visual-cognition "scale up" to something as complex as an updateable scene representation?

Different pieces of the visuospatial cognition puzzle have been discussed in previous chapters. Chapter 5 presented a DF model of visual attention that represented multifeature objects in a neurally grounded manner. The model also captured behavioral data showing the influence of working memory (WM) on attention and saccade planning. This highlighted the interplay between selection at the level of working memory and integration at the level of a retinal representation. Next, in Chapter 6, we presented a neural system that could detect changes in object features by operating on lower-dimensional fields. This captured the behavioral details of how people actively compare, for instance, a WM for colors to a new percept.

Critically, however, objects don't live in the abstract world of features—objects are integrated wholes anchored to spatial positions in an "allocentric" or world-centered frame. In Chapter 7, we examined the spatial side of this problem: How do people know where something is, given changes in the position of the eyes, head, and body? Here, we discussed a mechanism for updating spatial positions—creating a body-centered or world-centered frame from given retinal information. And we showed the power of this transformation mechanism by using it to understand aspects of how humans use spatial language.

Of course, the spatial-language examples were limited in a fundamental way—we don't just use objects to establish a reference frame; we also want to fill that frame with content. Thus, we need to bring together spatial and featural information at the level of an allocentric or scene-centered frame. That's one goal of this chapter—to integrate "what" and "where." And in the process, we will shed light on how people build a fast, flexible representation of a local scene such that they can detect changes in the world and update their WM in a few hundred milliseconds.

More generally, this chapter tackles the theoretical challenge of scaling up from simpler to more complex neural systems. We do this here by integrating the piecemeal accounts from Chapters 5–7 and demonstrating that an integrated system can form a fully functional scene representation that interfaces with human behavioral data. Chapter 9 will continue this arc, asking whether the integrated theory of visuospatial cognition can be extended into the real world in the context of an autonomous robot. This highlights the broad functionality that emerges from the integrated system. Together, Chapters 8 and 9 demonstrate that DFT does, in fact, scale up from simpler to complex neural systems.

TOWARD AN INTEGRATED THEORY OF VISUAL WORKING MEMORY

The concrete goal we have set for ourselves in this chapter is to build a DF architecture that supports the active representation of integrated objects in an allocentric reference frame, that is, a scene representation that identifies which object is where. Our approach derives from the large literature exploring the nature of object and scene representations using novel objects with simple features (e.g., colored squares, oriented lines). Researchers within this tradition have examined how representations of simple novel objects are formed (Vogel, Woodman, & Luck, 2006); the role of attention and WM in the encoding, maintenance, and retrieval of objects (Luck & Vogel, 1997; Rensink, 2000, 2002); how objects are linked to configurations and scenes (Hollingworth, 2006, 2007); and how object representations are updated (Moore, Mordkoff, & Enns, 2007).

We are focused here on this literature for three main reasons. First, this literature presents some daunting theoretical challenges. For instance, Luck and Vogel (1997) showed that people can form multiple object representations in 100 ms that are sufficiently detailed to detect a change in 1 of 16 feature values (4 simple feature values for each of four objects) 900 ms later. This clearly requires a fast and flexible cognitive system.

A second reason for focusing on novel, simple objects is that we are ultimately interested in forming a theory of visuospatial cognition that speaks to developmental origins. Evidence suggests that infants have developed relatively well-structured cortical fields for simple features like color and orientation by midway through their first year

(Banks, Shannon, & Granrud, 1993; Bornstein, Krinsky, & Benasich, 1986; Teller & Bornstein, 1987). Such fields might then serve as the foundation on which objects are built. Thus, by focusing on the representation of novel, simple objects, we hope to connect our interests in adult visuospatial cognition with those related to the very early integration of "where" with "what." We will pick up on this theme in Part 3 of the book.

A third reason for focusing on novel, simple objects is to tap into a rich literature on the neurophysiology of object representations. Neurophysiological evidence suggests a functional and anatomical segregation of the visual system into dorsal and ventral streams that represent spatial location ("where") and object property information ("what"), respectively (Ungerleider & Mishkin, 1982). The dorsal pathway extends from early visual cortex through the parietal lobe into the frontal cortex. Recall from Chapter 7, for instance, that regions of the parietal cortex are critically involved in spatial transformations.

Regarding the ventral pathway, converging evidence from electrophysiological recording studies in monkeys (Felleman & Van Essen, 1991; Livingstone & Hubel, 1988) and functional imaging studies in humans (Pessoa & Ungerleider, 2004; Todd & Marois, 2004; Tootell et al., 1998) suggests that object properties such as color, form, size, and direction of motion are coded in a distributed manner through the parallel activation of large numbers of neurons across different neural populations (Fujita, Tanaka, Ito, & Cheng, 1992; Komatsu & Ideura, 1993; Llinas & Paré, 1996). As one progresses through this pathway from primary visual area V1, through extrastriate areas V2–V4, and on to areas TEO and TE of the inferior temporal lobe, there are several clear changes in neural response properties (Luck, Girelli, McDermott, & Ford, 1997). First, there is an increase in the complexity of the features coded. For example, whereas neurons in V1 respond preferentially to rather simple stimuli such as oriented line segments, cells in TE may respond to complex stimuli such as faces (Desimone, Albright, Gross, & Bruce, 1984; Desimone & Gross, 1979; Tanaka, 1996). Second, there is an increase in receptive field sizes and an accompanying decrease in the spatial resolution of receptive fields for individual neurons (Desimone & Gross, 1979; Gross, Rocha-Miranda, & Bender, 1972). Note that even though spatial resolution decreases, object representations in the ventral pathway are still anchored to spatial positions. For instance, studies show that position dependence persists throughout the ventral visual pathway, even into areas such as the inferior temporal cortex, which was once thought to be spatially invariant (Aggelopoulos & Rolls, 2005; DiCarlo & Maunsell, 2003; Op De Beeck & Vogels, 2000; for review, see, Kravitz, Vinson, & Baker, 2008).

Although this type of distributed encoding can be computationally efficient, as discussed in Chapter 5, it can be difficult to determine which features belong together as attributes of a single object (Damasio, 1989; von der Malsburg, 1996; Treisman, 1996, 1999). In Chapter 5, we discussed a solution to this problem which is conceptually tied to Treisman's feature integration theory (Treisman & Gelade, 1980): By allocating visual-selective attention to occupied regions of retinal space, the features of a given object can be linked by virtue of a shared spatial dimension.

But what then—how are objects represented beyond this retinal frame at the level of the scene? According to Treisman, once the features of an object are linked, attention helps construct a limited number of multifeature object representations (e.g., the object files of Kahneman, Treisman, & Gibbs, 1992). Such object representations make it possible to maintain the experience of a unified object across changes in position or physical properties through time. According to feature integration theory, once attention is withdrawn from an object, feature bindings come undone, and the representation of the object disintegrates into its constituent features (Horowitz & Wolfe, 1998; Rensink, 2000, 2002).

This raises a fundamental question about visual short-term memory: Are features maintained as integrated object representations or independently in separate feature maps? Luck and Vogel (Luck & Vogel, 1997; Vogel & Luck, 1997; see also Vogel, Woodman, & Luck, 2001) investigated this question in a series of change detection experiments using visual arrays composed of simple colored shapes. Participants were shown arrays of 1 to 12 items 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). Same/different judgment accuracy sharply declined for arrays containing more than four items, which suggests that visual working memory (VWM) has a limited capacity of approximately three to four items (Cowan, 2001; Irwin & Andrews, 1996;

Sperling, 1960). Surprisingly, when participants viewed stimuli defined by a combination of four different features—color, shape, orientation, and the presence/absence of a gap—with the possibility that any one of these features could vary at testing, participants were just as accurate as when they looked for changes along a single dimension (Irwin & Andrews, 1996). Based on these findings, Luck and Vogel proposed the *integrated object hypothesis*—that individual features are bound into object representations by perceptual processes and that these representations remain integrated in VWM without requiring attentional resources. The capacity limitations then act on the level of bound object representations, not on the level of individual feature values memorized.

What neural processes support the active maintenance of integrated objects in VWM? Empirically, this has been probed using functional neuroimaging as adults perform a standard change detection task. Research shows activation in a distributed network of frontal and posterior cortical regions in this task. In particular, WM representations are actively maintained in the intraparietal sulcus, the dorsolateral prefrontal cortex, the ventral-occipital cortex for color stimuli, and the lateral-occipital complex for shape stimuli (Todd & Marois, 2004, 2005). In addition, there is suppression of the temporoparietal junction during the delay interval in the task, and activation of the anterior cingulate cortex during the comparison phase (Mitchell & Cusack, 2008; Todd, Fougnie, & Marois, 2005). Moreover, there is greater activation of this network on change than on no-change trials, and the hemodynamic response on error trials tends to be less robust (Pessoa, Gutierrez, Bandettini, & Ungerleider, 2002; Pessoa & Ungerleider, 2004).

At a theoretical level, there is currently no unified theory that effectively integrates "what" and "where" in a way that interfaces with these neural and behavioral data. Several neurally plausible models have been proposed that address the integration of "what" and "where" in some way (Deco, Rolls, Horwitz, 2004; Lee, Mumford, Romero, & Lamme, 1998; Van der Veld & de Kamps, 2001). These models are generally quite sophisticated on the "what" side, providing a detailed account of ventral stream processes that, for instance, integrate multiple features together into objects (Deco & Rolls, 2004; Olshausen, Anderson, & Van Essen, 1993). Nevertheless, these models provide a limited view of dorsal stream processes. For instance, several models use the concept of a "salience map" that tags specific locations in space as important for attention or WM (Itti & Koch, 2000; Mozer & Sitton, 1998; Treisman & Gelade, 1980). However, the salience map is not linked to a particular frame of reference, nor is it updated as eyes, head, and body are moved. Conversely, there are detailed models of the spatial aspects of planning sequences of saccades and scanning a visual scene (Dominey & Arbib, 1992; Fix, Rougier, & Alexandre, 2011), but these provide no or only a very rudimentary account of the processing of visual surface features necessary to form a scene representation. Moreover, many of these models have rather limited ties to behavior because they use a biophysical approach to neural function. Here, theoreticians attempt to build neurally realistic models of single neurons (Durstewitz, Seamans, & Sejnowski, 2000; Salinas, 2003), which are then coupled together into populations. Although the biophysical approach has led to new insights into brain function and neural dynamics, these models do an excellent job capturing the behavior of neurons but do poorly at the level of behavior (Finkel, 2000).

In the next section, we provide an overview of the first theory that effectively integrates "what" and "where" to form a WM of integrated objects in a scene (an earlier variant of this model was presented in Johnson, Spencer, and Schöner (2009). We discuss how this model was inspired by the neural literature on object representations. We also demonstrate that the theory effectively captures a suite of behavioral findings from the canonical probe of object representations—the change detection task. Chapter 9 then builds on the concepts introduced here, taking the integrated theory into a real-world, embodied context to demonstrate that the theory not only captures behavioral data with humans but can organize the behaviors of an autonomous robot.

BUILDING A SCENE REPRESENTATION IN DYNAMIC FIELD THEORY

To represent integrated objects, we need to bind the many features that characterize an object together. In principle, this can be done in high-dimensional dynamic fields, with one dimension for each feature value. In Chapter 5, however, we saw how this leads to a combinatorial-explosion problem in which astronomical numbers of neurons would be required to represent any possible combination of feature values. Chapter 5 showed how this problem can be avoided by representing individual feature

dimensions in separate DFs. The separate DFs are then bound across a shared dimension, which in Chapter 5 was retinal visual space.

Chapter 5 also showed how information about an individual object can be selected from a multipeak pattern. Here, we selected information using lower-dimensional fields, in that case, one-dimensional fields. Peaks in these selective attention fields projected ridges into the multidimensional field localized along only one dimension. The intersection of these ridges pulled a spatially aligned pattern of peaks into the attentional foreground. In visual search, this provides a way to bring any object into the attentional foreground that matches the expected feature values. When driven by VWM, this mechanism implemented a form of biased competition to selectively direct attention at objects that match the feature value in working memory.

Although attentional selection was effective at selectively operating on different types of information, there was a critical limit: These processes of selection and integration only work when objects are attended one by one. If multiple objects are brought into the foreground at the same time, then misbindings can occur: It would be unclear, for instance, which feature value was associated with which spatial location. Moreover, using retinal space as a binding dimension was computationally efficient but also fallible, as revealed by illusory conjunctions that may occur, for instance, when spatial overlap and brief stimulus presentation lead to spurious correspondences among different objects.

The more dramatic limitation of using retinal space as a binding dimension occurs when one considers that the eyes make, on average, 170,000 saccades per day. It would obviously not make sense to use retinal space to keep track of the locations of objects, as these would change with every gaze shift. To build a representation of which objects are where, locations instead have to be represented in the space in which objects reside—an allocentric or world-centered frame. Here, spatial positions remain invariant across gaze changes. A scene representation is thus an integrated representation of the visual features of objects grounded in an allocentric frame.

Chapter 7 introduced a neural dynamic mechanism for how information in a retinal frame can be transformed into a body-centered frame. Recall that this mechanism exploits higher-dimensional dynamic fields that combine spatial information in the retinal frame with a representation of the gaze angle relative to the body. Spatial information in a body-centered frame can be projected out from this integrated representation. We also described how this same mechanism can be used to create an object-centered frame anchored to a reference object for a model of spatial language.

What would it take to transform an integrated object representation distributed across multiple space-feature fields from Chapter 5 into a body- or object-centered frame? In the complete case, each field would be minimally three-dimensional (one feature dimension and two retinal spatial dimensions). A five-dimensional transformation field would then be required to associate the two retinal coordinates with the two gaze coordinates, while carrying the feature dimension along for the ride. Unfortunately, however, *every feature dimension would need its own transformation field!* This clearly would be a huge waste of neural resources because the same computation would be done over and over for each feature dimension. In effect, this would undo the gains obtained when we split up feature dimensions into separate neural fields.

Fortunately, the concepts first discussed in Chapter 5 guide us to a solution. Remember that if we select one item at a time, we can extract its spatial position and its feature values into separate, lower-dimensional neural fields without losing any information. Using this idea, we perform the reference frame transformation on a purely spatial representation, such as the retinal spatial attention field used in Chapter 5, to obtain an allocentric spatial representation. We then recombine this transformed spatial information with the feature information of the selected object. As a combined representation for scene memory, we use another stack of space-feature fields, analogous to the retinal level, but now in an allocentric frame of reference. Again, remember that the recombination of space and feature values is possible as long as we treat only a single item at a time. It can then be implemented through the intersection of ridge inputs in a higher-dimensional field, whereas this would lead to misbindings if done for multiple items in parallel. The upside of this approach is the neural computational savings: We do not need a transformation field for each feature dimension—transforming the purely spatial representation is sufficient. The downside is that this form of integration requires that the items to be memorized are attended sequentially, one object at a time.

The integrated DF architecture representing integrated objects in an allocentric frame is shown

in Figure 8.1. To keep the system simple, we have again limited space to a single dimension. The above considerations are nonetheless valid; a scene representation system with full two-dimensional visual space is presented in Chapter 9. Moreover, we consider only the two simple feature dimensions of orientation and color in this architecture, and do not employ any hierarchical system with increasingly complex visual features. This allows us to focus on the integration of feature and spatial information in scene perception, although it limits the possible visual stimuli we can deal with to simple oriented bars. In Figure 8.1, the lavender-shaded region in the lower right of the figure shows the visual attention model from Chapter 5. There are two retinal fields (that correspond to the visual sensory fields in Chapter 5): one defined over retinal space and a color (hue) dimension; the other defined over retinal space and an orientation dimension. These fields project to a one-dimensional spatial attention field, and two one-dimensional feature attention fields. Note that, as in Chapter 5, both retinal fields are coupled to the shared spatial attention field to enable the binding of features into an integrated object representation. Moreover, the attention fields have global inhibition to ensure that only one peak is built at a time.

The rose-shaded region in the top panel of Figure 8.1 shows the transformation field from Chapter 7. This field transforms spatial information in the retinal frame into a body-centered (or world-centered) frame using an estimate of gaze direction from a gaze field. The result is a peak in the allocentric spatial attention field (which, again, has global inhibition). Recall that in Chapter 5, we implemented a specific variant of saccade planning and generation to capture findings from the remote distractor paradigm. In Figure 8.1, we show a "gaze

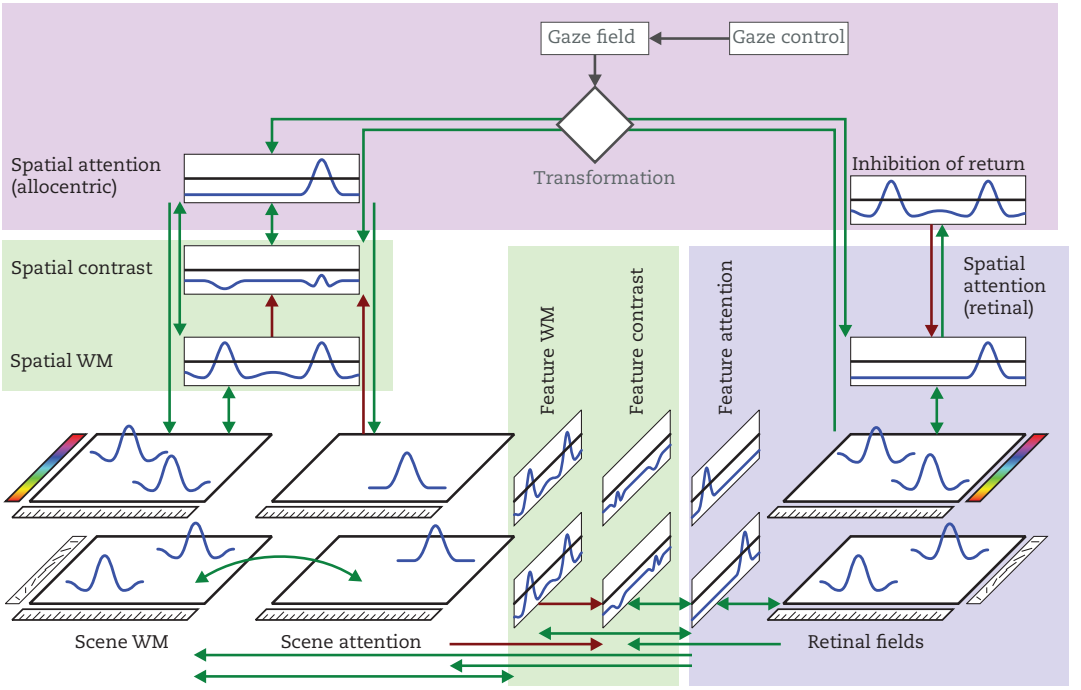

FIGURE 8.1: Dynamic field architecture for scene representation and change detection in multiple feature dimensions (color and orientation). The figure shows the one-dimensional and two-dimensional fields of the architecture with activation peaks and the connections between them. Green arrows indicate excitatory projections, red arrows inhibitory projections. The gray arrows and boxes serve as placeholders for elements not implemented through neural fields. The connectivity along the feature pathways is only shown for the bottom row (representation of orientation), but the same projections are also implemented for the color dimension. Likewise, the connections depicted as vertical arrows connecting fields in the spatial pathway to the two-dimensional integrated fields are implemented equally for all feature dimensions, not only for the color representation. For the two-dimensional fields, arrows ending at one edge of the field create ridge inputs, while the double arrow on the left that extends into the fields indicates a localized projection between the two-dimensional fields.

control" field as a placeholder for these details. Note that we have also included an inhibition-of-return (IOR) field coupled to the retinal spatial attention field. We use this IOR field in the demonstrations and exercises that appear later in the chapter to implement an autonomous version of covert attention. We will return to these details in the next section.

The unshaded region of Figure 8.1 in the lower left shows the scene-level WM and attention fields. The WM representation should be multi-item to enable functional interaction with multiple objects in the world in a way that remains invariant over time as gaze shifts. Moreover, data suggest that this WM must integrate features *and* spatial positions—that is, people robustly remember which objects were where in the scene. Accordingly, the scene-level WM is implemented as a stack of space-feature fields, with different feature dimensions bound through the shared, allocentric spatial dimension. These WM fields receive ridge inputs from the allocentric spatial attention field and the feature attention fields to form localized peaks at the intersection points. In addition, the fields are coupled bidirectionally with the one-dimensional WM fields (discussed later) to form a distributed WM representation over multiple feature spaces. The peaks in the scene-level WM fields are formed sequentially, one item at a time, and then remain self-sustained when attention is shifted to another item or the stimuli are removed.

The scene-level attention fields are used to select one item from the scene WM, for instance, to compare it to a selected perceptual item. They also have a role in indicating when the memorization or comparison for one item is complete and attention should be allowed to shift to the next item. Each scene-level attention field receives localized input from the corresponding scene WM field, with each WM peak inducing a hill of subthreshold activation in the scene attention field. Ridge inputs from the one-dimensional attention fields (which may specify either a spatial location or a feature value, depending on the task) can induce peaks from these localized activation hills and thereby select one WM item. This can be seen in the Figure 8.1, where the system has formed a WM representation of the present stimuli and has currently selected the item on the right both in retinal and in the scene-level attention fields. Like the one-dimensional attention fields, the scene-level attention fields feature global inhibition and allow only a single peak.

Once a scene representation has been created by sequentially attending to items and forming peaks in the scene WM field, it can be used for different tasks. In Chapter 9 we will describe how a scene representation in working memory can be used for planning actions in a robotic context. Here, we will focus on change detection tasks, which have played a prominent role in psychophysical experiments for probing the properties of working memory. Chapter 6 showed how change detection may arise within DFT. Here, change detection was based on comparing a WM representation of metric featural information with current sensory input. This was achieved using a three-layer architecture in which WM peaks inhibited associated sites in a contrast layer via a shared layer of inhibitory interneurons. The contrast field then became active only when current sensory inputs failed to match the contents of working memory.

This principle is implemented in the green-shaded portions of Figure 8.1. Each feature dimension has a feature WM field and a feature contrast field. Similarly, the allocentric spatial pathway has a spatial WM field and a spatial contrast field. We do not employ separate inhibitory fields here as in Chapter 6, but replace them by direct inhibitory projections from the WM fields to the contrast fields in order to limit the complexity of the architecture. Note that the contrast fields receive direct input from the retinal fields as well as input from the attention fields. The retinal connection enables the system to detect changes in spatial or featural information in parallel. As we discuss later in the chapter, this captures key aspects of behavioral data from the change detection task (see Chapter 6). The reciprocal connection between the contrast and attention fields allows the system to direct attention to changes it has detected.

Two other aspects of the green-shaded regions in Figure 8.1 are worth highlighting. First, note that the WM fields have reciprocal connections to the attention fields. These implement the biased competition effects explored in Chapter 5. Second, peaks in the one-dimensional WM fields are built via input from corresponding attention fields, and the WM fields are reciprocally coupled to the scene WM fields. The former connection ensures that peaks are built in WM—both in the one-dimensional WM fields and in the scene WM fields—only when an item is moved into the attentional foreground. The latter connections ensure that the pattern of WM peaks remain consistent between the higher-dimensional and

lower-dimensional fields. In particular, the bidirectional coupling with the spatial WM field ensures that the peaks in the scene WM fields remain spatially aligned, which stabilizes the spatial binding of the different feature values that belong to one item.

The change detection process captured by the three-layer architecture will detect the introduction of a new value along any dimension in parallel (and draw attention to the change). But how do people detect changes when no new features are introduced? For instance, how do people detect that the red cup of coffee seen a few seconds ago is now in a new spatial position (but is the same cup), but the blue cup of tea has been replaced with a blue cup of coffee? This requires comparing the conjunction of features in the current retinal fields with the conjunction of features in the scene WM fields.

One approach to this challenge would be to replicate the three-layer structure at the level of the scene representation, that is, to add a stack of allocentric contrast fields. The problem is that, even with this kind of structure, we could not compare multiple items in parallel, because we would still need to bring the items into the allocentric frame one by one. The alternative is to compare items based solely on the individual feature dimensions, but to augment the mechanism so that the comparison can be focused on individual items. Conceptually, the idea is to bring each item from the current stimulus array into the attentional foreground one at a time and to select an appropriate candidate item for comparison from the scene WM field. This selection takes place in the scene attention field, and depending on the task, it can be based either on the position or on the features of the attended perceptual item. The actual comparison between the perceptual and the memory item and the detection of changes then takes place in the feature contrast fields. These fields receive input from the feature attention fields (excitatory) and the scene attention fields (inhibitory) and will form a peak if a mismatch occurs between these inputs. The feature contrast fields, therefore, play a double role in change detection: They perform both a parallel change detection for multiple feature values (between the multipeak retinal fields and the multipeak WM fields) and a sequential change detection for selected items (between the single-peak feature attention fields and the single-peak scene attention fields).

The sections that follow go through a series of simulations to demonstrate the functionality of the integrated model in different variants of the change detection task. Before proceeding to that discussion, however, it is useful to consider how the architecture in Figure 8.1 was inspired by the neural literature on object representations.

In Chapter 5, we discussed the neural basis for the biased competition architecture (see lavender-shaded region of Figure 8.1). Briefly, the retinal fields capture key aspects of early visual cortical representations (e.g., in V1 and V4), the gaze control system captures aspects of saccade planning and execution in the frontal eye fields and superior colliculus (see Chapter 5), and the spatial and featural projections off the retinal fields mimic properties of visual-selective attention (e.g., in areas of parietal cortex for spatial attention). More generally, the attention fields are the starting point for two clear visual pathways in Figure 8.1—a "dorsal" stream for "where," or spatial, information, and a "ventral" stream for "what," or featural, information.

Moving along the dorsal pathway, the model implements the spatial transformation needed to move from a retinal frame to a body- or world-centered frame. As discussed in Chapter 7, this captures evidence of gain-modulated neurons in area LIP. Continuing along the dorsal pathway into the green-shaded region, we see allocentric representations of space involved in change detection. These fields will mimic aspects of neural activation in the intraparietal sulcus (IPS). For instance, Todd and Marois (2004) reported that IPS activation increased across set sizes in a change detection task as people remembered one to four objects. Critically, the neural signal showed an asymptote beyond four items, indicating a capacity-limited neural representation. The DF model discussed in Chapter 6 shows a similar capacity limit (see Johnson, Spencer, & Schöner, 2009).

The ventral pathway in Figure 8.1 also captures aspects of the neural literature on object representations. As in neural data, this pathway is divided into different neural populations for different feature dimensions. Moreover, the scene-level fields are sensitive to both featural and spatial information, like many neural populations in the ventral pathway, including regions of the inferior temporal cortex. Finally, the WM fields in the ventral pathway will show a capacity-limited increase in neural activation. This has been observed in multiple cortical fields in the ventral pathway in fMRI studies of change detection (see Pessoa & Ungerleider, 2004; Todd & Marois, 2004).

In summary, then, the model in Figure 8.1 captures many aspects of the visual-processing

pathways revealed by neurophysiological and neuroimaging studies. In the sections that follow, we examine whether this same model can also capture behavioral constraints from studies of VWM.

SEQUENTIAL FORMATION OF VISUAL WORKING MEMORY FOR OBJECTS IN A SCENE

In the next sections, we demonstrate the behavioral functionality of the DF model shown in Figure 8.1 using an interactive simulator that implements the model. The simulator is the focus of the exercises for this chapter. You may want to use the simulator as you work through this chapter to illustrate and explore the different functions of the model. We employ one significant simplification for these simulations, in that we assume that gaze direction always remains fixed. This is permissible for the experimental tasks we want to emulate in the model, since these can generally be solved by shifting covert attention instead of making overt eye movements. As a result, the transformation field and gaze control system have been omitted from the architecture, and a one-to-one projection between the retinal and allocentric spatial dimensions is implemented. Figure 8.2 shows a snapshot of the simulator. Note that we have aligned the allocentric spatial attention and contrast fields with the scene attention fields to highlight their spatial correspondence when an object is attended. Similarly, we have aligned the allocentric spatial WM field with the scene WM fields to highlight their spatial correspondence. Keep in mind that all of these fields actually share the same allocentric spatial dimension, but other constraints on the arrangement of fields in the figure do not allow us to present them all aligned with each other.

The buildup of a WM representation for a visual scene is performed sequentially in the model. To this end, one item at a time is selected and a distributed representation of its features and its location is formed. Attentional selection of an item occurs at the retinal level (lavender-shaded parts of model in Figure 8.1). It results from the properties of the one-dimensional attention fields (for space and feature) and their coupling to each other via the retinal fields. In particular, the inputs from the retinal fields vie for attention through competitive interactions in the attention fields. When one item begins to gain strength in one attention field, this influences competition in the other attention fields via the coupling through the retinal fields (this is directly equivalent to the biased competition model detailed in Chapter 5). The result is a consistent selection of the feature values and spatial location of a single item from the visual scene. In Figure 8.2, the left item has been selected, resulting in a single peak in the retinal spatial attention field, the color attention field at the hue value for red, and the orientation attention field at an orientation of 135°. Note that the WM and contrast fields can also play a role in the attentional selection. We will ignore the contrast fields for now and take a detailed look at their function and influence on attention in the next section.

When an object has been selected in the attention fields at the retinal level, activation is projected along the spatial and feature pathways. The retinal spatial attention field projects to the allocentric spatial attention field via the (hypothetical) transformation mechanism that establishes a bidirectional mapping between the two frames of reference. As can be seen in Figure 8.2, a peak has formed in the allocentric spatial attention field on the left. The peaks in the one-dimensional attention fields now drive the formation of WM peaks: The allocentric spatial attention field drives the one-dimensional spatial WM field and induces a peak there. At the same time, each feature attention field induces a peak in its corresponding feature WM field. Now both the one-dimensional attention fields and the one-dimensional WM fields project ridge inputs into the scene WM fields: The spatial fields create vertical ridges, the feature fields create horizontal ridges. The ridges from the attention fields and the corresponding WM fields lie on top of each other (although the WM fields may induce additional, weaker ridges once several items have been encoded in WM). The combination of all four inputs induces an activation peak at the intersection point between the horizontal and vertical ridges in each scene WM field. These peaks provide the integrated representation of the allocentric position and the features of the attended item, bound together via the spatial dimension.

The peaks in the scene WM fields project activation to the scene attention fields. This is a full two-dimensional projection, meaning that it does not induce ridges of activation but localized activation hills in the scene attention field. These activation hills form at the same positions as the peaks in the WM fields. Like the scene WM field, the scene attention field additionally receives ridge inputs from the one-dimensional attention fields (for features and allocentric space). These ridge inputs

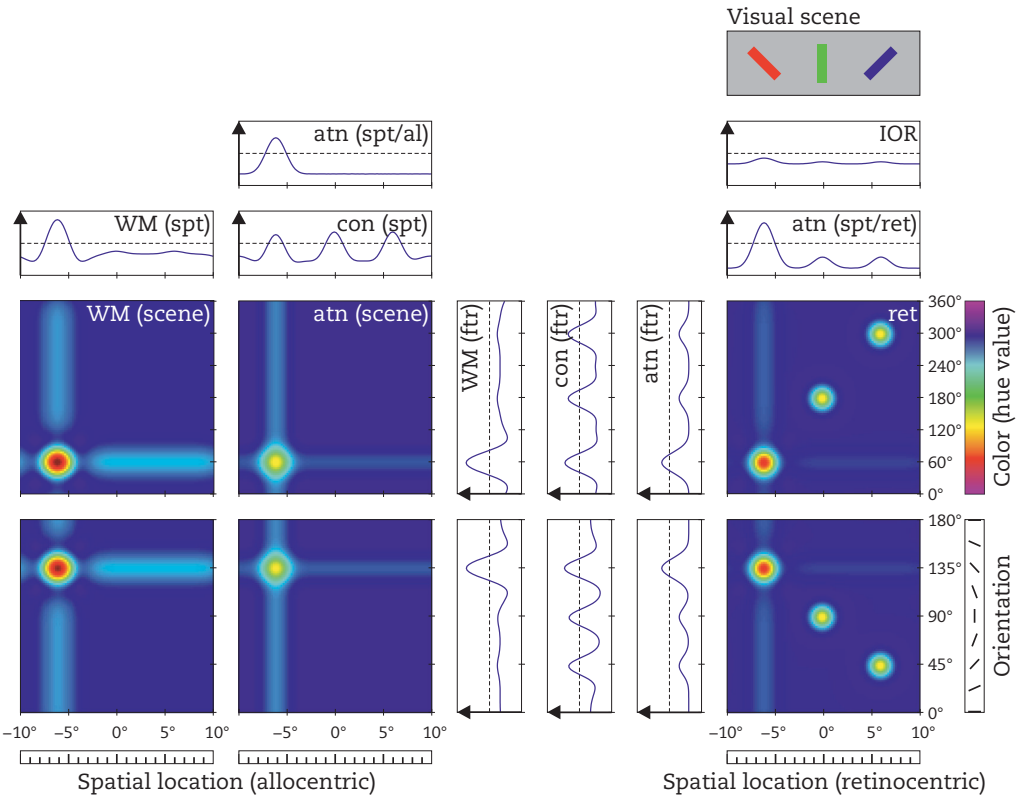

FIGURE 8.2: Activation patterns in the scene representation architecture during memorization of the first item. The fields are arranged analogous to Figure 8.1, only the placement of the allocentric spatial fields (top left) is slightly changed: The allocentric spatial attention field and the contrast field are spatially aligned with the scene attention fields, the spatial WM field is spatially aligned with the scene WM fields. The current visual scene is depicted in the top right; it provides localized inputs to the retinal fields. In the depicted situation, the leftmost item (red diagonal bar) is selected in the retinal fields and one-dimensional attention fields. Feature and spatial information is then transmitted via separate paths, and representations of the item's features and position are formed in the one-dimensional WM fields. They are then combined again in the two-dimensional scene WM fields to memorize the conjunction of features in this specific item. Abbreviations: atn, attention field (scene, feature [ftr], retinal spatial [spt/ret], or allocentric spatial [spt/al]); con, contrast field (feature or spatial [spt]); IOR, inhibition of return field; ret, retinal field; WM, working memory field (feature, spatial, or scene).

alone are not sufficient to induce peaks, but once the additional localized input from the scene WM field arrives at the intersection point of these ridges, the activation in the scene attention field reaches the output threshold and a peak forms.

Since the peak in the scene attention field can only form after the WM representation is established, it can be used as a confirmation signal that the currently attended item has been memorized. It is therefore used in the model to drive a disengagement of attention from the current item, which allows the selection of another item in the scene. This is implemented via a set of dynamic nodes (not shown in the figure). One peak detector node is associated with each scene attention field, which receives globally summed output from the field. These nodes act as binary switches that become active (i.e., produce an output signal) whenever the total output of the corresponding field exceeds a fixed threshold—that is, when a sufficiently strong peak has formed there. These two peak-detector nodes drive a third node, called the condition-of-satisfaction (COS) node. This node becomes activated only when both peak detector nodes are active, thus indicating that the memorization is complete for all features. The COS node

boosts the IOR field, which forms a peak for the currently attended object location. This peak is self-sustained, and it suppresses activation for this spatial position in the spatial attention fields. The effect is a disengagement of attention from the item at that position (which is facilitated by global inhibition from the COS node to the feature and scene attention fields). Moreover, because the IOR peaks are sustained, attention will not be redirected toward previously attended items.

After this sequence of events, one item from the scene is encoded in WM. There is one peak in each scene WM field, and one in each of the feature and spatial WM fields. The peaks are self-sustained through lateral interactions in each of the fields (local excitation and local surround inhibition) and mutual excitation between the fields. Each one-dimensional WM field projects a weak ridge input to the scene WM field and receives a weak input back from it. This coupling keeps the peak positions in all of the WM fields aligned. The two scene WM fields are coupled indirectly via the spatial WM field. There is also a continuous coupling of the one-dimensional WM fields to the retinal scene through weak parallel inputs from the retinal fields. The peaks in the attention fields have disappeared after the activation of the COS node, which in turn causes the COS node itself to return to its resting level. The whole process can now start anew for the next item in the scene, with the sustained peak in the IOR field ensuring that the same item is not attended twice.

Since all of the WM fields support multiple peaks (using only local surround inhibition), additional peaks can form for subsequent items. This is illustrated in Figure 8.3. Here, WM representations have already been formed for two of the items in the scene, and the third item is now attended. New peaks emerge

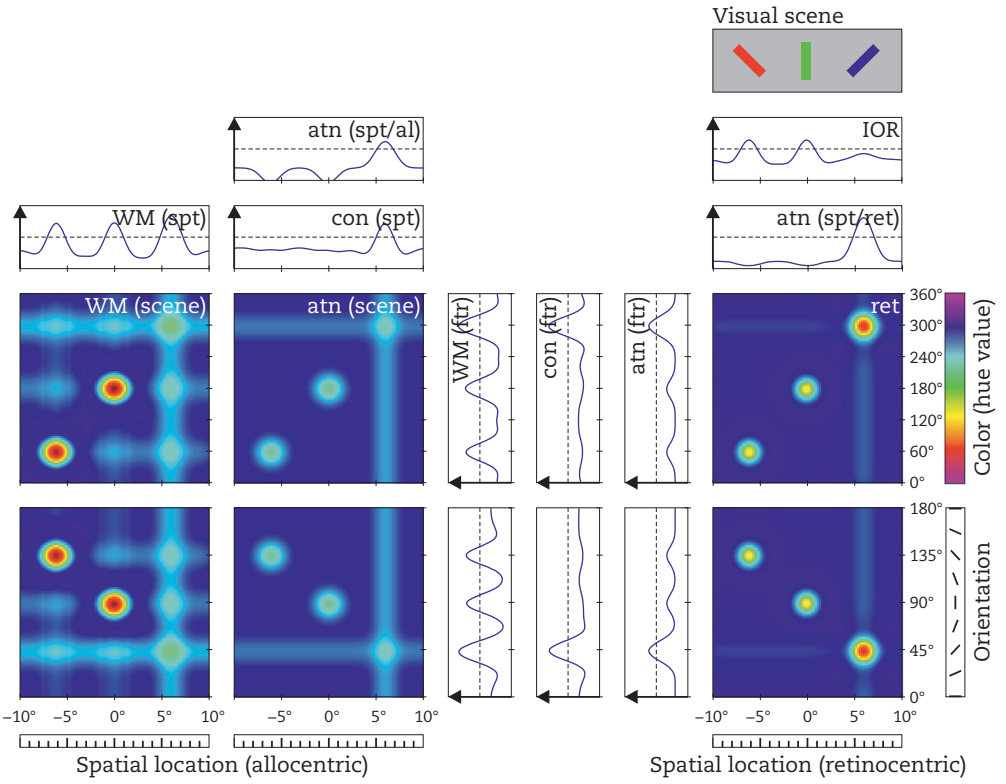

FIGURE 8.3: Adding a third item to a partially formed scene representation in WM. Here, WM peaks have already formed for the left and middle item in the visual scene (see peaks in scene WM matching the retinal field), and now the item on the right is encoded in WM. The item is selected through spatial and feature attention, and peaks form in the one-dimensional WM fields for location and features. The one-dimensional attention and WM fields then project to the scene WM fields. While the input from the spatial and feature WM fields is ambiguous, the single-ridge inputs from the attention fields uniquely determine a position for a new peak in each scene WM field. Abbreviations as in Figure 8.2.

in the one-dimensional WM fields for the attended item due to input from the one-dimensional attention fields. The attention and WM fields together now project to the scene WM fields. Note that in addition to the ridges for the currently attended item (one horizontal and one vertical ridge in each scene WM field), there are additional, weaker ridge inputs from the other peaks in the space and feature WM fields. However, new peaks form in the scene WM fields only from the intersection of the strongest ridges, representing the position and features of the currently attended item. Also note that there are multiple hills of activation in each scene attention field, each reflecting one peak in the corresponding scene WM field. Again, only one of them can turn into an actual peak, the one where the localized input from the scene WM field and ridge inputs from the one-dimensional attention fields come together.

In this fashion, the items are memorized sequentially, and the binding between the individual features of each item can be retained even though they are transmitted via different pathways. This does not mean that the WM representations of the items are isolated from each other. Both in the one-dimensional and two-dimensional WM fields, individual peaks may interact in the same ways as described in Chapter 6. Peaks may repel each other due to lateral inhibition, and in some cases they may merge when particularly close. These interactions are what primarily limits the precision of the memory as well as its total capacity in the model.

The result after a scene with three items has been processed is shown in Figure 8.4. When the stimulus array is turned off, the peaks in the IOR field decay. Unlike WM peaks, they are only sustained as long as they receive some input from the

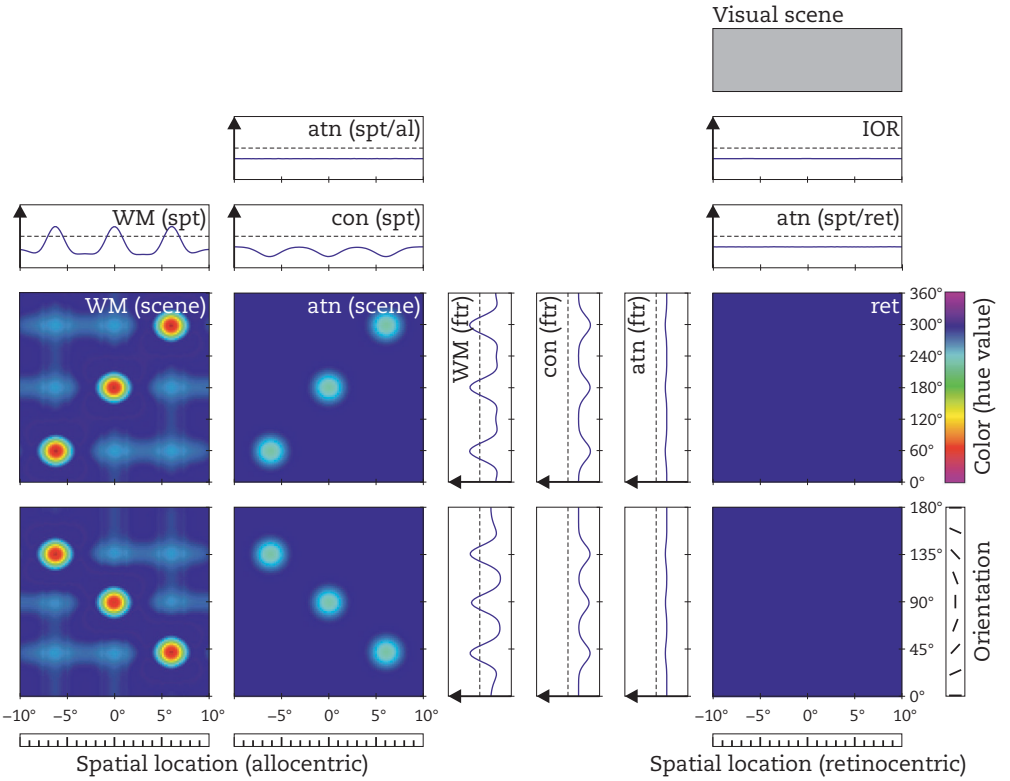

FIGURE 8.4: Scene representation as in Figure 8.3, after the sequential memorization of items is completed and the stimulus array has been turned off. Self-sustained activation peaks for each item are present in the scene WM fields, as well as peaks for the individual features and locations in the one-dimensional WM fields. The peaks in the scene WM fields induce subthreshold hills of activation in the scene attention fields, which will be used in the subsequent tasks to select individual items from WM. The contrast fields show depressions for the memorized feature values and locations, since they only receive inhibitory input from the WM fields. The peaks in the IOR field have decayed after the visual input was turned off, such that the system is ready to attend to the same locations again. Abbreviations as in Figure 8.2.

retinal fields. The model is now ready to process a new stimulus array and compare it to the WM representation. We will use the WM representation shown in Figure 8.4 as the basis for comparison in all tasks described in the next sections.

PARALLEL DETECTION OF FEATURE CHANGES

Change detection tasks are a prominent tool to investigate the properties and capacities of VWM, and they come in many different forms. We have already discussed such tasks in Chapter 6, where we presented a DF model for detecting changes in a single feature or spatial dimension. Here, we extend this discussion to include multiple feature dimensions. In the sections that follow, we extend things further to examine how people detect changes in conjunctions between space and features, and between different feature conjunctions.

The first challenge for the extended change detection architecture is to reproduce what was described in Chapter 6—the detection of changes in a single dimension (either space or feature). For instance, in one of the experiments presented by Treisman and Zhang (2006), subjects were first presented with a sample array of three colored shapes, which they were asked to memorize. After a delay period, the test array was shown. It either contained the same colors and shapes as the first one, or one of the colors or shapes was replaced by a new value not contained in the sample array. Note that in this setting, the locations of the individual items as well as the feature conjunctions are irrelevant for the correct response.

The extended DF model solves this task in a fashion directly analogous to the mechanism explained in Chapter 6. Changes can be detected in parallel, that is, without sequentially attending to each item. The different contrast fields in the model constitute the places where the actual change detection is happening. We have ignored the contrast fields in the previous section since they are not necessary for initial formation of the scene memory. However, they do influence the attentional mechanisms, even during the memorization phase, in a meaningful way. We shall briefly revisit here their function during memorization, which is directly related to the change detection task itself.

Each contrast field—both for surface features and for space—receives direct excitatory input from the corresponding dimension of the retinal fields. As can be seen in Figure 8.2, they immediately form peaks when a new stimulus array is presented (as long as there is no matching representation in the WM fields). These peaks indicate the novel features present in the visual input. Peaks are stabilized by moderate lateral interactions to allow a distinct transition between peak and no-peak states. Lateral inhibition is only local, so multiple peaks can form simultaneously in response to the parallel inputs from the retinal fields.

The main antagonist to this multi-item excitatory input from the retinal fields is an inhibitory input from the one-dimensional WM fields. As these WM fields can likewise have multiple peaks, they can also project inhibition to multiple locations in the contrast fields. The contrast fields then perform a comparison between the memory and perceptual representations in each dimension (features and space), simply through a summation of these inputs: Where the inhibitory memory input and the excitatory retinal input match, they cancel each other out; where the retinal input is not matched by a memory input, it can generate a peak. This is the same principle as in the three-layer model described in Chapter 6, although here it is implemented via a direct inhibitory projection.

The effect of this can be seen when comparing Figures 8.2, 8.3, and 8.4, . In Figure 8.2, at the beginning of the memorization process, there are peaks for all three items in each contrast field. In Figure 8.3, two items have already been memorized. The peaks for these items' feature values in the contrast fields are gone, the activation patterns in those regions are relatively flat (excitatory and inhibitory inputs cancel each other out). The features of the third item, for which memorization is not yet complete, are still considered "novel." Finally, in Figure 8.4, the memorization is complete, and the visual stimulus and the resulting excitatory inputs are gone. As a result, we see the inhibitory troughs in the contrast fields created by active inhibition from the WM peaks. Note that there are other inputs that affect the contrast fields (excitation from the feature/spatial attention fields and inhibition from the scene attention fields). Some effects of these can be seen in Figures 8.2 and 8.3, but they do not qualitatively alter the behavior during memorization and parallel change detection. We will discuss these in detail in the next sections.

The same mechanisms described for the memorization period also take effect when a novel stimulus array—the test array—is presented after the previous one—the sample array—has been memorized. This enables the model to solve the basic change detection task for simple feature changes.

If the memorized features match the features present in the test array, the inhibitory and excitatory inputs to the contrast fields will cancel each other out. If there is a mismatch in one feature, the corresponding contrast field will receive excitation in a region not suppressed by WM input. This can happen either when there is a qualitatively new feature in the display or if there is sufficient quantitative deviation between feature values in the memory and sample array.

The latter case is depicted in Figure 8.5. As in Figure 8.4, three objects have been consolidated in WM (see scene WM fields). The test array presented now is identical to the sample array, with one exception: The color of the leftmost item has been changed from red to orange. Consequently, when the test array is presented, the hue value of this item does not match any of the memorized hue values. This allows the excitatory input to the color contrast field to form a peak immediately after the stimulus presentation. As in Chapter 6, this peak in the contrast field is the basis for change detection in the model.

Note how detection of change in Figure 8.5 does not depend on the binding of feature values to spatial locations or to each other. The comparison only takes place among the values within each individual feature field. This reflects the task requirements, where changes in the features' locations and their conjunctions should be ignored.

Detecting changes by forming peaks for novel features is the key role of the contrast fields, but it is not their only function in the architecture. Each contrast field also projects to the corresponding feature or spatial attention field in an excitatory fashion, thereby guiding attention to novel stimuli. This additional input to the attention fields is not very strong, but it can bias the attentional selection

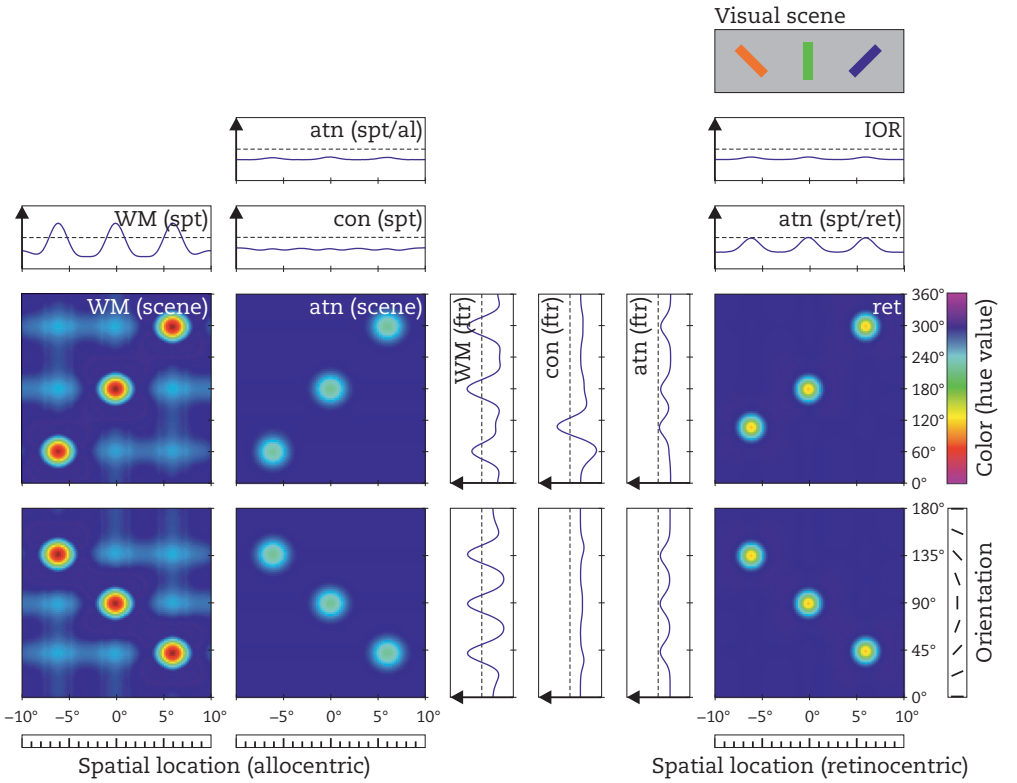

FIGURE 8.5: Detection of a single-feature change. A new array of visual stimuli is presented with one feature changed (the red bar is replaced by an orange bar). The mismatch of memorized and perceived features in the color dimension leads to formation of an activation peak in the color contrast field. The peak forms immediately after the new stimulus array is presented without requiring attentional selection of the item first (parallel detection of feature changes). In the orientation dimension, the present values in the stimulus array still match the memorized orientations, and excitatory and inhibitory inputs cancel each other out in the contrast field. Abbreviations as in Figure 8.2.

toward one item: If several visual items vie for attention, a small additional input to one of the feature attention fields is often sufficient to decide over the outcome, given the reciprocal coupling of all attention fields to each other via the retinal fields.

There is evidence for such an autonomous allocation of attention and parallel feature change detection in the psychophysical literature. Hyun and colleagues (2009) measured the onset of attentional selection during a single-feature change detection task using event-related potentials (ERPs). They found a fast change in spatial allocation of attention to the side where the feature change had occurred. The onset of this change was independent of the number of items in the display, indicating a parallel process. There was also another component in the ERP waveform that was indicative of a second, iterative process. We will describe such processes for other tasks later in the chapter. It is possible that humans tend to employ iterative or sequential processing after a change has been detected, even in situations where they are not necessarily needed.

The attentive reader may have noticed that there is a potential conflict here in the allocation in attention. In Chapter 5, we discussed in detail the biased competition effect for VWM (Hollingworth, Matsukura, & Luck, 2013): Stimuli that match a memorized color are more likely to be selected as targets for timed saccadic eye movements, even when stimulus colors are irrelevant in the saccade tasks. This indicates an autonomous allocation of attention to stimuli that match memorized features. The coupling in Chapter 5 between attention fields and WM fields used in the model of this effect is also present here in the scene representation model. In contrast, the allocation of attention to novel items just described indicates a bias in the opposite direction. It is likewise incorporated in the model, through projections from the contrast fields to the attention fields.

How can these opposite effects be reconciled? While the model cannot give a definite answer, the implementation presented here suggests a possible route to account for both seemingly contradictory experimental results. First, there is difference in the time course between the two effects: The bias toward memorized features is a result of the sustained WM peaks. Their effect on the attention field is present even when there are no visual stimuli, and they start to influence the attentional selection immediately after a new stimulus array is presented. A dominant bias toward memorized features can therefore be expected in tasks where a fast selection decision directly after stimulus onset is encouraged, as is the case in the timed saccade task of the biased competition experiments.

The situation is different for the bias toward novel features, which only emerges after peaks have formed in the contrast fields. This happens quickly after a stimulus with a novel feature value is presented, but still not instantly. Often, by this time, the competition for attention between stimuli has already resulted in a decision for one item. However, if this fast attentional selection is suppressed—for instance, by globally lowering the activation level of the spatial attention field—the influence of the contrast field has time to emerge. This may be appropriate in the change detection task of Hyun and colleagues, where it is likely more efficient for subjects to first take in the whole stimulus array rather than to quickly focus their attention on a single item. In the model, the impact of the contrast fields on attentional selection is stronger than that of the WM fields, so it can dominate the selection process once the contrast field peaks have formed. This would explain the autonomous allocation of attention to items with novel features.

Modulating the global activation values of different fields in the architecture offers a mechanism to significantly alter the behavior of the model and adjust it to different task requirements. The lowering of the activation in the spatial attention field just mentioned basically turns off the sequential attentional processing of items and brings the model into a purely parallel processing mode. In this mode we can even perform a parallel memorization of pure feature values (but not feature conjunctions!), by increasing the resting levels of the feature WM fields. We might also eliminate change detection and the influence of feature novelty on attention by tuning down the contrast fields. This adjustment may further help to explain why no novelty preference was observed in the biased competition experiments.

Experimental evidence for such task-dependent adjustments of attentional mechanisms has been found for the IOR effect (also implemented in our model). This effect can be observed when subjects are required to make a saccade to a location they have recently inspected, for example, during a visual search task. Saccade latencies are then on average longer than for saccades to novel locations (Posner & Cohen, 1984). However, subsequent experiments (Dodd, Van der Stigchel, & Hollingworth, 2009) have found that, depending

on the task requirements, the IOR effect can be replaced by facilitation of return. Inhibition tends to occur in tasks where re-fixating an item is disadvantageous (as in visual search); facilitation occurs when re-fixations may be helpful for the task (e.g., during memorization of a complex scene). While these effects are clearly task-dependent, they cannot really be said to be intentional (the subjects have no direct incentive to make faster or slower saccades). Instead, they likely reflect global adjustments in the neural system in response to a concrete task, which in turn also alters the response to stimuli not part of the task. This matches the possible adjustments in the model by changing global activation levels of specific fields.

CHANGE DETECTION FOR SPACE-FEATURE BINDING

In another variant of the change detection task, participants are asked to detect whether the same features are still present at the same locations (Johnson, Hollingworth, & Luck, 2008). Here, the items in the sample array and in the test array always occupy the same locations. Moreover, the same feature values are present in both displays, but the conjunctions of features and locations may change. In a typical "change" trial, two items' feature values in one feature dimension (e.g., color) are swapped between the sample and test arrays. According to the feature integration theory (Treisman & Gelade, 1980), detecting changes in space-feature bindings requires focused attention on the changed objects. Performance in such tasks is lower than in comparable tasks in which simple feature changes have to be detected (Wheeler & Treisman, 2002), indicating that additional processes are required here.

The DF model can solve this task as well, in a form consistent with the feature integration theory and using the same parameters as those used in the previous task. The encoding of the sample array in the WM fields is performed exactly as before, yielding the result shown in Figure 8.4. Thus, we will focus on the series of events that takes place when the test array is presented, shown in Figure 8.6. If no new feature values are detected (which would generate a novelty peak in the contrast layer and attract attention as described earlier), the system will begin by attending to one randomly selected item in the visual scene, here the one on the left. This takes place through competition in the spatial and feature attention fields, and their coupling to each other via the retinal fields.

Next, the feature and spatial attention fields project ridges into the scene attention fields along the separate pathways. The retinal spatial attention field induces a peak in the allocentric spatial attention field, which then projects a vertical ridge into the scene attention fields. Similarly, the feature attention fields project horizontal ridges into the scene attention fields for the corresponding feature values. In the model, the spatial pathway is overall somewhat dominant over the feature pathways. The peaks in the spatial attention fields will form a little faster and be slightly stronger, and they provide stronger input to the scene attention fields. This is useful in general to achieve a robust binding of the other feature dimensions via space, and it is necessary in particular for the current task.

We use the currently attended spatial location to select a specific WM item in the scene attention field. This is achieved as follows. After a scene has been memorized, there are localized, subthreshold hills of activation in the scene attention field (see Figure 8.4). These are the result of a projection from the scene WM fields, and each of them reflects the combination of spatial position and feature value of one memorized item. The vertical input ridge that now arrives from the allocentric spatial attention field is sufficient to lift one of them to the output threshold and form a peak. (Remember that in this task, the items in the sample and test arrays occupy the same locations, so the ridge input will always coincide with one of the memory peaks). Consequently, in Figure 8.6, the left item selected in the retinal scene is now also selected in the scene attention fields. Once a peak has formed in each of these fields, it suppresses the remainder of the field through global inhibition. The weaker ridge inputs from the feature attention fields therefore have little influence on the selection process in the scene attention fields.

This coupled spatial selection process now allows direct comparison of one selected item in the current scene with the item at the same location in working memory. The feature values of the current visual item have been selected in the feature attention fields. The feature values of the corresponding memorized item can be read out from the scene attention fields. The actual comparison again takes place in the contrast fields for the individual features. We have already described how the contrast fields perform change detection by comparing a retinal and a WM input; this will not produce any peaks in the current scenario, as there are no entirely novel feature values in the test array.

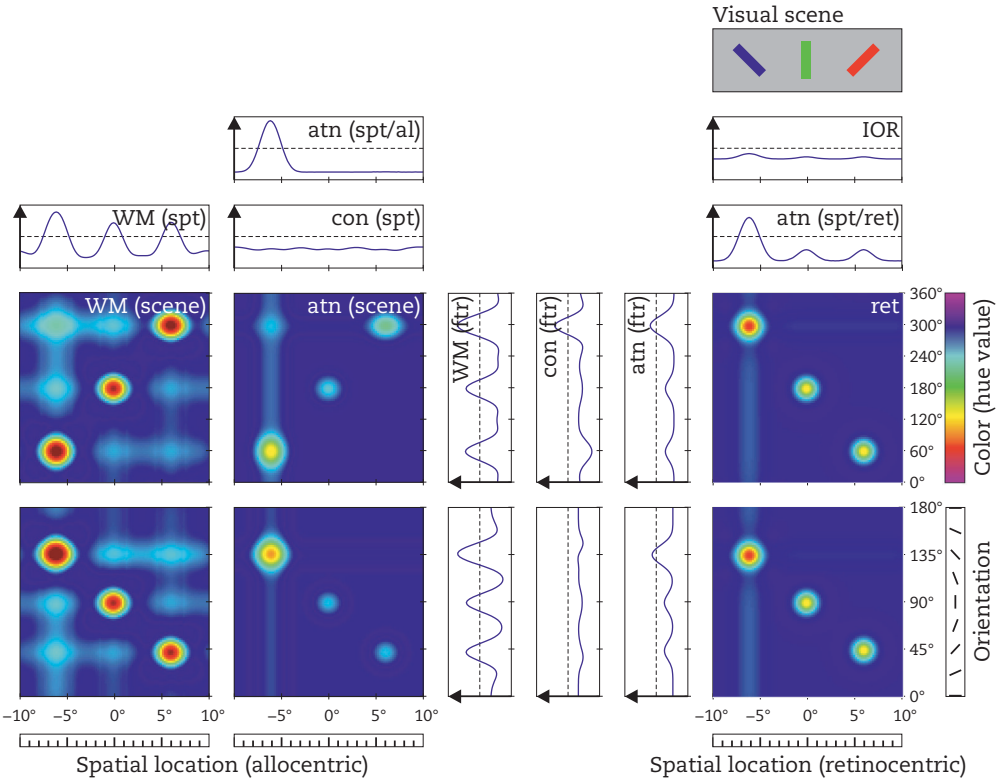

FIGURE 8.6: Detection of changes in feature location. In the stimulus array presented here, the colors of the two outer items have been swapped compared to those in the memorized sample array. Competition in the spatial and feature attention fields has led to attentional selection of the left item. The spatial selection is transmitted via the allocentric spatial attention field to the scene attention fields. Here, the WM item at the same location is selected through spatial ridge inputs, and the feature values of the WM item can be read out. These are compared to the features of the attended visual item in the contrast fields. In the orientation contrast field, the feature values match, and the inputs to the field cancel each other out. In the color contrast field, the excitatory input from the feature attention field is not matched by inhibitory input from the scene attention field, and a peak can form. This indicates detection of a change in feature location. Abbreviations as in Figure 8.2.

However, each contrast field receives an additional pair of inputs: An excitatory input is received from the corresponding feature attention field, and an inhibitory one is received from the scene attention field. These are the basis for change detection in the current task.

The comparison process and its result can be seen in Figure 8.6. Here, the two outer items in the test array have swapped their colors compared to the sample array. The leftmost item in the current array has been selected by the attentional process for the comparison. As seen in the figure, for the orientation dimension the same value is selected in the feature attention field and the scene attention field. The two corresponding inputs cancel each other out in the orientation contrast field, so no peak forms here. In the color dimension, however, there is a mismatch: In the scene attention field, the color of the selected memory item is red, while in the feature attention field there is a peak for blue. Consequently, a peak can form at the hue value for blue in the color contrast field, and a depression is visible at the hue value for red. The system has thus detected the difference in the color dimension.

Note that to actually decide whether two arrays of objects are the same or different, the system must sequentially attend to each item in the test array one at a time and compare it to the corresponding memory item. These sequential transitions in covert attention are driven by the same mechanism as during the memorization of a scene: When sufficiently strong peaks have formed in the scene attention fields, their associated peak detector nodes become active. By this time, the contrast fields will already

have formed a peak if there was any feature mismatch, so we may assume that comparison for the attended item is complete. The peak detector nodes activate the COS node, which boosts the IOR field, and the IOR field suppresses the spatial selection of the current item and prevents it from being selected again. When a feature mismatch is found in any one item, the two scenes are different and the comparison process can be terminated (this is not yet implemented in the model). When all items have been processed without any change being found, we may conclude that the two stimulus arrays are the same.

CHANGE DETECTION FOR FEATURE CONJUNCTIONS

The third type of change detection addressed here deals with feature conjunctions. This can be seen as the laboratory version of a task we face in everyday life: Are two sets of objects the same, irrespective of their location? Imagine, for instance, that you have a few writing tools you typically use for work, like a blue fountain pen, a red ballpoint pen, and a green pencil. Now if someone shows you a set of writing implements lying on a desk and asks whether they are yours, you must compare these items with what you remember. Is there a fountain pen that is blue and has the right size and shape? Is there a red ballpoint pen and a green pencil? Critically, the locations of the items are not informative. It is quite possible the objects have been moved since you last saw them. But the conjunctions of the different features—form, color, size, and so on—must remain the same. Real-world objects don't swap their colors, for instance.

In the laboratory version of this task, participants are again shown two arrays of simple novel objects—a sample array and a test array. Critically, the items in the two arrays may now be spatially scrambled, either switching places or occupying novel locations (Wheeler & Treisman, 2002). The task is to determine whether the feature conjunctions in the test array match the feature conjunctions in the sample array, irrespective of location. Note that although this laboratory task is analogous to the real-world example, there is a key difference—the laboratory task uses completely novel feature combinations. Thus, participants cannot rely on a longer-term memory of the blue pen. Rather, they must quickly build a WM representation for the novel feature conjunctions on the fly and detect changes in these conjunctions a few seconds later. It is quite remarkable that people can do this, given the neural computational constraints discussed previously.

When one thinks about how this task might be solved by participants, it is less clear-cut than the previous variants of the change detection task. Even given the constraints set by our architecture—separate spatial and feature pathways, binding through space for feature conjunctions in working memory—there are several different cognitive strategies that might be used to approach this problem. For instance, one might compare each item in the scene with every WM item. If a match is found for each of them, then test and sample array can be said to be the same. Obviously, this approach would take a lot of time. Alternatively, one might extract the features of an attended item and directly check whether they occur at the same location in working memory. This test is not trivial, however, and would require additional elements in the model architecture.

The strategy we will pursue here to solve this task is the following: We sequentially pick one item in the scene and then select a candidate item for comparison from working memory, based on a feature match. Then we compare these two selected items for differences in their feature values. The assumption is that if there is a matching item in the WM representation, that item will win in the feature-based selection, and the subsequent comparison will yield no differences. If there is no perfectly matching item in working memory, then some imperfect match will be selected as a candidate (e.g., matching only in one feature dimension), and the subsequent comparison will reveal the mismatch. This process can be implemented in the model without adding any new elements.

To carry out this strategy in the model—and, in fact, any of the possible strategies described here—we need to decouple the spatial selection in the retinal and allocentric frames. This reflects the task instruction to ignore the items' locations and will allow us to select items at different locations in the current scene representation and WM representation. It is achieved in the model by inactivating the projections between the retinal and allocentric spatial fields. In a more complete architecture, this might be achieved by de-boosting the activation level of a transformation field that provides the coupling between the two reference frames. To compensate for the resulting loss of inputs for some of the fields, we globally increase the resting level of the allocentric spatial attention field and the scene attention fields. These adjustments—which would be relatively easy to achieve in a biological neural system—are the only changes made to solve the

feature conjunction task. All other connections and parameters in the model remain the same as in the two previous tasks.

The comparison process in this scenario for a "same" trial is illustrated in Figures 8.7 and 8.8. In these figures, the positions of the two outer items have been exchanged between sample and test array, but the feature conjunctions have been retained. As in the previous scenario, the system has to attend to each item in the current scene sequentially. This is again achieved by competition in the coupled one-dimensional attention fields, which leads to selection of the rightmost visual item in Figure 8.7. With the spatial pathway inactivated, only the feature attention fields provide ridge inputs to the scene attention fields. When these ridges overlap with localized inputs from the scene WM fields, they induce (relatively weak) activation peaks. Note that this happens in the two scene attention fields independently. At this early stage of the selection process, there is effectively no coupling between the two fields via the spatial dimension. This coupling only comes about when peaks have formed in the scene attention fields and start projecting to the allocentric spatial attention field.

For a visual item that has a perfect match in the WM representation, the input from all scene attention fields will converge on one position in the spatial attention field. The result is shown in Figure 8.8. The peaks in the two scene attention fields have formed at the same spatial location (albeit a different one than in the retinal fields), and they quickly induce a peak in the allocentric spatial attention field. This field now projects a vertical ridge input back to the scene attention fields and reinforces the existing peaks. In this case, no peaks will form in the contrast fields and thus no change signal is generated: In both feature dimensions, the peak in

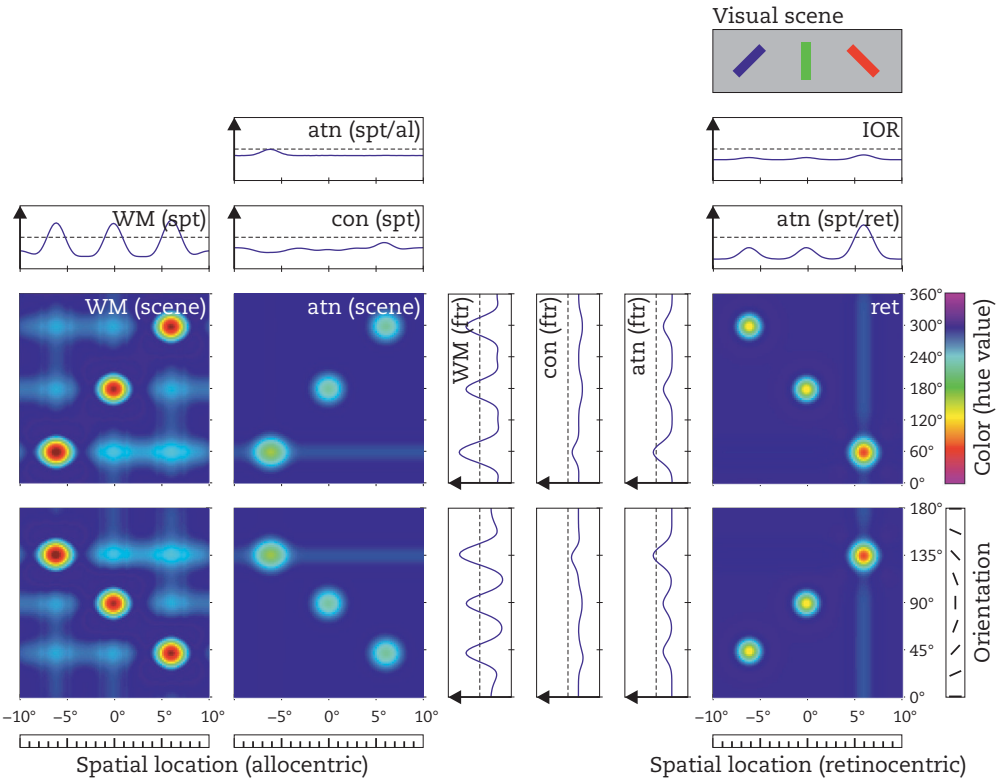

FIGURE 8.7: Detection of feature conjunction changes (early phase of a "same" trial). The test array presented here contains the same items (defined by feature conjunctions) as the sample array, although the locations of the outer items have been swapped. The projections between retinal and allocentric spatial representations have been inactivated for this task. The rightmost item is selected from the retinal field. Its features are projected by the feature attention fields into the scene attention field (as horizontal ridge inputs). They induce weak peaks here, by which the matching item from WM is selected (based on the feature match). The fact that its location has changed has no effect on the selection. Abbreviations as in Figure 8.2.

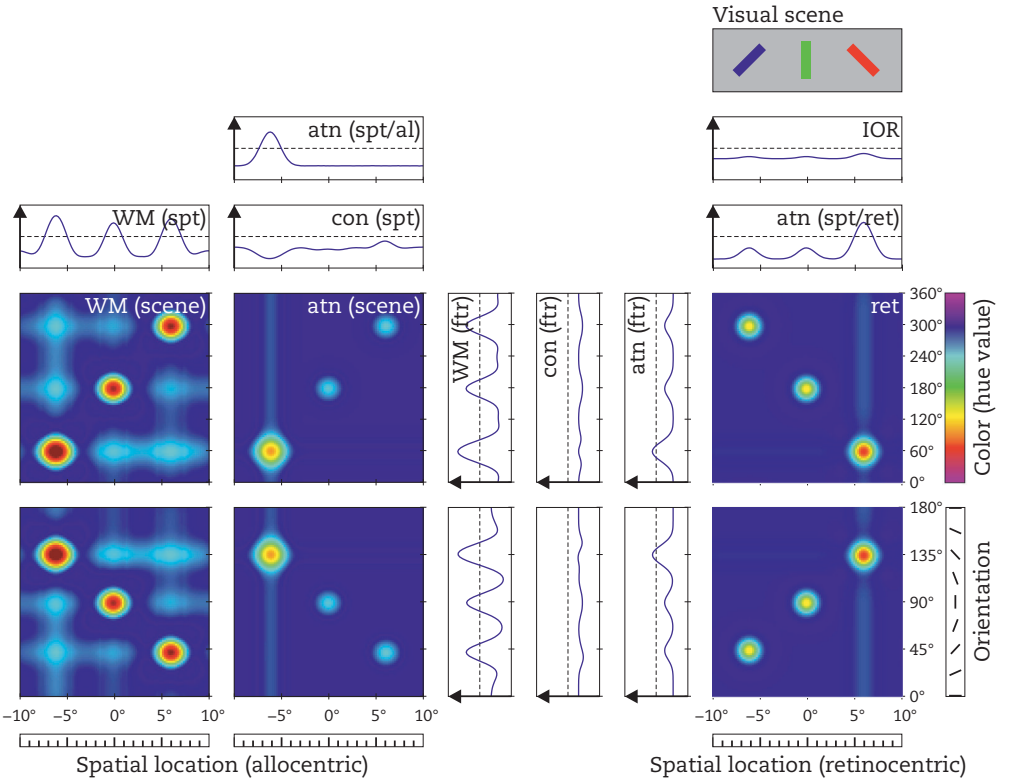

FIGURE 8.8: Detection of feature conjunction changes (late phase of a "same" trial). Activation in the scene attention fields has induced a peak in the allocentric spatial attention field. This peak in turn strengthens the spatially aligned selection in the scene attention fields. Since the selected feature values in the scene attention fields match those in the feature attention fields, no peaks can emerge in the contrast fields. Abbreviations as in Figure 8.2.

the scene attention field matches the peak in the feature attention field, such that the excitatory and inhibitory inputs to the contrast fields cancel each other out. As in the previous scenario, the formation of strong peaks in the two scene attention fields triggers the COS node, which effects the release of attention from the current item and transition to the next one.

Figures 8.9 and 8.10 show the situation for a "different" trial. Here, the colors of the two outer items have been swapped, but the orientations remain the same, so that the feature conjunctions are different between sample and test array. The rightmost item has been selected in the visual scene. This item does not have a perfect match in working memory. Again, the feature attention fields project ridge inputs to the scene attention fields and induce weak activation peaks where these ridges overlap with localized WM inputs (Figure 8.9). These peaks are now at different spatial locations in the two scene attention fields, and they project to different points in the allocentric spatial attention field. So far, no peaks form in the contrast fields, as the peaks in the feature attention fields and scene attention fields are necessarily aligned.

In the next step, a selection process takes place in the allocentric spatial attention field: Under the influence of lateral interactions in the field, an activation peak forms at the location of one of the inputs, while the other one is suppressed (Figure 8.10). The selection is random here, though one could also adjust the system such that one feature dimension is slightly dominant and determines the outcome. The peak in the spatial attention field then again projects back to the scene attention fields. In one of these fields, it will overlap with the existing peak and reinforce it. In the other scene attention field, however, it will not match. Instead, it will overlap with another localized input from the scene WM field. The peak in this scene attention field consequently switches to a new location (compare the scene attention field for color in Figures 8.9 and 8.10). The scene attention fields thereby make the transition from just reflecting the individual

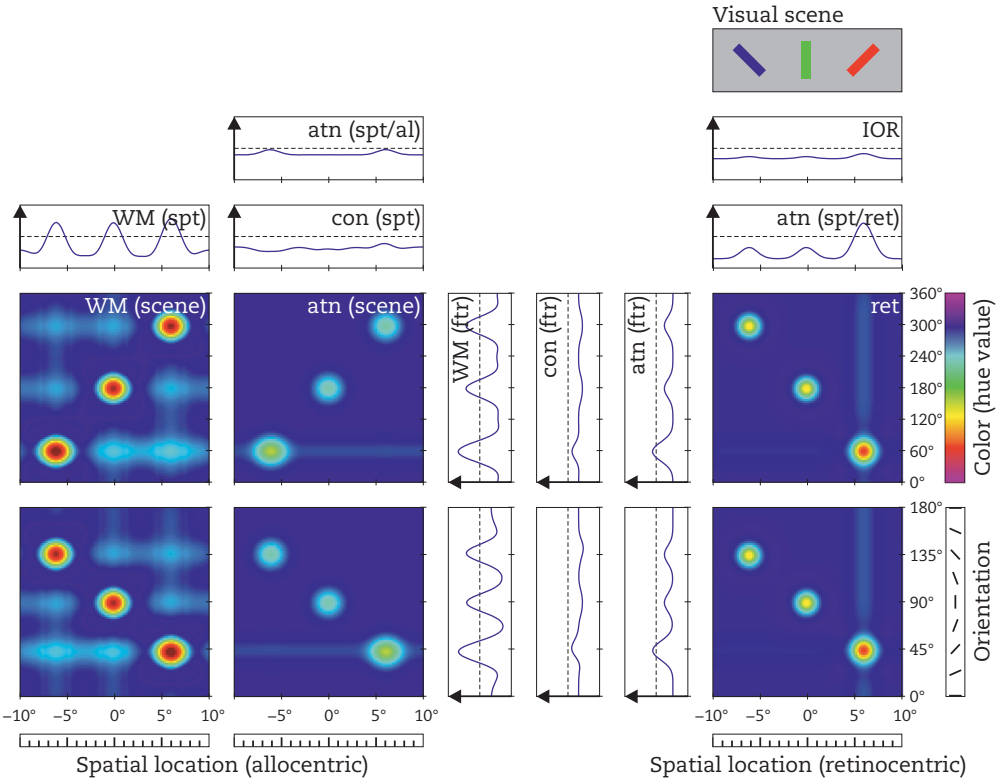

FIGURE 8.9: Detection of feature conjunction changes (early phase of a "different" trial). In the test array presented here, only the colors have been swapped between the outer items, thereby changing the feature conjunctions in the array. The attentional mechanism has again selected the rightmost item from the retinal fields. As before, the feature attention fields project ridge inputs into the scene attention field, in this case inducing peaks that are not spatially aligned. So far, no change is detected, since the weak peaks in the scene attention fields match the peaks in the feature attention fields. Abbreviations as in Figure 8.2.

features of the attended item in the visual scene to reflecting the features of a single, consistent item from working memory, bound together via space. The selected item from working memory matches the features of the attended visual item as much as possible (since it was selected on the basis of these features), but if no perfect match is found, an imperfect one is chosen.

This mismatch between the attended visual item and the selected WM item can now be detected in the contrast fields: In Figure 8.10, the initial peak in the scene attention field for color has been replaced, thus the excitatory and inhibitory inputs to the color contrast field no longer match. A peak can form, indicating that a change in feature conjunctions has been detected between sample and test array. As in the previous task, this process has to be applied sequentially for the items in the scene until a change is found or all items have been processed. The shift of attention from one item to the next occurs autonomously as in the previous task.

The mechanism we employ in this task highlights the central role that the spatial dimension plays in our model. Even though object locations are to be ignored in this task, space still is critical in binding the feature dimensions together. Experimental evidence supports this special role of space in WM representations. Pertzov and Husain (2014) employed a change detection task with sequential presentation of the sample items. Memory performance, particularly with respect to retaining the correct feature conjunctions, was impaired when sample items occupied the same location on the screen. If the items all shared some surface feature, such as color, no analogous decrease in performance was observed. This indicates that object location is used in keeping the memorized

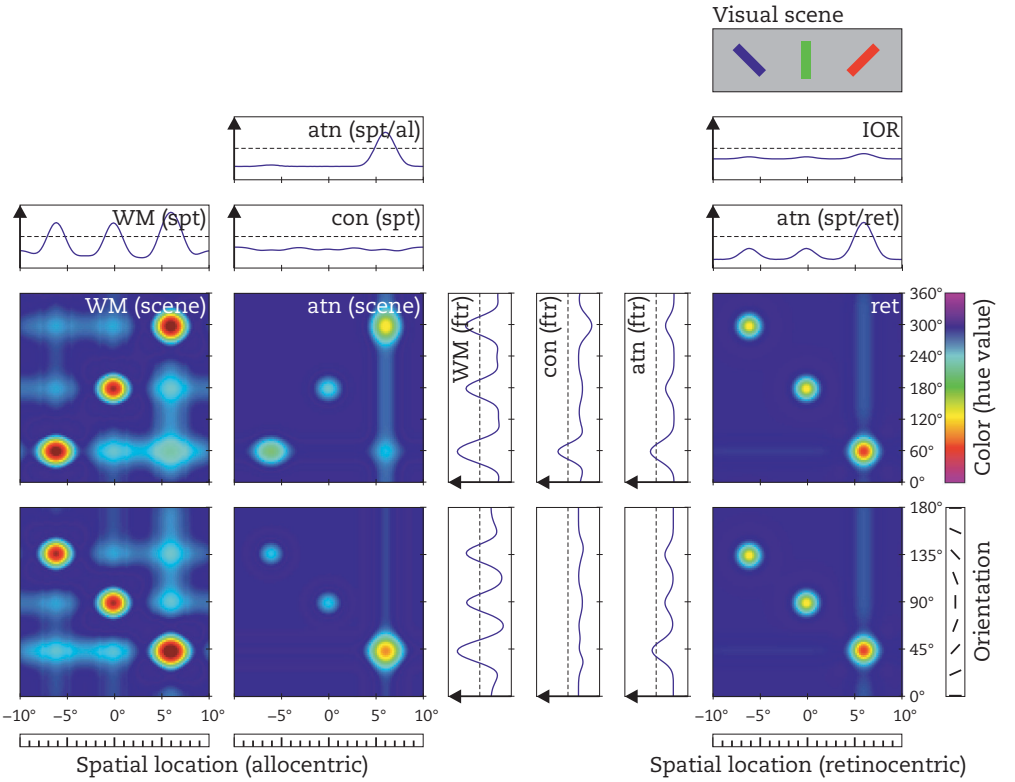

FIGURE 8.10: Detection of feature conjunction changes (late phase of a "different" trial). A peak has formed in the allocentric spatial attention field, selecting one of the competing inputs from the two scene attention fields. This peak projects back to the scene attention fields, strengthening the peak for the orientation dimension but replacing the peak in the color dimension with a new peak. This implements selection of a single consistent WM item based on an (imperfect) feature match. After this has happened, the color contrast field detects the mismatch in the color dimension by forming a peak. Abbreviations as in Figure 8.2.

surface features of each individual object bound together and separate from the features of other objects. These findings are analogous to similar results for the level of visual perception (Nissen, 1985) referred to in Chapter 5.

DESIGNING LARGE DYNAMIC FIELD ARCHITECTURES

The DF architecture for scene representation and change detection presented in this chapter is one the largest, most intricate models covered in this book. It is also, at the time of this writing, quite fresh and still in the process of being tested and refined. For these reasons, we will describe the process of designing and implementing this architecture, and the steps that have lead us to the model in its current form.

As already pointed out by frequent references to previous chapters, this architecture was not designed from a blank slate, but formed as a combination of several previous models—the models presented in Chapters 5, 6, and 7. Among these predecessors we must also count the robotic scene representation architecture presented in Chapter 9. This was, at least in its basic form, already completed before we started work on the change detection model.

The design of a DF model can be structured into three phases(for additional discussion, see Chapter 15): the conceptual planning (what should the model entail, what effects should it produce or explain, and how should these come about?); the design of the architecture (what fields are needed, what is their role in the architecture, and how are they connected?); and, finally, the tuning of parameters to achieve the desired model behavior. Ideally, one would progress through the phases in that order. In practice, however, it may be necessary to

return to an earlier phase when an insurmountable problem is encountered at a later phase.

For the scene representation model, the conceptual planning phase was strongly guided by the existing models. We knew that we wanted to use the biased competition/illusory conjunction architecture as the "front end" of the model, and that the existing mechanism for detecting feature changes should be integrated with it. The goal was then to combine, expand, and, where necessary, adjust these components to create a more general change detection model capable of emulating a larger number of experimental tasks.

A key design decision for this model was the structure of the WM representation. This representation has to fulfill several requirements. It must be able to store, in some form, values of surface features, associations of features to locations, and conjunctions between features (since humans can retain all of these, as is evident from a large number of experiments). This might be achieved in different ways, for instance, by a single high-dimensional field over all feature and spatial dimensions, or by a fixed number of slots for individual memorized items, each with a single one-dimensional field over each spatial and feature dimension. We opted for the stack of space-feature fields, which can be considered a middle ground between these other two options. Several reasons led us to this decision. The stack of space-feature fields mirrors the analogous structure at the retinal level, which in turn is based on well-investigated feature maps in visual cortex. A stack of separate fields requires significantly fewer resources than a single, high-dimensional field over all feature dimensions (as discussed in Chapter 5), and at the same time implements the special role for location in scene memory that is supported by experimental evidence. Finally, it can implement a capacity limit as observed for VWM as a naturally emerging feature (through mutual suppression of peaks), without requiring an inflexible and seemingly artificial definition of WM slots for a fixed number of items.

More generally, one central design decision in the conceptual planning phase is the choice of dimensions over which DFs should be defined. This determines what can be represented in the model and therefore limits what effects can be covered by it. One must also decide whether these dimensions should be covered by actual continuous fields or can be sampled by a few discrete values. In those dimensions that are only sampled by a set of discrete values, no metric effects can be generated in the model. This approach was chosen for the color dimension in the spatial language model (Chapter 7). In that case, feature similarity along the color dimension was not relevant in the covered tasks, and the reduction of the color dimension to three discrete values significantly reduced the computational demands for the simulations. In the scene representation model, we opted for a one-dimensional spatial representation (instead of two-dimensional one) for similar reasons. Conceptually, the model is intended to work in the same way with two spatial dimensions, as demonstrated in the robotic implementation in the next chapter.

Once it is clear what the model should entail and what representations are required, the next step is to design the concrete architecture. One has to consider what fields are needed (and what dimensions every individual field should cover), what the role of each field is in the architecture, and how they must be connected to implement these roles. In specifying the fields and their function, one can focus on a simple classification, based on the bifurcations treated in Chapter 2: Should activation peaks in the field be self-sustained or depend on external input? And should the field support multiple peaks or enforce the selection of a single peak through competition? These two questions are typically sufficient to specify the general behavior of each field in this design stage. When conflicting requirements exist for a representation—for instance, it should have multi-item working memory in one situation, but perform a selection decision in another one—this indicates that at least two separate fields are required (or one has to rethink the requirements). Even if there is no explicit conflict, it can be advisable to separate a representation into multiple fields if it has to fulfill a large number of requirements. This can greatly facilitate the tuning of parameters in the end.

In specifying the fields, one has to think about the sequence of events that should take place in the architecture: When should peaks form in each field, when and where should selection decisions take place, and when may peaks disappear again? To illustrate this, reconsider a somewhat simpler architecture—that of the biased competition model from Chapter 5. We started that model from the assumption that visual stimuli are initially represented in a feature map over space (the two-dimensional visual sensory field). We then used task requirements and empirical results to guide us in what additional components were needed and how they should behave. For instance,

in the biased competition task, we had to consider the color memory cue, which is only presented at the beginning of each trial but affects behavior later. This clearly tells us that there must be some sustained effect of the memory cue, so we added a color memory field that allowed self-sustained activation peaks. The response in the task is a saccadic eye movement, requiring a selection of a spatial location when multiple stimuli are presented. Thus, we added a spatial field with competitive interactions. We then connected the fields so that color and spatial representations can interact via the two-dimensional field, and that led to the basic structure for the model architecture (though two more fields were added to refine the model behavior).

For the scene representation model, specification of the fields and the desired sequence of bifurcations was more complex. We started with the understanding that we would need both a parallel processing of feature values (for the detection of novel features, as in the single-feature change detection model) and a selective, sequential memorization and comparison of individual items (to account for the complexity of feature binding, as discussed earlier). The first sketches of the model had largely independent paths for parallel and selective processing in each feature dimension. We felt that this was unsatisfactory, since it meant that the different tasks would in effect be solved by different, nearly separate systems. A stronger unification was achieved by merging initially separate contrast fields. This resulted in the double role of the contrast fields in the current architecture to detect feature changes for multiple items in parallel and conjunction changes for selected items sequentially.

In designing this large architecture, a complete plan of the sequence of bifurcations (detection, selection, and memory decisions) for all tasks was made before work on the implementation even began. This allowed us to check whether the architecture could work at least in principle and solve the tasks we had selected. For instance, the scene attention field is expected to form a peak in all tasks when it receives a localized input from the scene WM fields and additionally one or two ridge inputs. This setup is consistent. In contrast, if a field has to form a peak from a certain input in one situation, but must not form a peak for the same or a stronger input in another situation, this presents a conflict that might require an adjustment in the architecture. We would note that the rather detailed plan developed at this stage was not fully realized in the final model. For instance, the original plan envisaged that during the memorization phase, a peak would form first in the scene attention field, and this in turn would drive peak formation in the scene WM field. We changed this sequence during parameter tuning when we found that it tended to require excessive mutual excitation between scene WM and scene attention fields (although we are still exploring this variant of the model and its ties to behaviors such as executive function; see Chapter 13).

For the third phase, specification of the model parameters, we implemented the architecture in the COSIVINA framework. The parameters were tuned by hand, a process facilitated by the interactive simulators that make it possible to adjust parameters and immediately see the effects of the change. In choosing the parameters, we were again guided by the classification of fields according to peak stability (input-driven or self-sustained) and mode of interactions (multipeak or competitive). One can find basic parameters for each of these modes from simpler single-field simulations and use these as starting values in the larger architecture. The planned sequence of bifurcations informs us about the required connection strengths between fields: If we want field A to induce a peak in field B, then the input strength must be sufficient to raise the activation in field B above the output threshold. We can do some arithmetic if multiple inputs are combined to form a peak: The input from scene WM field to scene attention field alone must remain below the output threshold, and the same is true for the ridge inputs from the one-dimensional attention fields. Localized input and ridge input together, however, should pierce the output threshold to form peaks.

When connecting fields through mutual connections, it is often necessary to adjust the lateral interaction strengths to compensate for the additional inputs. This is especially true if two fields mutually excite each other. In that case, it can easily happen that activation grows excessively in both fields as soon as they have formed peaks. To compensate, one can increase the lateral inhibition and thereby limit the growth of activation. Because of such effects, the tuning of a model becomes generally more complex with more interconnections, since a change in one field will then affect the behavior of many other fields. That said, these interconnections can also be the source of flexibility as the model is placed in different task contexts.

A particular issue in the scene representation model was the autonomous organization of the sequential processing of items. This involved relatively long sequences of bifurcations that are not driven by any change in the external input but only by the internal dynamics of the model. This adds an additional layer of complexity. During tuning of the model parameters, we first operated the model in a non-autonomous mode: The correct order of bifurcations was created by sequentially boosting fields that were intended to form a peak and de-boosting them once a peak was to be extinguished (this is reproduced in the exercises). This considerably relaxes the requirements for the individual fields, since the precise amount of input each field receives is much less critical in this mode of operation. A field will always form a peak if it is boosted sufficiently, and it is easy to limit the input such that a field will never form a peak without an additional boost. The obvious drawback is that in this mode of operation, the system does not perform any work without constant control inputs from a user.

To obtain autonomous behavior, we had to achieve the same sequence of bifurcations without the boosts. We further tuned the connection strengths between the fields such that inputs would be strong enough to induce peaks only in the desired situations, not in others. The peak detector nodes, COS node, and IOR field were added at this stage (previously, the sequence of items was also induced manually by setting small biasing inputs for different locations). What made tuning more complex in this mode of operation was the behavior of the model in the transition phases. For instance, during the change detection task for feature locations, the contrast fields not only have to show the correct behavior once an item has been selected in both the retinal and the WM representation, they also have to show the right behavior while the selection is still taking place and, in particular, not form a peak prematurely when an item has been selected in the retinal representation but not yet in the WM representation. This requires a sufficient buffer between peak-inducing and non-peak-inducing inputs, such that the right order of bifurcations is retained even when there is some variability in the states of the fields.

There is an alternative approach to creating complex sequences of bifurcations in DF architectures: Rather than removing the boost inputs (or at least most of them) and finely tuning the connection strengths, one may also retain the boosts and add an additional layer to the architecture which autonomously generates the needed sequence of control inputs. In order to achieve a robust autonomous performance, this new control layer has to not only generate the sequence of boost inputs but also check that they have the desired effect before proceeding to the next sequence step. This kind of mechanism will be presented in Chapter 14. We believe that this form of top-down control is appropriate for arbitrary or learned sequences, whereas autonomy from internal interactions is more appropriate for the relatively low-level operations in the scene representation architecture. Tasks like memorizing a scene or detecting changes are performed constantly in everyday life and are not the result of explicit training. It is possible, however, that a mode of operation with more dominant top-down control may be employed in certain situations for the scene representation mechanisms as well. This may, for instance, be a way to improve performance when there is ample time to complete a task.

CONCLUSIONS AND OUTLOOK

The goal of this chapter was to build an integrated neural system that could construct a VWM for novel objects in a scene such that the system could remember which object was where and detect changes in those objects after short delays. We accomplished this goal, presenting the first integrated theory of VWM for objects in an allocentric frame of reference. This DF model was inspired by neurophysiological studies of non-human primates and neuroimaging studies of human adults. Moreover, we demonstrated that the theory can capture behavioral findings from different variants of the canonical task used to probe VWM—the change detection task. To date, no other formal theory has captured data from all of these variants within a single neural system.

Importantly, the theory we developed built on innovations described in Chapters 5–7. This shows that DFT can scale up from simpler systems to a larger-scale, integrated neural architecture. This is an important proof of concept. Our sense is that models are often treated in isolation. This is unfortunate, because it can lead to a proliferation of disconnected accounts, when the promise of formal theories is integration across phenomena. Here we have not only brought together three variants of the change detection task but have also embedded this account in a neural system that captures the

details of biased competition effects in saccade orienting (Chapter 5), illusory conjunctions (Chapter 5), spatial recall and position discrimination (Chapter 6), reference-frame alignment and the characteristics of gain-modulated neurons (Chapter 7), and—at least in principle—the ingredients for spatial language (Chapter 7). In this final case, work remains to clarify precisely how the spatial language model from Lipinski, Schneegans, Sandamirskaya, Spencer, and Schöner (2012) can be realized in the integrated DF model presented here. At face value, our sense is that peaks in the scene attention fields operate much like the target field in the spatial language model, bringing the target into the foreground, while peaks in the scene WM fields operate like an object-based frame of reference.

Consideration of the spatial language model also points toward another key issue we are poised to tackle with the integrated DF model: We have the potential to explain not only how people use spatial language to refer to target and reference objects, but also how people remember the details of these object-based scenes. In particular, by adding in a memory trace to the scene-level fields, we can establish a long-term memory for visual scenes in addition to the short-term memory. Indeed, if we were to store multiple "copies" of the memory trace pattern—one for each "context"—we could flexibly reinstantiate these scene-level patterns in a context-dependent manner. Perhaps this could explain the finding that people can detect often subtle changes in the details of objects in complex scenes after viewing more than 400 total objects—up to 24 hours later (Hollingworth, 2005).

This chapter also re-emphasizes a point initially raised in Chapter 5—that cognition often occurs via a sequence of bifurcations, with the formation of one peak (or peaks) causing a cascade of other neural events. We will return to this notion in Chapter 14, when we introduce behavioral sequence generation. Next, however, we continue the arc started in this chapter. Chapter 9 instantiates an integrated visual cognitive architecture in an autonomous robot. This fully implements the real-world autonomy captured in a cursory way here using the IOR field (see exercises). Moreover, Chapter 9 highlights new types of functionality that emerge when the integrated model is placed in the real world within an autonomous agent—functionality that extends well beyond the change detection setting probed here.

REFERENCES

Aggelopoulos, N. C., & Rolls, E. T. (2005). Scene perception: inferior temporal cortex neurons encode the positions of different objects in the scene. *European Journal of Neuroscience*, *22*, 2903–2916.

Andersen, R. A. (1995). Encoding of intention and spatial location in the posterior parietal cortex. *Cerebral Cortex*, *5*, 457–469.

Baker, M., & Olson, C. R. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nature Neuroscience*, *5*(11), 1210–1216.

Banks, M. S., & Shannon, E. (1993). Spatial and chromatic visual efficiency in human neonates. In C. E. Granrud (Ed.), *Visual perception and cognition in infancy* (pp. 1–46). Hillsdale, NJ: Erlbaum.

Bornstein, M. H., Krinsky, S. J., & Benasich, A. A. (1986). Fine orientation discrimination and shape constancy in young infants. *Journal of Experimental Child Psychology*, *41*(1), 49–60.

Bridgeman, B., Gemmer, A., Forsman, T., & Huemer, V. (2000). Processing spatial information in the sensorimotor branch of the visual system. *Vision Research*, *40*, 3539–3552.

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

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, *33*, 25–62.

Darling, W. G., & Miller, G. F. (1993). Transformations between visual and kinesthetic coordinate systems in reaches to remembered object locations and orientations. *Experimental Brain Research*, *93*, 534–547.

Deco, G., & Rolls, E. T. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, *44*(6), 621–642.

Deco, G., Rolls, E. T., & Horwitz, B. (2004). "What" and "where" in visual working memory: A computational neurodynamical perspective for integrating fMRI and single-neuron data. *Journal of Cognitive Neuroscience*, *16*, 683–701.

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, *4*, 2051–2062.

Desimone, R., & Gross, C. G. (1979). Visual areas in the temporal cortex of the macaque. *Brain Research*, *178*, 363–380.

DiCarlo, J. J., & Maunsell, J. H. (2003). Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *Journal of Neurophysiology*, *89*, 3264–3278.

Dodd, M. D., Van der Stigchel, S., & Hollingworth, A. (2009). Novelty is not always the best

policy: Inhibition of return and facilitation of return as a function of visual task. *Psychological Science, 20*(3), 333–339.

Dominey, P. F., & Arbib, M. A. (1992). A cortico-subcortical model for generation of spatially accurate sequential saccades. *Cerebral Cortex, 2*(2), 153–175.

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature Neuroscience, 3*, 1184–1191.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*, 1–47.

Finkel, L. H. (2000). Neuroengineering models of brain disease. *Annual Review of Biomedical Engineering, 02*, 577–606.

Fix, J., Rougier, N., & Alexandre, F. (2011). A dynamic neural field approach to the covert and overt deployment of spatial attention. *Cognitive Computation, 3*(1), 279–293.

Fujita, I., Tanaka, K., Ito, M., & Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature, 360*, 343–346.

Gross, C. G., Rocha-Miranda, C. E., & Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex. *Journal of Neurophysiology, 35*, 96–111.

Hollingworth, A. (2005). The relationship between online visual representation of a scene and long-term scene memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(3), 396–411.

Hollingworth, A. (2006). Scene and position specificity in visual memory for objects. *Journal of Experimental Psychology: Learning, Memory & Cognition, 32*, 58–69.

Hollingworth, A. (2007). Object-position binding in visual memory for natural scenes and object arrays. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 31–47.

Hollingworth, A., Matsukura, M., & Luck, S. J. (2013). Visual working memory modulates rapid eye movements to simple onset targets. *Psychological Science, 24*(5), 790–796.

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature, 394*(6693), 575–577.

Huttenlocher, J., Hedges, L., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review, 98*, 352–376.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General, 129*, 220–241.

Hyun, J. S., Woodman, G. F., Vogel, E. K., Hollingworth, A., & Luck, S. J. (2009). The comparison of visual working memory representations with perceptual inputs. *Journal of Experimental Psychology: Human Perception and Performance, 35*(4), 1140.

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* (pp. 125–155). Cambridge, MA: MIT Press.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*, 1489–1506.

Jackendoff, R. (1996). The architecture of the linguistic-spatial interface. In P. Bloom, et al. (Eds.), *Language and space*. Cambridge, MA: MIT Press.

Johnson, J. S., Hollingworth, A., & Luck, S. J. (2008). The role of attention in binding features in visual short-term memory. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 41–55.

Johnson, J. S., Spencer, J. P., & Schöner, G. (2009). A layered neural architecture for the consolidation, maintenance, and updating of representations in visual working memory. *Brain Research, 1299*, 17–32.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology, 24*, 175–219.

Komatsu, H., & Ideura, Y. (1993). Relationship between color, shape, and pattern selectivities in the inferior cortex of the monkey. *Journal of Neurophysiology, 70*, 677–694.

Kravitz, D. J., Vinson, L. D., & Baker, C. I. (2008). How position dependent is visual object recognition?. *Trends in Cognitive Sciences, 12*(3), 114–122.

Lee, T. S., Mumford, D., Romero, R., & Lamme, V. A. (1998). The role of primary visual cortex in higher level vision. *Vision Research, 38*, 2429–2454.

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory & Cognition, 38*(6), 1490–1511.

Livingstone, M. S., & Hubel, D. H. (1988). Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science, 240*, 740–749.

Llinás, R., & Paré, D. (1996). The brain as a closed system modulated by the senses. In R. Llinás & P. S. Churchland (Eds.), *The mind-brain continuum*. Cambridge, MA: MIT Press.

Luck, S. J., Girelli, M., McDermott, M. T., & Ford, M. A. (1997). Bridging the gap between monkey neurophysiology and human perception: An ambiguity resolution theory of visual selective attention. *Cognitive Psychology, 33*, 64–87.

Luck, S. J., & Vecera, S. P. (2002). Attention. In S. Yantis (Ed.), *Stevens' handbook of experimental*

psychology: Sensation and perception (Vol. 1, pp. 235–286). New York: Wiley.

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

McNamara, T. P., Halpin, J. A., & Hardy, J. K. (1992). Spatial and temporal contributions to the structure of spatial memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 18*, 555–564.

Mitchell, D. J., & Cusack, R. (2008). Flexible, capacity-limited activity of posterior parietal cortex in perceptual as well as visual short-term memory tasks. *Cerebral Cortex, 18*(8), 1788–1798.

Moore, C. M., Mordkoff, J. T., & Enns, J. T. (2007). The path of least persistence: Evidence of object-mediated visual updating. *Vision Research, 47*, 1624–1630.

Mozer, M. C., & Sitton, M. (1998). Computational modeling of spatial attention. In H. Pashler, H. (Ed.), *Attention* (pp. 341–393). New York: Psychology Press.

Nissen, M. J. (1985). Accessing features and objects: Is location special? In M. I. Posner & O. S. M. Marin (Eds.), *Attention and performance Xl* (pp. 205–219). Hillsdale, NJ: Erlbaum.

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience, 13*, 400–419.

Op De Beeck, H., & Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. *Journal of Comparative Neurology, 426*, 505–518.

Pashler, H. (1988) Familiarity and visual change detection. *Perception & Psychophysics, 44*(4), 369–378.

Pertzov, Y., & Husain, M. (2014) The privileged role of location in visual working memory. *Attention, Perception, & Psychophysics, 76*(7), 1914–1924.

Pessoa, L. Gutierrez, E., Bandettini, P. A., & Ungerleider, L. G. (2002). Neural correlates of visual working memory: fMRI amplitude predicts task performance. *Neuron, 35*(5), 975–987.

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

Pick, H. L., Montello, D. R., & Somerville, S. C. (1988). Landmarks and the coordination and integration of spatial information. *British Journal of Developmental Psychology, 6*, 372–375.

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. Attention and performance X: Control of language processes, *32*, 531–556.

Rao, S. C., Rainer, G., & Miller, E. K. (1997) Integration of what and where in the primate prefrontal cortex. *Science, 276*, 821–824.

Rensink, R. A. (2000) The dynamic representation of scenes. *Visual Cognition, 7*, 17.

Rensink, R. A. (2002) Change detection. *Annual Review of Psychology, 53*, 245–277.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8*(5), 368–373.

Salinas, E. (2003). Background synaptic activity as a switch between dynamical states in a network. *Neural Computation, 15*(7), 1439–1475.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review, 5*, 644–649.

Soechting, J. F., & Flanders, M. (1989) Errors in pointing are due to approximations in sensorimotor transformations. *Journal of Neurophysiology, 62*(2), 595–608.

Spencer, J. P., & Hund, A. M. (2002) Prototypes and particulars: Geometric and experience-dependent spatial categories. *Journal of Experimental Psychology: General, 131*, 16–37.

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

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience, 19*, 109–139.

Tanaka, K. (2000). Mechanisms of visual object recognition studied in monkeys. *Spatial Vision, 13*, 147–163.

Teller, D. Y., & Bornstein, M. H. (1987). Infant color vision and color perception. *Handbook of Infant Perception, 1*, 185–236.

Todd, J. J., Fougnie, D., & Marois, R. (2005). Visual short-term memory load suppresses temporo-parietal junction activity and induces inattentional blindness. *Psychological Science, 16*(12), 965–972.

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

Todd, J. J., & Marois, R. (2005). Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity. *Cognitive, Affective, & Behavioral Neuroscience, 5*(2), 144–155.

Tootell, R. B., Hadjikhani, N., Hall, E. K., Marrett, S., Vanduffel, W., Vaughan, J. T., & Dale, A. M. (1998). The retinotopy of visual spatial attention. *Neuron, 21*, 1409–1422.

Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology, 6*, 171–178.

Treisman, A. (1999). Solutions to the binding problem: Progress through controversy and convergence. *Neuron, 24*(1), 105–110.

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

Treisman, A., & Zhang, W. (2006). Location and binding in visual working memory. *Memory & Cognition, 34*(8), 1704–1719.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549–586). Cambridge, MA: MIT Press.

Van der Veld, F., & de Kamps, M. (2001). From knowing what to knowing where: Modleing object-based attention with feedback disinhibition of activation. *Journal of Cognitive Neuroscience, 13*, 479–491.

Vogel, E. K., & Luck, S. J. (1997). ERP evidence for a general-purpose visual discrimination mechanism. *Society for Neuroscience Abstracts, 23*, 1589.

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.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 1436–1451.

von der Malsburg, C. (1996). The binding problem of neural networks. In R. Llinás & P. S. Churchland (Eds.), *The mind-brain continuum* (pp. 131–146). Cambridge, MA: MIT Press.

Wheeler, M., & Treisman, A. M. (2002), Binding in short-term visual memory. *Journal of Experimental Psychology: General, 131*, 48–64.54.

Wolfe, J. M. (1998). Visual search. In H. Pashler (Ed.), *Attention* (pp. 13–73). London: University College London Press.

EXERCISES OF CHAPTER 8

Start the simulator for this exercise by running the file `launcherSceneRepresentation`. The graphical user interface (GUI) shows the fields of the scene representation architecture in the same layout as that used in the figures throughout this chapter. In the control area at the bottom of the GUI window you will find sliders to boost or de-boost all fields, buttons to activate or deactivate input patterns, and sliders to give an extra input to one location in the retinal spatial attention field. These latter sliders can be used to bias the system toward the attentional selection of a specific stimulus.

Exercise 1: Sequential Memorization of Items in a Scene

a) Run the simulator and activate stimulus pattern A. Observe the sequence of peak formations as the WM representation is built up. Use the `Pause` button to slow down the simulation when necessary, and use `Reset` to view the process repeatedly.

b) Now use the sliders to de-boost all fields to −5 (leave the three spatial input sliders a_{s1} to a_{s3} in the lower right at zero!). This will prevent formation of peaks in most of the fields (except for the contrast fields, but you can ignore them for now). Reset the simulation (click the `Reset` button) and activate one of the stimulus patterns again. Now manually create the sequence of peaks for the memorization of one item (as observed before), by setting the boost values of individual fields back to zero in the correct order. (Note: Click on the slider bars to the left or right of the slider to change the boost value by steps of 1.)

c) It is explained in the text that the formation of peaks in the scene attention fields is used as signal that the memorization of an item is complete and that attention should be shifted to the next item (via the peak detector nodes, COS node, and IOR field). Why does the system not use peak detectors in the scene WM field directly to detect when an item has been memorized?

Exercise 2: Parallel Detection of Feature Changes

a) Reset the simulation again and activate stimulus pattern A. Now observe the activation pattern in the contrast fields while the WM representation is built up. What do the peaks in these fields indicate during the memorization phase?

b) Form only a partial representation of the stimulus array in WM by turning off the stimulus pattern once one or two items are memorized. Now turn the same stimulus pattern on again. What can you observe in the contrast fields?

c) After the whole pattern is memorized again, modify the stimulus pattern by changing the feature value of one stimulus. Turn the stimulus off first, then open the parameter panel. Select the element "i1 for vis_f1" (scroll down almost to the end of the drop-down list), and change the parameter `positionY` to 30. Now turn the stimulus pattern on again, and observe how the change is detected through a peak in the feature contrast field. (You should then

turn the changed stimulus pattern off again before the WM representation is updated so you can use that for the subsequent exercises).

Exercise 3: Change Detection for Space-Feature Bindings

a) After pattern A has been memorized, deactivate it, wait a moment for the IOR peaks to decay, then activate pattern B or C. Observe how the model performs the sequential change detection for feature locations by forming peaks in the feature contrast fields when the location of a visual feature has changed. Note where the difference lies between "same" and "different" items.

b) Turn off the stimulus pattern and wait for the peaks in the IOR field to decay. De-boost all attention and contrast fields to −5 using the sliders (leave the WM fields untouched, otherwise the WM representation will be lost). Now reactivate the stimulus pattern and set the boost value of the fields back to zero in the right order to create the feature location change detection for one item. (You may use the spatial input sliders a_{S1} to a_{S3} to bias the attentional selection to a certain item. A small input like 0.2 is typically sufficient to achieve that.)

c) Why is it important for this task that the spatial input to the scene attention field is stronger or arrives earlier than the feature input?

Exercise 4: Change detection for feature conjunctions

a) To perform this task, first have the model memorize pattern A (this should still be present from the previous exercises), then deactivate the spatial coupling between the retinal and allocentric reference frame (click the corresponding button in the bottom center of the GUI to switch between active and inactive coupling). By deactivating the spatial coupling, the system is no longer sensitive for changes in feature location and instead detects changes in feature conjunctions independent of location. Now activate either pattern B ("same" for this task) or pattern C ("different"). Observe the sequence of events in the model, and note the differences between "same" and "different" items.

b) Once more, reproduce the sequence of events for the comparison of one item manually by de-boosting all attention and contrast fields, then boosting them again in the right order.

c) You may notice that the formation of peaks in the scene attention field happens in two phases in this task: first relatively weak peaks form, then stronger ones after the allocentric spatial attention field projects additional input into the field. What happens when you give the scene attention field an additional positive boost when it first forms peaks, so that strong peaks form immediately? Why does the detection of feature conjunction changes not work under these conditions?