Dynamic Field Theory of Visuospatial Cognition

Dissertation zur Erlangung des Grades eines
Doktor-Ingenieurs der Fakultät für Elektrotechnik und
Informationstechnik an der Ruhr-Universität Bochum

vorgelegt von

Sebastian Schneegans

geboren in Duderstadt

Bochum, Dezember 2014

Kurzfassung in deutscher Sprache

Ein wesentliches Merkmal der menschlichen visuellen Wahrnehmung ist das aktive Sehen, also das Erfassen der visuellen Umwelt durch Abfolgen zielgerichteter Augenbewegungen. Die Steuerung dieses Prozesses und die Interpretation der so gewonnenen sensorischen Daten stellt spezifische Anforderungen an das Nervensystem. Wesentliche Signaturen des aktiven Sehens, insbesondere die sequentielle Fokussierung der Aufmerksamkeit auf einzelne Elemente, zeigen sich zudem auch in kognitiven Prozessen.

In dieser Arbeit stelle ich eine Reihe von neurodynamischen Modellen vor, die aufeinander aufbauend verschiedene Aspekte des aktiven Sehens behandeln. Zunächst untersuche ich die neuronale Steuerung zielgerichteter Augenbewegungen sowie die Verarbeitung der visuellen Information zur Erzeugung blickrichtungsinvarianter räumlicher Repräsentationen. Dieselben neuronalen Mechanismen nutze ich zur Entwicklung kognitiver Modelle des menschlichen Arbeitsgedächtnisses für visuelle Szenen sowie der Verankerung räumlicher Sprache in der visuellen Wahrnehmung.

Die theoretische Grundlage für diese Modelle ist die Theorie Dynamischer Neuronaler Felder. Diese Theorie beschreibt neuronale Prozesse auf einer mittleren Abstraktionsebene, die einerseits die grundlegenden Funktionsweisen biologischer neuronaler Systeme widerspiegelt, andererseits aber auch eine direkte Verbindung zu beobachtbarem Verhalten herstellt. Dazu wird die zeitlich kontinuierliche Entwicklung von Aktivitätsverteilungen auf der Ebene neuronaler Populationen durch Integro-Differentialgleichungen beschrieben. Neuronale Felder können unmittelbar an visuelle Eingänge angekoppelt werden, wie ich in Teilen der Arbeit durch Demonstrationen auf Videobildern zeige.

Der Übergang von reaktivem sensomotorischen Verhalten zu kognitiven Prozessen in den neurodynamischen Modellen wirft spezifische theoretische Probleme auf. Zum einen müssen in den parallel und kontinuierlich arbeitenden Systemen Abfolgen von diskreten kognitiven Operationen erzeugt werden. Dies wird unter Ausnutzung von Attraktorzuständen in der Dynamik der neuronalen Felder erreicht. Diese erlauben es, die kontinuierliche Veränderung der Aktivitätsmuster in geordnete Abfolgen von Zustandsübergängen zu gliedern. Zum anderen müssen aus dem komplexen Fluss visueller Informationen spezifische Objekte als Argumente für die kognitiven Operationen ausgewählt werden. Hierzu werden Attraktorzustände in räumlichen Repräsentationen verwendet, welche die Position eines ausgewählten Objektes wiedergeben und es ermöglichen, auf weitere Eigenschaften des Objektes zuzugreifen. Die Modelle sind so in der Lage, komplexe kognitive Operationen flexibel auszuführen, und sie erklären zahlreiche experimentell beobachtete Charakteristika menschlichen Verhaltens im Bereich des aktiven Sehens.

Contents

Chapter 1

Introduction

Visual perception provides us with a detailed sense of the space around us and informs us where we are in this space. It allows us to identify objects, to judge their structure and perceive their motions. And it guides our actions, when we grasp a cup, kick a ball, or duck under a branch hanging in our way. Due to its vital role in human perception and action, the visual system and the neural mechanisms that are involved in the processing of visual information are in the focus of research in neuroscience and psychology, and large efforts are aimed at reproducing the capabilities of human vision in artificial systems.

One central aspect of the biological visual system that is often given little consideration in computer vision approaches is that vision in humans is a highly active process. We do not just passively *see*, we *look*. We select points of interest in our surrounding and direct our gaze at them, constantly making eye and head movements to inspect different parts of a scene. Humans can shift their fixation point several times per second. Yet we are most of the time completely unaware of this behavior and how it shapes our perception. Only when we actively pay attention to our eye movements or intentionally suppress them do we get an idea of how important they are for the way we see our environment.

The most obvious reason that we employ such frequent eye movements lies in the structure of the retinas in our eyes. The receptor cells that form the basis for high-acuity shape and color perception are clustered in the central region of the retina, the fovea, and are sparse in the periphery. As a result, our ability to perceive shapes and colors outside of the foveal region is quite limited, much more than we are typically aware of. When reading a text like this, we can reliably decipher only a handful of letters to the left and right of our current fixation point without moving the eyes. And our ability to discriminate colors decreases drastically for objects at a visual angle of greater than 30° from the fixation point. Nonetheless, our subjective impression when viewing a scene is not that of a sequence of disconnected

images, each with just a small window of the scene clearly perceivable. Instead, our typical impression is that we constantly perceive a fixed visual scene surrounding us, colorful and highly detailed everywhere.

Active vision creates significant challenges for the processing of visual information, and it structures to a significant extent the neural mechanisms involved in visual perception. Many visual areas of the brain are tightly coupled to motor areas involved in the control of gaze shifts, so that behaviorally relevant parts of the environment can be selected as target for an eye movement. Visual representations of space have to take into account the variable gaze direction, and mechanisms for flexible reference frame transformations are needed to determine and keep track of object locations in the world. And finally, special processes are needed to form a coherent internal representation of the visual surrounding from a series of fixations of single objects—although this internal representation is more limited in many respects than we are typically aware of, as I will discuss in a later chapter.

But importantly, it does not seem to be only the limited region of high acuity in the retina that necessitates this kind of processing. In experiments on visual scene perception and memory, participants are sometimes required to view a limited array of simple stimuli without making any eye movements, with all stimuli close enough to a central fixation point that their details can be perceived simultaneously. Even under these conditions, the participants typically inspect the items one by one (Vogel et al., 2006). They do this without actually making eye movements to fixate them, but by focusing one item at a time through spatial attention.

It has been proposed that the key reason for this sequential selection of visual items even in the absence of eye movements lies in the *binding problem*, more concretely in the problem of feature binding. The problem arises from the fact that specialized populations of neurons represent different visual features. There are neural populations whose activity reflects stimulus colors, others represent stimulus shape or visual motion. The question arises now how in such a distributed system, the features that belong to different objects can be kept apart, while the features in different feature dimensions (such as color and shape) that belong to the same object can be grouped together. Assume, for instance, that you are shown a simple stimulus display containing a red circle and a green square. This will induce activity in the neural populations for color that indicate the presence of red and green in the visual scene, and activity in shape-sensitive neural populations indicating the presence of a square and a circle. But how can the neural system determine from such activity patterns that there is in fact a red circle and a green square present, rather than a green circle and a red square?

One approach to deal with this problem is a sequential and selective processing of visual items, as it is achieved by individual fixations or attentional selection. If each feature representation contains only the single feature

value of the currently selected item, the binding problem can be avoided. This has been prominently suggested in the Feature Integration Theory of visual attention (Treisman and Gelade, 1980). The theory states that while individual features can be processed in parallel (e.g., in visual search tasks for a green item among red ones), the binding between different features requires focused attention on a single item (e.g., in a visual search for a red horizontal line among red vertical and green horizontal lines). The theory is supported by a host of evidence from visual search, change detection, and image segregation tasks (Treisman, 1988).

The feature integration theory describes how the conjunctions of features that make up a visual item can be perceived in a bound form, but it does not explain how these bound features for multiple objects can be stored simultaneously in working memory to create an internal representation of a visual scene. An influential extension to address this question was proposed by Kahneman et al. (1992) in the form of the Object File Theory. This theory proposes that when an item in a visual scene is first attended, an *object file* is created in visual working memory. In this object file, the perceived visual features of the object are stored in a bound fashion, as well as more abstract information such as an object category. The object file then remains linked to one object in a scene over movements of the object and changes in gaze direction. It may be addressed via the object location when the object is attended again, which acts as a kind of pointer to the object file, and new features may be added or present features updated based on new sensory input. To account for the severe limitations in human scene memory, it is proposed that only a fixed number of object files—between three and five—can be maintained at a time in working memory, and the content of older object files is lost when this limit is reached.

A related idea was presented by Pylyshyn (2001, 2007), whose work focused on visual tracking and related tasks. He addressed the problem of object individuation, asking for instance how humans can keep track of an object and think or communicate about it before the object is identified, or how we perceive that something still remains the same object when it changes its locations or even its visual features. Pylyshyn suggests that a limited number of *visual indexes* are maintained by the visual system, which act as pointers to objects in the world. These pointers remain the same even when the object changes, and thereby provide individuation and the perception of persistent object identities.

This idea of pointers is taken further by Ballard et al. (1997), who interpret the visual fixation of an object as *deictic strategy*, a form of pointing out an object for cognitive operations. As a very simple example, the fixation of an object with the eyes can mark it as the target of a reaching movement. This is proposed as a general concept for both motor and cognitive behaviors, directly likened to the use of pointers in computer programming. The

role of the fixation can also be taken by spatial attention, and it is proposed that a limited number of such pointers can be held in working memory. Ballard et al. (1997) claim that "The concept of pointers changes the conceptual focus of computation from continuous to discrete processing." The pointers are suggested to be used in "cognitive programs" to solve the problem of variable binding, in that they provide the targets for elementary cognitive operations by linking to content in cortical representations. By using such pointers, the cognitive program can be made much more general and flexible than if the targets of the operations would have to be provided explicitly.

All of these cognitive theories are formulated using analogies from computer programming, referring to files and pointers and cognitive programs. But critically, these concepts cannot easily be transferred to neural systems. Files and pointers require an addressable, general-purpose memory structure. Neural representations, however, can only communicate with each other via the synaptic connections between them. They have no abstract addresses, and their connection patterns define their functional role—determining, for instance, what content can be retained in a specific neural working memory representation. It is therefore unclear how a pointer, which provides access to different representations at different times, could be implemented in such a system. Moreover, Object File Theory does not specify how the problem of feature binding is to be solved for the working memory representations of objects, if different features are still represented in separate neural populations. And finally, many of these theories require ordered sequences of operations (most prominently in the cognitive programs of Ballard et al.), without specifying how these sequences can arise in a biological neural system that is structurally set up to operate entirely in parallel.

In this thesis, I aim to overcome this discrepancy between cognitive models and neurophysiology. The approach that I use is the formulation of neurodynamic process models, based on the Dynamic Field Theory. In this theoretical framework, neural activity at the level of populations of neurons is described through Dynamic Neural Fields (DNFs), continuous distributions of activation that are defined over a space of behavioral variables. The change of these activation distributions over time is specified by a set of differential equations, which describe the effects of external stimulation and lateral interactions. Particular emphasis is put on the analysis of stable states in the field dynamics that arise due to the lateral interactions, and of the instabilities as transitions between these stable states. These stability properties are critical to support robust perceptual representations, decision making, and working memory in DNF models, and are the basis for autonomous behavior generation.

The DNF models constitute neurally plausible, integrated dynamical systems that do not employ abstract concepts like files which can hold arbitrary information. They operate in a continuous and inherently parallel fashion,

and all interactions within and between DNFs are mediated by fixed excitatory and inhibitory projections, consistent with synaptic connections in the neural system. These projections implement all operations within the DNF model, from the processing of sensory input to the generation of overt behavior. Neural representations over different feature spaces are modeled by separate DNFs, so that the neurodynamic models face the same types of binding problems as the neural system.

With this type of model, I will address several core problems from the field of active vision, and account for human performance in a variety of tasks. I will present four concrete DNF models. The first two address more elementary neural mechanisms, whereas the later two models combine these mechanisms to generate more complex cognitive behaviors. The first model I will describe addresses the interactions between spatial and surface feature information in early visual processing, and it introduces the core problem of feature binding as well as a first part of a solution. Concretely, the model aims to explain recent experimental findings on the effects of holding a color in working memory on the selection of target locations for eye movements (Hollingworth et al., 2013a). The DNF model captures the processes underlying visual working memory formation and maintenance, allocation of spatial and feature attention, and the planning and execution of saccadic eye movements, and thereby covers several core elements of active vision behavior in general.

The second DNF model deals with neural mechanisms for spatial transformations. Since every eye movement shifts the whole visual image, all spatial information that is retained in the reference frame of the retina is made obsolete whenever the fixation point changes. To obtain a stable representation of object locations in a visual scene, location information must be transformed from the retinal to a gaze-invariant frame of reference. This transformation can be described as a variable mapping, parametrized by the current gaze direction. In the DNF model, this mapping is realized as a continuous bi-directional coupling between spatial representations in different reference frames. This mechanism can account for neural findings of peri-saccadic remapping in retinal spatial representations. It also provides a fundamental capability for forming stable scene representations from multiple fixations.

The third model that I will present combines mechanisms from the first two models to capture the formation of a scene representations in working memory, and it directly addresses the problem of feature binding. In this model, one stimulus is selected at a time by coupled spatial and feature attention. The stimulus location and its surface features are then conveyed on separate pathways to a scene working memory representation and recombined there. These separate pathways are necessary to efficiently apply the reference frame transformation to the location representations, and are

consistent with established views of visual processing in the brain (Mishkin et al., 1983). The model is then able to perform a variety of change detection tasks, a key experimental paradigm to asses working memory properties and capacity in humans. Depending on the type of task, either parallel processing along the separate pathways or sequential processing relying on combinations of the two pathways is used. These different types of processing are consistent with human performance measures and reaction times in different types of change detection tasks.

Finally, in the fourth model, the previously introduced mechanism are applied to address the use of spatial language in humans. The model describes how relational spatial expressions (like "the cup is to the right of the monitor") can be generated from a perceptual representation of a visual scene, and how spatial expressions can be resolved to select specific visual items. In the underlying mechanism, a reference object and a target object for a spatial relation are sequentially selected from a scene, and the their relative position is then determined via a reference frame transformation into an object-centered spatial representation. The metric relative position is then mapped onto a symbolic spatial term, using a set of weight patterns that reflect spatial semantics. Due to the fact that all projections in this model are implemented as bidirectional, the system can be used to perform a variety of spatial language tasks simply by giving different sets of inputs. The model has succesfully accounted for the results of spatial term rating tasks (in which subjects rate the appropriateness of a certain spatial term for the spatial relation shown in a visual scene), and it has reproduced human reference object selection behavior in a task that requires the free generation of a spatial description.

The DNF models have to meet two key conceptual challenges. First, an architecture, composed of individual DNFs defined over different feature spaces, must be found with sufficient representational power to solve the core problems of active vision and explain human performance. For the task of forming a scene representation, the architecture must contain the memory representations of different visual features in such a fashion that they can be bound together for individual objects, but held separate for different objects. In addition, these working memory representations must still remain linked to sensory input and motor behavior, supporting for instance that a previously formed working memory representation is activated when the same stimulus is visually inspected again at a later time; or conversely, that the memory can be used to guide an eye movement to re-fixate a stimulus. This implements the pointer function discussed above. The model of spatial language must additionally address the problem of variable binding. To generate and interpret spatial expressions, the system must be capable of binding different objects to the grammatical roles of target and reference object in a flexible manner.

The second key challenge is that the models must implement the required sequential operations. For forming a scene representation, the model must sequentially select individual items, moving on to the next item once the currently focused one is memorized. Similarly, the spatial language model must sequentially select the target and the reference object of a spatial relation and assign them to their respective semantic roles. What would be achieved trivially by a sequence of commands in an imperative programming language must here be achieved by the internal dynamics of the neural architecture. These dynamics, initially set into motion by an external input (e.g., from the visual stimuli), drive the transition of the activation patterns in the system through a series of stabilized states. In these transitions from one stabilized state to another one, the system can implement different types of decisions, like the selection of one stable state out of a set of qualitatively different potential successor states. The stability properties that are emphasized in the Dynamic Field Theory are critical here to prevent this series of state transitions from degenerating into chaos, especially in the presence of random noise in the visual input and the internal representations.

The DNF architecture with its internal dynamics can thus be viewed as implementing an algorithm, performing a sequence of processing steps on the visual input. These processing steps are however not given by explicit instructions, but instead emerge from the dynamics of the activation patterns. And unlike classical algorithms, the dynamical system is not started at a fixed time and then terminates, but is conceptually a continuously running system, in which operations are initiated simply by the presentation of external stimuli. This mode of operation makes DNF architectures particularly suitable for robotic applications, where a high degree of autonomy is desirable. The value of DNF models thus lies not only in their ability to explain the operations of neural systems and account for psychological findings, but they are also valuable for designing artificial autonomous agents. They allow continuous operation and responsiveness to external stimuli without the need for any higher-order system to control and monitor them.

All of the models presented in this thesis provide this kind of autonomy to a certain degree, making them viable as building blocks for autonomous agents. The model of saccade target selection continuously reacts to visual stimulation, executes simulated gaze-shifts to fixate salient stimuli, and can modulate its behavior in different ways to support goal-directed active vision. The spatial transformation model provides continuous coupling and updating of spatial representations, although it does not produce overt behavior. The model of scene representation has direct applications in guiding robotic action, and a robotic version of this DNF model has been successfully employed. Likewise, the model of spatial language has obvious uses for verbal interfaces in human-machine interactions, and closely related models have been used for this task. In this scenario, using a model that is based

on neurobiology and explicitly aimed at reproducing psychological findings also has a concrete advantage: It allows easy and intuitive communication even in ambiguous situations since the behavior of the model follows characteristics of human behavior. These practical applications highlight the value that biological models offer to the field of applied sciences.

Chapter 2

Dynamic Field Theory

2.1 Overview

Dynamic Neural Fields (DNFs) are a class of models that describe neural and behavioral processes through the continuous change in distributions of neural activation. They allow a direct mapping of neural activation patterns to behavioral variables that are measured in psychophysical experiment, such as movement directions or reaction times. This makes DNFs particularly suited for behavioral modeling. DNF models have been used successfully to explain a variety of psychophysical findings, such as biases in working memory (Simmering et al., 2006) and performance in change detection tasks (Johnson et al., 2009a), as well as developmental changes in working memory capacity (Spencer et al., 2001; Perone et al., 2011). At the same time, DNF models have been employed to reproduce and explain measured neural activation patterns, in particular in the planning of saccadic eye movements (Trappenberg et al., 2001; Marino et al., 2012) and the preparation of reach movements (Bastian et al., 2003; Cisek and Kalaska, 2005). Finally, the same type of model has been used in robotics to actively perform cognitive tasks and generate behavior. Applications include object recognition (Faubel and Schöner, 2009), scene representation (Zibner et al., 2011a), and interpretation of spatial language (Lipinski et al., 2009).

2.2 Biological motivation

The biological basis of DNFs lies in neural population codes. In the nervous systems of animals, sensory signals, motor plans, and cognitive states are encoded in the activity of neurons. This activity is expressed through action potentials (APs, also referred to as spiking or firing of the neuron). An AP is a rapid, transient de-polarization of the neuron's membrane potential, typically lasting around 1 ms. Highly active neurons produce APs in quick succession, whereas less active neurons generate APs more sparsely or not

at all. These APs are transmitted to other neurons via the cell's axons, long and thin projections of the cell body. They terminate in synapses that make contact with other neurons, and they affect those receiving neuron's activity by the release of chemical neurotransmitters that either excite or suppress the activity of the target cell.

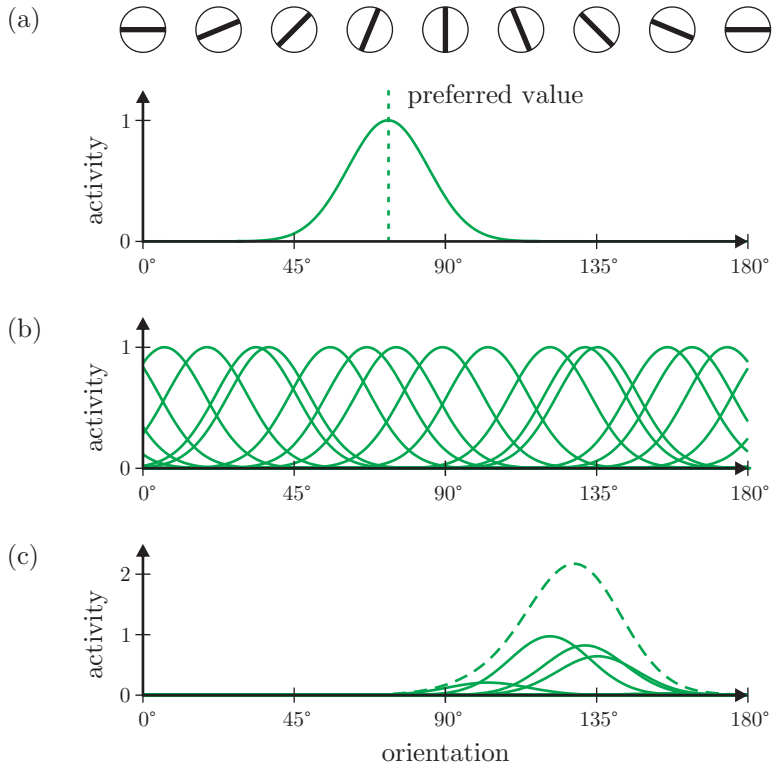


Figure 2.1: Derivation of continuous activation distributions from neural tuning curves. (a) Idealized tuning curve of an orientation-sensitive neuron in visual cortex, plotting the activity of the neuron (as normalized firing rate) against the orientation of a visual stimulus (shown on top). (b) Tuning curves from a sample of orientation-sensitive neurons, covering the space of possible orientations. (c) Continuous activation distribution as superposition of scaled tuning curves. Tuning curves from (b) are scaled with the neuron's activity in response to a specific stimulus (solid lines). The dashed line is the sum of these scaled tuning curves.

To generate goal-oriented behavior, neural systems often have to encode metric values of a sensory stimulus or a planned action. This may include the position of a visual stimulus, the direction of a perceived motion, or the parameters for a motor action. In neural systems, particularly in the cortex of the vertebrate brain, such metric values are typically represented

in the form of population codes (Erickson, 1974). In this form of coding, the metric value is reflected in the distribution of activity over a large assembly of neurons. The activity of each individual neuron of this population can be described by a tuning curve. For instance, a sub-group of neurons in the primary visual cortex is sensitive to edge orientations in visual stimuli. Each of them will show high activity for a certain orientation (the neuron's preferred value), moderate activity for similar orientation values, and low activity for dissimilar values. Plotting the neuron's activity (e.g., as the firing rate) against the stimulus orientation yields the tuning curve, which in many cases takes the approximate shape of a Gaussian centered on the neuron's preferred value (Swindale, 1998; see Figure 2.1a).

Individual neurons in a population have different preferred values, distributed over the space of possible metric values that the population can represent. (In the following, I will refer to this space as the *feature space* for the neural population, even if the space does not strictly reflect any visual surface feature.) The tuning curves of the neurons typically show considerable overlap, such that they cover the whole feature space (Figure 2.1b). In certain cortical areas (particularly in sensory areas), there is a topological organization of the neurons, such that cells that are located close to each other on the cortical surface tend to have similar tuning curves. This property is however not necessary for a population code, and the physiological arrangement of neurons is not addressed in Dynamic Field Theory.

The activity of any single neuron in a population is generally not informative about what is encoded. The tuning curves are often very broad, and the firing rate of each neuron is subject to noise. Furthermore, a neuron's activity may depend on additional factors beside the similarity between the stimulus feature and the cell's preferred value (such as stimulus contrast for orientation-sensitive neurons). To interpret the population code, one must look at the distribution of activity over all neurons. A simple estimate $\widetilde{v}$ of the encoded value may be obtained as a weighted average of the neurons' preferred values p_i, with the average firing rates r_i as the scaling factors:

$$\widetilde{v} = \alpha \sum_i p_i r_i \tag{2.1}$$

Here, α is a normalization factor (e.g., $\alpha = \frac{1}{\sum_i p_i}$). Note that for circular feature spaces (like orientation), the computation of the average has to be adjusted accordingly. This form of analyzing population activity has been used successfully in interpreting neural firing patterns during the preparation and execution of reach movements (Georgopoulos et al., 1986). However, it relies on certain assumptions about the shape of the tuning curves (symmetry and monomodality) and the shape of the activity pattern. In particular, the weighted average will only be meaningful if only a single value is encoded in the neural population.

However, the population code representation goes far beyond the encoding of a single metric value. First, a population of neurons may encode multiple values in parallel, for instance the features of different stimuli or different possible motor plans. This was shown in the empirical and theoretical work of Cisek and Kalaska (2005). In this case, several groups of neurons (with overlapping tuning curves within each group, but dissimilar between groups) are active at the same time. In addition, the intensity of the activation in each group of neurons and the width of the activity distribution can be informative. It can make a difference, for instance, whether there is low activity in many neurons with relatively different tuning curves, or high activity in a small group of neurons with strongly overlapping tuning curves, as has been shown for the formation of reach plans in the motor cortex (Bastian et al., 2003). These results indicate that the full activity distribution is relevant and cannot be reduced to a single value without loss of important information.

Finally, the population can also represent the absence of any specific values by a uniformly low activity profile. While this case may appear trivial, it is critical for a sensory representation to be able to signal that no salient stimuli are present, or for a motor representation to be able to not generate a movement command at certain times. In these cases, reducing the activity distribution to a single value (using, e.g., the weighted average as described above) will yield a misleading result, as this single value cannot reflect the absence of stimuli or motor plans in an intuitive way.

2.3 From population activity to DNFs

Dynamic Field Theory views the activity distributions over neural populations and the evolution of these distributions over time as the central element of neural representation and the basis for neural processing and behavior generation (Wilson and Cowan, 1973; Amari, 1977). Accordingly, DNF models are aimed at capturing these activity distributions, explaining how they form and how they shape overt behavior. In doing so, certain abstractions from biological neural populations are employed. First, the discrete APs are replaced by a continuous activation variable. This is a basic simplification that is used in many neural models, including classical neural networks. It relies on the assumption that the relevant information in neural processing is carried by the rate of APs generated over a brief time window, and not by the exact timing of individual APs. This greatly simplifies the simulation and analysis of the evolution of population activity over time. There are, however, a few known exceptions to the assumption that AP timing is not critical for neural processing, such as the use of AP synchronicity in binaural sound localization. These can consequently not be modeled adequately with DNFs.

The second and more fundamental abstraction is that Dynamic Field Theory also abandons the modeling of the discrete neurons that make up a population. DNFs instead describe a distribution of activation over the continuous feature space. The activation at one location in feature space is interpreted as directly supporting the corresponding feature value, removing the intermediate step via a neuron's tuning curve. This abstraction is based on the belief that the individual neural cells in a population code merely provide a discrete implementation of a functionally continuous representation.

A continuous activation distribution over the feature space can be formally derived from neural firing rates as a superposition of the weighted tuning curves, as shown in Figure 2.1c. Here, each tuning curve from Figure 2.1b is scaled with its activity in response to a specific stimulus (with many neurons having an activity close to zero). The sum of these scaled tuning curves, shown as dashed line, is then a continuous representation of the population activity. Note that additional normalization steps are necessary when using this approach as an analytical tool in order to compensate for inhomogeneities in the sampling of the feature space (Erlhagen et al., 1999).

A special focus of Dynamic Field Theory is on the dynamics of neural representations, in particular the formation of stable states in the activation distributions. Such stable states emerge in DNFs from lateral interactions, which reflect the interactions that occur in biological neural populations through synaptic connections between the neurons. A general interaction pattern that is frequently found in neural systems is described as local excitation and surround inhibition. This means each neuron in a population tends to form excitatory connections to neurons that have similar tuning curves, such that these neurons mutually excite each other. (Such projections are "local" in a physiological sense if the neurons are organized topologically, otherwise the are "local" only in the sense that the neurons' preferred values are close to each other in feature space). At the same time, each neuron acts in an inhibitory fashion on those neurons that have dissimilar tuning curves, typically through an indirect projection via inhibitory interneurons.

These lateral interaction patterns shape the activity distributions in biological neural populations (Jancke et al., 1999). If a group of neurons with overlapping tuning curves is activated by an external stimulus, the local excitation will further increase these neuron's activation, while surround inhibition will suppress activation of other neurons in the population. This stabilizes the distribution of activation in that population against fluctuations in the external input. Dynamic Field Theory views such stabilization of activation states as critical to generate robust goal-oriented behavior in a world where the sensory input used to guide this behavior is often noisy or transient. Of course, this stability must be balanced by sufficient flexibility so that behavior can be adjusted to changing situations. This flexibility is

achieved through transitions between different stable states. I will describe the possible stable states in DNF models and the transitions between them in detail below.

2.4 Mathematical formulation

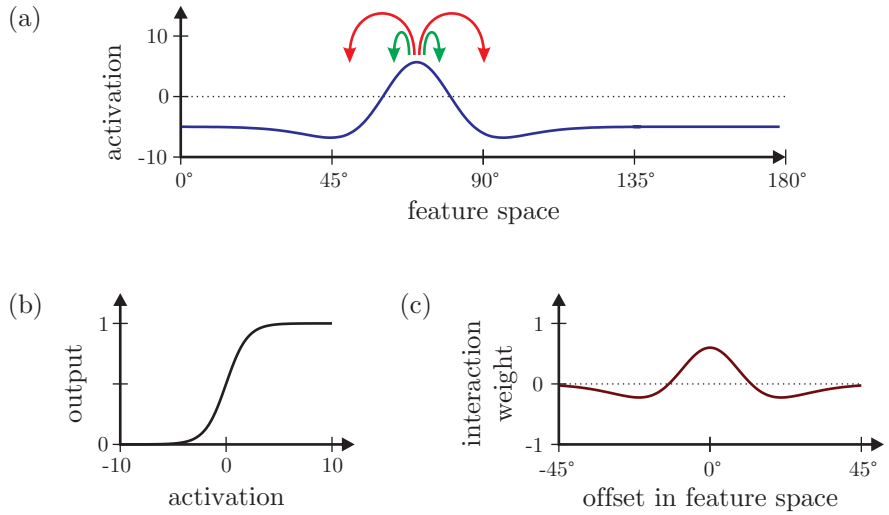


Figure 2.2: Dynamic Neural Field, sigmoid output function, and interaction kernel. (a) Dynamic Neural Field as distribution of activation over a feature space. Arrows indicate lateral interaction effects of regions with suprathreshold activity (green for excitatory, red for inhibitory interactions). (b) Sigmoid (logistic) output function. (c) Lateral interaction kernel with a difference-of-Gaussians profile.

The temporal evolution of the field activation is governed by a differential equation that determines the change of activation for each position depending on the current state of the field, the effects of lateral interactions, and the external input to the field. This equation has the general form (Amari, 1977)

$$\tau \dot{u}(x,t) = -u(x,t) + h + i(x,t) + \int k(x-x')f(u(x',t))dx' + q\xi(x,t). \quad (2.2)$$

Here, $u(x,t)$ is the field activation at position x in the feature space and at time t (Figure 2.2a). The rate of change $\dot{u}$ (the derivative of the field activation over time) is scaled with a time constant τ. The parameter h defines a global resting level for the field activation. By convention, the resting level is negative, and the threshold for the output function (see below)

is always at zero. The time-dependent external input to the field is given by $i(x,t)$. The following term describes the lateral interactions in the field: An interaction kernel k is convolved with the output $f(u)$ of the field. Finally, $\xi(t)$ is a Gaussian white noise process, with a scaling factor q determining the noise level.

The output function is a sigmoid nonlinearity, typically implemented here as a logistic function with a steepness parameter β (Figure 2.2b):

$$f(u(x)) = \frac{1}{1 + exp(-\beta u(x))} \tag{2.3}$$

The output is close to zero for low activation values, grows for activation values around zero, and saturates at one for high activations. The effect of using an output function of this kind is that only those regions in a field that have sufficiently high activation levels contribute in a significant way to the lateral interactions in a field. In addition, the saturation of the output function limits the maximal strength of the interaction effects. For $\beta \to \infty$, the sigmoid approaches a step function with threshold at zero. I will sometimes refer to this (soft) threshold of the output function to distinguish between activation levels that produce significant output (and therefore induce interaction-driven instabilities, as described below) from activation levels that do not.

The lateral interactions in the field are homogeneous, and the interaction strength between two points in the field only depends on the signed distance between them (in feature space). This dependence is described by the interaction kernel $k(x - x')$. Reflecting the connection patterns in biological neural populations, the interaction kernels used in DNF models typically feature lateral excitation over short distances in feature space, and inhibitory interactions over longer distances. This is implemented with a difference-of-Gaussians kernel (with a Mexican hat shape) with an optional global inhibitory component (Figure 2.2c). For a one-dimensional feature space, the kernel function can be given in the general form

$$k(x) = \frac{c_{\text{exc}}}{\sqrt{2\pi}\sigma_{\text{exc}}} \exp\left(-\frac{x^2}{2{\sigma_{\text{exc}}}^2}\right) - \frac{c_{\text{inh}}}{\sqrt{2\pi}\sigma_{\text{inh}}} \exp\left(-\frac{x^2}{2{\sigma_{\text{inh}}}^2}\right) - c_{\text{gi}} \tag{2.4}$$

Here, c_{exc}, c_{inh} and c_{gi} are interaction strengths, and the widths σ_{exc} and σ_{inh} determine the range of local excitation and surround inhibition (with $\sigma_{\text{inh}} > \sigma_{\text{exc}}$). The strength of either the local surround inhibition, c_{inh}, or the global inhibition, c_{gi}, may be zero.

2.5 Attractor states and instabilities in DNFs

The lateral interaction patterns in DNFs qualitatively shape the possible attractor states in the field dynamics and the possible transitions (instabilities) between them. In particular, the typical interaction patterns of

short-range excitation and long-range inhibition promote the formation of localized peaks of activation. These serve as units of representation for individual feature values in neural fields. The transitions between stable states with qualitatively different configurations of activation peaks are the elementary operations in fields, and serve as building blocks for complex cognitive processes. For this reason, I will describe the attractors and the instabilities in some detail here.

Other attractor states than localized activation peaks may also arise from the field equation, in particular when fields are defined over multi-dimensional feature spaces (as described in the following chapter). These include for instance stabilized ridges or rings of activation and different forms of repetitive patterns. These are not typically employed in DNF models, and I will not further discuss them here. The reasons are that such attractors are likely not physiological (measured activation patterns in neural populations can typically be described by one or more localized activation hills, see Cisek and Kalaska, 2005), and that they do not offer a straightforward interpretation in terms of what they represent.

2.5.1 Detection instability and self-sustained peaks

Consider a field that does not receive any external input, and in which the activation at all points is well below the output threshold. In this case, the lateral interactions do not take effect, and the deterministic part of the field equation can be reduced to

$$\tau \dot{u}(x,t) = -u(x,t) + h. \tag{2.5}$$

These field dynamics drive the activation at every point in the field toward the resting level h: If the activation at one point x is higher than h, Equation 2.5 yields a negative rate of change, if it is lower, it yields a positive rate of change. If the resting level itself is also well below the output threshold—which is the typical configuration in the DNF architectures treated here—the field activation exponentially relaxes toward the value h. The activation will also return to that level when slightly perturbed. This is a first attractor state of a field, which I will call the *sub-threshold attractor*.

If an external input is now applied to the field, the field activation will move toward the sum of this input and the resting level (Figure 2.3a). As long as the field activation remains so low over the whole field that the field output is negligible (and therefore no significant interaction effects occur), the field activation will continue to track the input in this fashion. This state is not qualitatively different from the input-free state, and is still an instance of the sub-threshold attractor.

Now assume that a single, localized external stimulus (e.g., with a Gaussian shape) is applied, and its strength is slowly increased over time. At

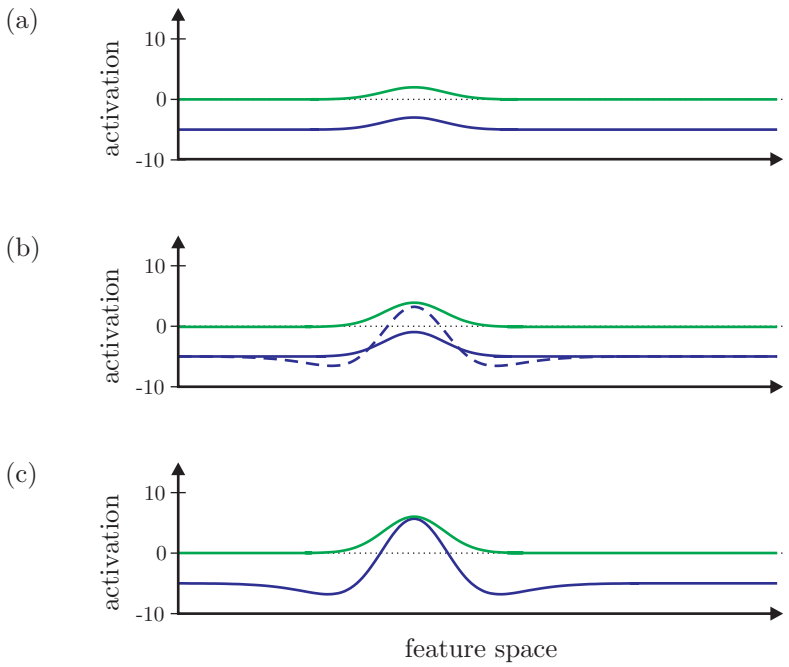


Figure 2.3: Sub-threshold and peak attractor in a DNF. (a) For weak localized input (green plot), the sub-threshold state is the only attractor state for the field activation (blue line). The activation then directly mirrors the input pattern, shifted by the field's negative resting level. (b) For moderate input strength, the field is bistable. The sub-threshold state (solid blue line) and the activation peak that is stabilized by lateral interactions (dashed blue line) coexist as attractor states. (c) When the input is strong enough to drive the field activation locally above the output threshold, the activation peak is the only possible attractor state.

first, the field activation will track this changing input as it did before. As the field activation rises further and approaches the threshold of the sigmoid function, the output signal increases around the location of the input. Now the lateral interactions begin to take effect. When the interactions take the typical shape of local excitation and surround inhibition, they will further increase activation at the input location, and decrease it in the surrounding regions. This brings about a qualitative change in the field's activation pattern: As the self-excitation drives activation at the input location, it further increases the output that is produced by the field, and this in turn further strengthens the self-excitation. The activation level around the input location rises significantly above the level induced by the input itself, and a stabilized *activation peak* forms (Figure 2.3c).

The growth of this peak is limited by two factors. First, as activation rises beyond the output threshold, the sigmoid output function goes into saturation; the output from any one point in the field does then not increase any further. Second, the growth of the activation peak in the feature dimension is limited by the shape of the interaction kernel. Since the excitatory effects are limited to a short range, the output of one point in the field only increases the activation value in its direct vicinity. If an activation peak grows wider, the points at one edge of the peak do no longer contribute positively to the activation of points at the opposite edge. In contrast, the inhibitory interactions have a longer range, and more points contribute to the inhibitory interactions as a peak grows in width. This increase of inhibitory effects for wider peaks limits the spatial expansion of an activation peak (Amari, 1977), and drives the activation pattern toward a stereotypical peak size.

It can be seen that the activation peak and the sub-threshold state are in fact qualitatively different attractor states if the external input is slowly decreased again after a peak has formed. The supra-threshold activation peak will remain even at input strengths that had not been sufficient to induce it in the first place. This means that for these values of input strength, both the sub-threshold activation pattern and the activation peaks are possible and distinct attractor states (Figure 2.3b). The system is bistable under these conditions, and which attractor state it relaxes to is determined by its history. When the input is increased sufficiently, as described above, the system becomes monostable with only the peak state remaining as attractor. This change of attractor states is mathematically described as a tangential bifurcation, in which the sub-threshold attractor collides with a repellor state and is extinguished. Conversely, if the external input is decreased sufficiently, the peak state will disappear as an attractor in another tangential bifurcation. (Depending on the parameters of the field, the latter may only occur for negative input, as described below.)

The transition from the sub-threshold state to an activation peak is referred to as *detection instability* in Dynamic Field Theory. It marks the point at which the presence of a certain feature value is sufficiently supported by the input signal for the system to start actively representing this value. In the sub-threshold state, the input signal is passively reflected in the field activation, but it is neither actively maintained nor transmitted to any downstream structures via the field's output signal. Once a peak has formed, the lateral interaction stabilize it against fluctuations in the input signal, even if the input strength decreases below the value that is necessary to induce a peak. Moreover, the activation peak will smoothly shift its location in response to small changes in the input location, and thereby provides the ability to track an input signal that varies over time. Only if the input strength decreases significantly or there is a sudden and large change in the

input location, the peak will become unstable and disappear. This is called the *inverse detection instability.*

If a DNF features strong self-excitation in its lateral interactions or a resting level that is close to the output threshold, the peak state may remain stable even in the absence of any input. Such self-sustained peaks are used to model working memory in behavioral models. This is consistent with neurophysiological findings that during retention of features in working memory, neurons sensitive to these features show sustained activity (Wang, 2001). Mathematically, the self-sustained peak (in the absence of any localized input) is an instance of a line attractor. While it is stabilized against decay, it is not stabilized against shift along the feature dimension. In DNF models of working memory, random drift of the self-sustained peaks occurs in the presence of random noise in field activation, and systematic biases in the memory representation may be introduced by non-homogeneous external input or by interactions with other activation peaks.

2.5.2 Selection decisions and averaging

Additional attractor states and instabilities can occur when multiple localized inputs are present. In the field equation, separate input components are simply combined additively, so the term $i(x,t)$ in Equation 2.2 would reflect the sum of all input components. If the lateral interactions in a field are purely local and the locations of the external inputs are sufficiently distant from each other, peaks can form or decay at multiple sites independent of each other. This is also true for self-sustained peaks in models of working memory.

The situation is different if the inhibitory interactions in a field are global. If the activation is increased toward the output threshold at two separate locations in such a field, the lateral interactions create a competition effect: The self-excitation acts only within a short range, so each location only excites itself. In contrast, the global inhibition originating from each location also acts on the other location. If the activation level is slightly higher at one of the locations at any point in time, the lateral interactions have the potential to amplify this difference. The more active location creates more self-excitation and therefore generates further increase in activation. At the same time, it has a stronger suppressive effect on the less active location. When interaction parameters are chosen appropriately, this will lead to the complete suppression of one location below the output threshold (Figure 2.4a). The activation peak at the prevailing location then no longer receives any inhibitory input from the competing location, and consequently has the same shape as if no competing input existed.

The transition from two competing active regions to a single activation peak is another instability, called *selection instability.* When the system goes through this instability, a decision occurs that determines which of multiple

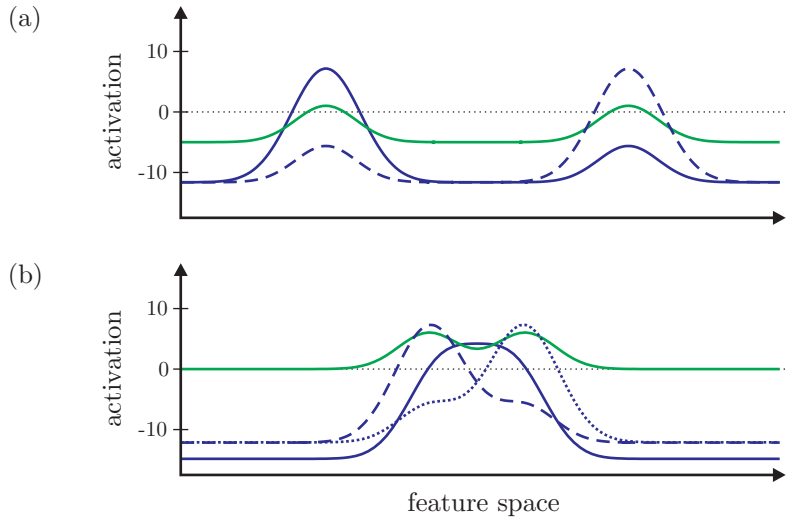


Figure 2.4: Selection instability in DNFs. (a) When two distant, localized inputs (green plot) are applied to a DNF with strong global inhibitory interactions, the field will undergo a selection decision in which an activation peak forms at one location while the other location is suppressed. The field is then in a bistable state (possible attractors shown as dashed and solid blue plots). (b) For two nearby localized inputs, an averaging peak can emerge as third attractor state (solid blue plot), in addition to the states where a peak is localized on one input (dashed and dotted blue plots).

possible attractor states the system settles in (shown as solid line and dashed line in Figure 2.4a). The details of input strengths and timing decide which attractor state is selected, that is, at which of multiple input locations a peak forms. After a peak has formed and activation at other locations is suppressed, this decision is stabilized against fluctuations of input strengths. Nonetheless, the state of the field is still continuously coupled to the input: As before, the activation peak can track the input location if it changes smoothly in feature space. And the activation peak can switch to another input location if the input strengths become sufficiently imbalanced.

This combination of stability and flexibility make the selection mechanism in DNFs suitable for behavioral modeling. As an example, consider a reaching or pointing movement with multiple possible target objects, conveyed by a noisy sensory signal. Assuming that only a single reach movement can be specified at any time, the system has to select one object and ignore the others. To create efficient, goal-oriented behavior, the selection should not switch for spurious reasons, such as when the input strengths fluctuate due to sensory noise. On the other hand, the selection should still be

allowed to switch if one of the non-selected objects suddenly appears significantly more appealing (indicated by an amplified input signal). Finally, the autonomous tracking that occurs in DNFs is valuable for motor planning if the target objects are allowed to move.

The above considerations on the selection decision assume that the inputs are always distant from each other, such that excitatory interactions between the input locations can be neglected. If this is not the case, activation peaks can merge in DNFs. The most obvious way this can happen is directly by merging of the input signals. For instance, if the centers of two Gaussian inputs are located closely together, they form a single-peaked input pattern. This will generally also induce a single activation peak in the field. Merging can also happen if the input has two distinct local maxima through lateral interactions (solid line in Figure 2.4b). In this case, two active regions may form in a field that are separated in feature space, but still so close that they mutually excite each other—and that both excite the region between them. Through mutual excitation of the proximate borders of the two active regions and the interjacent space, the active regions may expand toward each other and merge to a single peak. If the inputs are symmetrical, the activation peak will be centered in the middle between them, even if this means that it does not overlap with either peak of the input pattern. For certain input conditions and field settings, this attractor state with an averaging peak may coexist in a multi-stable dynamic regime with attractor states in which a peak has formed over one of the input locations (these three attractor states are shown in Figure 2.4b). As in the pure selection decision, the details of input strengths and timing decide which attractor state the system settles in.

The mechanisms of selection and averaging in DNFs have been used in the modeling of movement planning, in particular for saccadic eye movements (Kopecz and Schöner, 1995; Wilimzig et al., 2006). If multiple visual stimuli are present (as is always the case in naturalistic visual scenes), one of them has to be selected as the saccade target. In psychophysical experiments with controlled stimuli, it can be shown that the selection changes into an averaging behavior if a group of stimuli is presented close to each other (and distant from the current fixation point; Van der Stigchel and Nijboer, 2011). DNF models can explain these effects and match neural activation patterns in the superior colliculus, a midbrain structure that is involved in eye movement control. The use of DNFs in modeling saccade behavior will be one topic in Chapter 3.

2.6 Constructing behavioral models

The attractors and instabilities described above form the elementary building blocks to generate more complex behaviors in the DNF models that I

will present in the following chapters. Activation peaks are used as stabilized representations of sensory inputs or cognitive states. The detection instability here creates the demarcation between non-informative sub-threshold activation states and the representation of specific values along the feature space. Activation peaks that are self-sustained serve as model of working memory, to retain feature values after the input that supported them has vanished. The selection instability implements decisions between different alternatives, either in the generation of a response or as an intermediate step to form or alter a cognitive state.

More complex states can arise when multiple fields are coupled to each other to form larger architectures. Coupling fields together is achieved by using the output of one field as an input to another field. Typically the output is first convolved with an additional interaction kernel that describes synaptic connection patterns between two neural representations. This coupling between different fields may be done in a bidirectional or circular fashion, such that the output that one field sends to another field acts back in an indirect manner on the activation distribution in the source field. Such architectures then act as coupled dynamical systems, in which attractor states can no longer specified individually for each DNF, but are distributed over the whole architecture.

The architectures receive external sensory input, either from actual sensors (such as a camera, see Lipinski et al., 2012) or in the form of simulated stimuli. They can produce responses in the form of motor behavior that is driven directly by the output of the DNFs, either for a simulated effector or on actual robotic hardware (e.g., Bicho et al., 2000). The differential equations of the DNFs generate activation time courses that describe the real-time neural processing. DNF models can thereby produce variable reaction times in their response generation based on the characteristics of the provided stimuli. For instance, if a model has to perform a selection decision between two stimulus locations to generate a response, the reaction times will be higher if the two stimuli are almost equal in strength, since in this case it takes longer to resolve the competition between them via the field interactions. Moreover, response distributions that model the variability in human behavior can be generated by adding random noise to field activations in every time step, which reflect non-task-specific fluctuations in neural activity in the brain.

The DNF architectures presented in this thesis are conceptualized as continuously operating neurodynamic models. This means that there is no fixed start and end point for their operation, but they react whenever stimuli are presented to them and can produce responses over an unlimited duration. This contrasts with certain other neurodynamic models (e.g., Denève et al., 2001) that are initialized to a certain state and then allowed to evolve until they reach an attractor state. To generate autonomous behavior, this

mode of operation would require additional control structures to create the initial state, detect when an attractor has been reached, and read out this attractor state. The models presented here are instead intended to run autonomously, and they should ultimately provide systems that can control an autonomously behaving agent. To model psychophysical experiments, these models are still operated in such a way as to perform single trials, with reinitialization between trials to ensure reproducibility. But the processing in the models is driven by the timing of the stimuli, and they would continue to operate in a meaningful way after the completion of a trial.

2.7 Numerical simulations

Since the activation time courses of DNFs with non-trivial interactions cannot be solved analytically, the main tool to explore the behavior of DNF models are numerical simulations. To this end, the DNF—which is conceptually continuous in time and feature space—has to be discretized. For all simulations of the models presented in this thesis, the feature spaces of the DNFs are sampled at equidistant points, and the differential equations are evaluated at fixed time steps to update the field activations using the Euler method.

The equation to update the field activation can then be formulated in its discrete form as follows:

$$\begin{aligned} u(x_j, t+\Delta t) =& \frac{\Delta t}{\tau}\left(-u(x_j,t) + h + i(x_j,t) + \sum_{l=1}^{n} k(x_j - x_l) f(u(x_l,t))\right) \\ &+ \sqrt{\Delta t} q \xi(x_j, t) \end{aligned} \tag{2.6}$$

Here, $x_1, \ldots, x_n$ are the equidistant sampling points in feature space, and Δt is the step size for the Euler step. The Euler method is used to update field activations in preference of other more elaborate methods with variable time steps because only the Euler method allows adequate treatment of random noise in the field activations.

Despite this discrete implementation used for numerical simulations, the DNFs are still conceptualized as being continuous. This is reflected in the discrete implementation in the following ways: The sampling rate for the feature space is chosen in such a way that the sampling does not determine the behavior of the model. In particular, using a finer sampling or shifting the sampling points should not produce any qualitative changes in the model's behavior. Numerical deviations from a (hypothetical) exact solution of the continuous field equation are of course unavoidable, but these do typically not produce any qualitative effects, especially given that there is often random noise added to the field activations that masks any small numerical differences.

Similar considerations are also used to determine an appropriate value for the time step Δt. The activation patterns should not change qualitatively within a single step, but evolve smoothly. Here, one has to take into account that qualitative deviations from the exact solution of the differential equation can occur if the time step is chosen too large, in particular in the form of overshoot of activation and oscillations around an attractor state. In this approximation of continuous changes in activation states, DNF models differ significantly from classical neural network models, which frequently undergo qualitative changes of their activation patterns and even perform complete operations in a single discrete time step.

All numerical simulations of DNF models for this thesis were done in the computing environment Matlab. The latest models were implemented using the *cosivina* toolbox, an open source, object-oriented framework that I have developed to simplify the design and simulation of DNF architectures in Matlab (available at *bitbucket.org/sschneegans/cosivina*).

Chapter 3

Feature Associations in Multi-Dimensional Neural Fields

3.1 Overview

The topic of this chapter is the attentional selection of items in a visual scene and the generation of saccadic eye movements. As briefly touched on in the general introduction, one key reason why humans make eye movements is the structure of the retina. The central region of the retina, the fovea, is tightly packed with cone-type light receptor cells, which form the basis for color vision and high-acuity shape perception. In the peripheral regions of the retina, rod-type cells are predominant, which are important for movement perception, but contribute little to color and shape perception. In order to perceive the detailed features of an object in the visual scene, it is therefore necessary to fixate that object, and thereby bring its image onto the fovea.

The main behavioral mechanism to achieve this is a saccadic eye movement. Saccades are rapid, coordinated movements that typically take less than 200 ms. Between the saccades there are periods of fixation during which the gaze direction remains fixed. Saccades are ballistic, meaning that the movement is fully specified before it is initiated. The eye movement is then executed with a stereotyped velocity profile, without any update or correction from visual guidance during the saccade. If the intended target point is not reached by the initial saccade, it is quickly followed by a correction saccade. For larger gaze changes, the saccade may be executed as a coordinated movement of head and eyes.

During free viewing, saccades are driven to a large degree by the visual saliency of image patches (e.g., due to high contrast or motion in the visual image). They are closely coupled to visual attention: Attention is obligatorily directed to the saccade target location (the new fixation point) before

the initiation of the eye movement (Hoffman and Subramaniam, 1995), and focused attention to a certain location is typically followed by an eye movement to fixate that spot. It is possible, however, to actively suppress the eye movement when directing attention at a location. This is referred to as *covert attention*, in contrast to the *overt attention* that is accompanied by an eye movement.

In order to plan saccades for goal-directed actions, bottom-up attention that is only driven by the saliency distribution in the visual image is not sufficient. To guide everyday actions such as making coffee, sequences of saccades to behaviorally relevant objects are used (Land and Hayhoe, 2001), for instance to precisely plan a reaching movement to the coffee pot or to check the water level in the coffee machine. These have to be controlled by top-down inputs that specify what kind of object to look for, since one cannot assume that the objects that are relevant for the current behavior are always the ones that are most salient in the visual scene. This top-down guidance of visual attention implements a form of visual search: Given the visual features of an object, the visual system has to determine its location in a visual scene. The search template for such behaviors is typically held in visual working memory (VWM), brought there for instance by activation of an item in long-term memory.

Recent experimental evidence shows that such guidance by visual features held in working memory is not purely the effect of a mental strategy for goal-oriented behavior, but rather an inherent property of the human visual system. In a series of psychophysical studies, Hollingworth and colleagues combined a color working memory task with a saccade task (Hollingworth et al., 2013a,b). Participants had to hold a color in working memory and then performed a simple saccade task to a sudden onset target, while ignoring possible distractor targets. The memorized color was not relevant for the saccade task—the target was unambiguously identified by its location, and stimulus color was not predictive of whether a stimulus was a target or a distractor. Nonetheless, saccade behavior was systematically influenced by the color held in working memory. Participants made saccades to targets faster when these matched the memorized color, were more strongly influenced by distractors of that color, and the exact amplitudes of their eye movements varied depending on color match. These effects were found even for rapid saccades, where such guidance effects had previously not been expected.

These results speak for the existence of continuous interactions through which the content of VWM influences saccadic motor planning even when it is not beneficial for the task, and indicate that this happens on an early level of visual processing. This contrasts with a class of existing models of visual search which propose separate processing stages: first a parallel, bottom-up stage that is not influenced by task requirements, then a selection process in which top-down inputs are incorporated (Wolfe, 1994; Bundesen et al.,

2005). It is consistent however with the *biased competition* approach for visual attention (Desimone, 1998; Deco and Lee, 2002) and for visual search tasks (Hamker, 2005b). This approach describes the attentional selection of an item as an emergent process that acts already on the first stages of visual processing and incorporates both bottom-up and top-down inputs.

In the present chapter, I will present a DNF model to explain the experimental findings of Hollingworth and colleagues. The model covers early visual processing, attentional selection, formation and maintenance of VWM, and planning and execution of saccadic eye movements. This work builds on and combines previous DNF models of VWM (Johnson et al., 2009a,b) and saccade planning (Kopecz and Schöner, 1995; Trappenberg et al., 2001). At the core of the model, however, lies another mechanism: The association between representations over different feature spaces using multi-dimensional DNFs. It is used here to reach a coupled attentional selection of a visual item in terms of both its surface features and its location, and to incorporate both color and spatial biases in this selection.

The DNF mechanism underlying this attentional selection is also central for other DNF architectures in which representations over different feature dimensions have to interact with each other. It directly touches on what is know as the *binding problem* in cognitive science, addressing the question of how neural representations of object features that are distributed over different cortical areas can create coherent object percepts. The psychophysical experiments treated in this chapter form an excellent test case for this mechanism, since they provide a variety of both categorical and metric effects of interactions, and they show interaction effects in different directions.

The full DNF model was tested on a variant of the original experimental study described above. The psychophysical experiment was designed collaboratively and conducted by Andrew Hollingworth. After fitting of the parameter values, the model was capable of emulating the experimental task, and it quantitatively reproduced the empirical results on saccade target selection, saccade amplitudes, and saccadic reaction time in different experimental conditions. These results provide strong support for the proposed interaction mechanism. Moreover, the model made specific predictions about the mechanism underlying observed variations of color memory performance in the empirical results. These predictions were tested in a separate experiment, and found confirmed.

Below, I will first motivate and explain the core mechanism of feature associations in DNFs in general terms. I will then describe the specific model used to address the experimental findings on interactions between VWM and saccade planning, referred to hereafter as the *biased competition model*. I will present simulation results from this model and compare them to the experimental data. Finally, I will discuss these results and the broader scope of the model, and give comparisons to other neurally inspired models. The

biased competition model and the results presented here have been published in a journal article (Schneegans et al., 2014), and the introduction of the general mechanism for feature associations follows the description published in Schneegans et al. (in press a).

3.2 Mechanism of feature association

3.2.1 Multi-dimensional DNFs

Multi-dimensional DNFs are DNFs that are defined over a feature space with more than one dimension. The field equation given in the previous chapter can be extended to multiple dimensions in a straightforward manner, and all attractor states previously introduced are retained. Due to the computational load imposed by sampling a high-dimensional space, the dimensionality that is feasible for numeric simulations is quite limited. In the models described in this thesis, only DNFs with no more than four dimensions are employed. Analogous limitations also exist for biological population code representations: The number of neurons required to adequately sample the underlying space with their tuning curves quickly becomes prohibitively high as the dimensionality of the feature space increases. It is therefore advantageous both in the models and in biological systems to use low-dimensional representations whenever possible, and only to employ higher-dimensional representations where they are necessary to achieve a certain function.

The space that is spanned by a multi-dimensional representation may be composed of multiple dimensions of the same type, or it may combine qualitatively different feature spaces. A biological example of the first situation can be found in maps over two-dimensional visual space. Such maps exist for instance in the superior colliculus (SC) in the midbrain, which is involved in selecting the target for saccadic eye movements. In a DNF model of such a representation, the interaction weights can be defined in a straightforward fashion as function of the distance in feature space (e.g., the Cartesian distance between two points in two-dimensional visual space).

A prominent example of inhomogeneous feature spaces in a biological population code representations can be found in the primary visual cortex. Neurons in this area show both a confined spatial receptive field, and a preference for certain surface feature values such as a specific orientation (Hubel and Wiesel, 1962). The population as a whole samples the full multi-dimensional feature space (Blasdel, 1992). Multiple intertwined maps for different surface features, including orientation, spatial frequency, and color, have been identified in the visual cortex (Issa et al., 2000; Livingstone and Hubel, 1984). I will use such maps that combine visual space with one surface feature as the prime example for DNFs with multiple inhomogeneous dimensions in this chapter. Note that the DNF model does not attempt in

any way to reproduce the physiological arrangement of feature maps in visual cortex, but only describes activation distributions over the abstract feature space.

To describe the lateral interactions in such a DNF, one has to define a metric on the combined feature space (at least implicitly). There is generally no natural conversion factor between distances in different feature spaces (for instance, a spatial distance cannot be translated into a distance in hue value). Therefore, the metric has to be based either on empirical data for synaptic connection strength in biological neural populations, or on the observed or the desired behavioral signatures of lateral interactions. The interactions can then be implemented in the same way as in fields with multiple dimensions of the same type. For instance, a DNF over two qualitatively different feature dimensions with Mexican-hat style lateral interactions can be defined by the equation

$$\tau \dot{u}(x,y) = -u(x,y) + h + i(x,y) + \iint k(x-x', y-y') f(u(x',y')) dx' dy' \quad (3.1)$$

with an interaction kernel

$$\begin{aligned} k(x,y) = & \frac{c^{\text{exc}}}{2\pi \sigma_x^{\text{exc}} \sigma_y^{\text{exc}}} \exp\left(\frac{x^2}{2\sigma_x^{\text{exc}}} + \frac{y^2}{\sigma_y^{\text{exc}}}\right) \\ & - \frac{c^{\text{inh}}}{2\pi \sigma_x^{\text{inh}} \sigma_y^{\text{inh}}} \exp\left(\frac{x^2}{2\sigma_x^{\text{inh}}} + \frac{y^2}{\sigma_y^{\text{inh}}}\right). \end{aligned} \quad (3.2)$$

The general notation is the same here as introduced in the previous chapter, with dependence on time omitted for brevity. The width parameters σ_x and σ_y set the relative scaling of distances in the two feature spaces x and y, and determine how broad or sharp the interactions are in each feature dimension. This relationship between the distances in each feature dimension may be different for different fields in an architecture, to reflect the different functional requirements for the field.

The field equation for the general case of a multi-dimensional field can be given in vector notation as

$$\tau \dot{u}(\vec{x}) = -u(\vec{x}) + h + i(\vec{x}) + [k * f(u)](\vec{x}). \quad (3.3)$$

Here, I use the notation $[k * f(u)]$ to denote an n-dimensional convolution between the output of an n-dimensional field and an interaction kernel of the same dimensionality.

3.2.2 Coupling between fields of different dimensionality

To create DNF architectures that can generate complex behaviors in an efficient way, it is often necessary to couple fields of different dimensionality

along feature dimensions shared by both fields. I will describe the basic operations needed for this coupling. As a concrete example, I will consider an architecture of a single two-dimensional field connected to two separate one-dimensional fields, as it will also be used in the biased competition model presented later in this chapter. The two-dimensional field is defined over one spatial dimension and one color dimension (using the circular space of color hue values). This reflects in a simplified form one feature map in visual cortex, omitting one spatial dimension to allow easier description and visualization of the connections and activation patterns. The one-dimensional fields are defined over one of these dimensions each, yielding one purely spatial field and one color field. The architecture is shown in Figure 3.1, with the one-dimensional fields aligned with the matching feature spaces of the two-dimensional field.

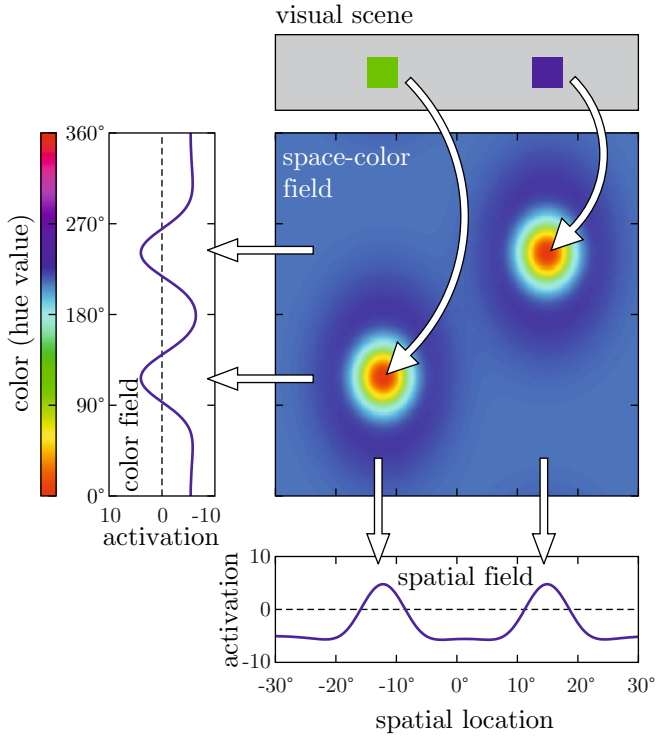


Figure 3.1: Read-out operation from a two-dimensional DNF into separate one-dimensional DNFs. One-dimensional DNFs are shown as blue activation plots over their respective feature space, activation levels in the two-dimensional DNF are color-coded (red indicating highest, dark blue lowest activation; yellow marks the output threshold at an activation level of zero). Curved arrows indicate localized inputs from visual stimuli, straight arrows indicate active projections between DNFs.

I assume that the two-dimensional space-color field receives excitatory input that reflects the presence of visual stimuli. Each individual colored stimulus in the visual scene is reflected by a single localized input, modeled for instance by a Gaussian pattern over the two-dimensions (as shown in the activation pattern of Figure 3.1). The input position along the spatial dimension reflects the stimulus location, its position on the feature dimension reflects stimulus color. The activation distribution induced by these inputs is then further modulated by lateral interactions within the field. Here, I will assume that these interactions comprise local excitation and local surround inhibition of moderate strengths. These interactions create stabilized activation peaks from the localized inputs, but do not create any selection or working memory effects within the space-color field.

The first type of connection between fields to be implemented in this architecture is now a projection from the two-dimensional space-color field to each of the one-dimensional fields. I will refer to this kind of projection (from higher- to lower-dimensional fields) as a *read-out*. It is implemented by integrating the field output of the higher-dimensional field over the disregarded dimension. Thus, to determine the input to the spatial field, the output of the space-color field is integrated over the color dimension; to determine the input to the color field, the output is integrated over the spatial dimension. Before these integrals are fed into the one-dimensional field, they are first smoothed by a convolution with a Gaussian interaction kernel. This operation reflects the synaptic spread that is found in projections between different cortical areas. The smoothing also counteracts the effects of the sigmoid output function, which can produce plateau-like patterns in the output distribution at the positions of activation peaks. The convolution with a Gaussian turns these back into smooth profiles with a localized maximum at the center of each peak. The result of this readout is shown in Figure 3.1, where the color field shows two activation peaks reflecting the colors of the two present visual stimuli, the spatial field showing two peaks that reflect only their spatial locations.

This DNF architecture with the read-out operation can be described by the following set of differential equations:

$$\dot{u}_v(x,y) = -u_v(x,y) + h_v + [k_{vv} * f(u_v)](x,y) \tag{3.4}$$

$$\dot{u}_s(x) = -u_s(x) + h_s + [k_{sv} * F_s(u_v)](x) + [k_{ss} * f(u_s)](x) \tag{3.5}$$

$$\dot{u}_c(x) = -u_c(x) + h_c + [k_{cv} * F_c(u_v)](x) + [k_{cc} * f(u_c)](x) \tag{3.6}$$

The terms $F_s(u_v)$ and $F_c(u_v)$ describe the output of the space-color field integrated over one dimension:

$$F_s(u_v)(x) = \int f(u_v(x,y))dy \tag{3.7}$$

$$F_c(u_v)(y) = \int f(u_v(x,y))dx \tag{3.8}$$

Here, u_v is the activation of the space-color field, u_s the spatial field activation, and u_c the color field activation. The lateral interaction kernel k_{vv} is a difference of Gaussians as in Eq. 3.2. The kernels k_{sv} and k_{cv} that mediate the read-out projection from the two-dimensional field to the spatial and color field are one-dimensional Gaussian functions. The lateral interaction kernels in the one-dimensional fields, k_{ss} and k_{cc}, may be difference-of-Gaussian kernels or Gaussians with a global inhibitory term.

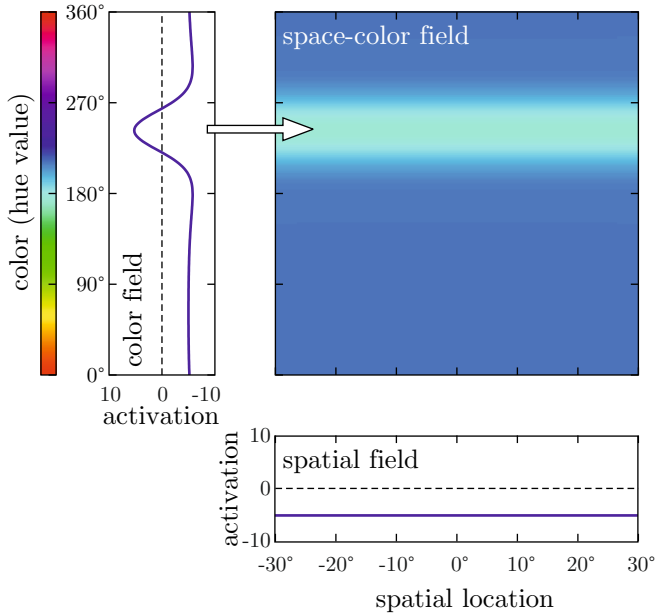


Figure 3.2: Ridge input from a one-dimensional DNF to a two-dimensional DNF. The induced activation pattern is localized along the shared feature dimension, and homogeneous along the dimension not covered by the input field.

In the reverse projection, each of the one-dimensional fields can provide input to the two-dimensional space-color field. Let us consider this projection first for the color field, and let us assume that a single activation peak is present in this field, produced by an external input (e.g., a top-down input reflecting the concept "blue"). The space-color field does not have any activation peaks for now (e.g., because there are presently no salient visual stimuli). This scenario is shown in Figure 3.2. The color field now projects to the space-color field. Since the representation in the color field does not specify any spatial locations, this input cannot be localized along the spatial dimension. Instead, it is a ridge of activation, localized in the color dimension, but homogeneous along the spatial dimension. Analogously, the spatial field can project an input to the space-color field that is homogeneous

along the color dimension.

This type of projection can be implemented by extending the field equation of the space-color field given above as follows:

$$\begin{aligned}\dot{u}_v(x,y) = &- u_v(x,y) + h_v + [k_{vv} * f(u_v)](x,y) \\ &+ [k_{vs} * f(u_s)](x) + [k_{vc} * f(u_c)](y)\end{aligned} \tag{3.9}$$

As in the read-out projection, the field output from the source fields is convolved with an interaction kernel, typically a Gaussian function, before being fed into the target field.

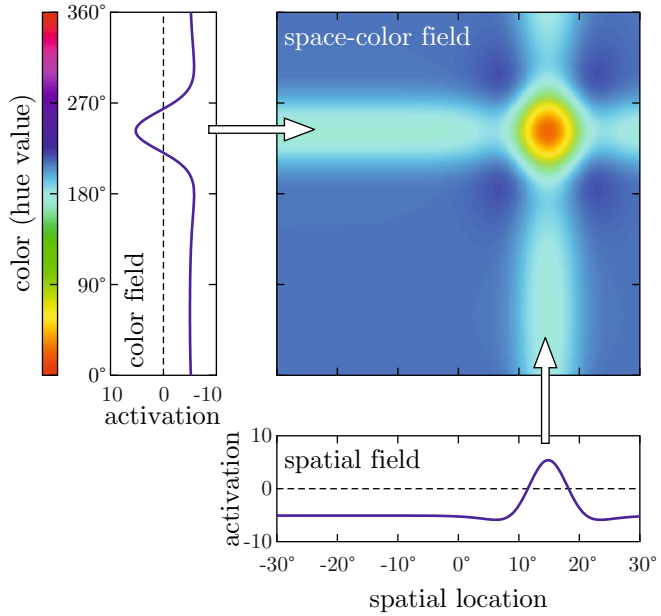


Figure 3.3: Induction of a localized activation peak in a two-dimensional DNF at the intersection point of two ridge inputs from separate one-dimensional DNFs.

An individual ridge input should typically not induce an activation peak in the two-dimensional field, since it does not fully specify a location for the peak. Although the lateral interactions in the two-dimensional field can be set up to force the formation of a single localized peak from a ridge input, the position of that peak along the ridge would be random. Therefore, the connection weights (in the kernels mediating the projection) are generally chosen to produce only sub-threshold activation in the target field. These sub-threshold activation ridges can be combined with other inputs to produce or modulate localized peaks, since the different inputs to a field are added up in the field equation. They can strengthen peaks that were induced by

localized input, such as external visual input in the examples used here. Alternatively, two ridges may be combined to form a localized peak at their intersection point, as shown in Figure 3.3. This will be used in the following chapters to create integrated representations from separate inputs.

3.2.3 Motivation for using multi-dimensional DNFs

The simple DNF architecture used above illustrates that multi-feature items can be represented in two different ways in DNF models: Either in a single field defined over the combined feature space (here, stimulus color and location), or in a distributed fashion in separate one-dimensional fields. These different types of representations can be transformed into each other using the operations described above. In terms of neural or computational resources, the separate one-dimensional fields have a significant advantage, as they provide a much more compact representation. For instance, if each feature dimension is sampled with 100 neurons (or 100 sampling points in numerical simulations of a neural field), it takes a total of 200 neurons to create the separate one-dimensional fields, whereas it would require $100 \cdot 100 = 10000$ neurons to sample the full two-dimensional field. And still, the separate one-dimensional fields can represent the individual features values with the same precision as a single combined field.

However, the representation in separate one-dimensional fields is missing one critical aspect: It does not reflect the conjunction of different features if multiple values are represented in each field. Consider again the case depicted in Figure 3.1, with a green stimulus on the left and another blue stimulus on the right. The representations in the one-dimensional fields can provide the information that there is green and a blue item in the visual scene, and that one of them is to the left and the other to the right. But one cannot tell from the peaks in these fields which color belongs to which location. This can be made explicit when each one-dimensional field projects two ridge inputs back into the two-dimensional field, as shown in Figure 3.4. In this example, four intersection points form, one for each possible combination of a color and a spatial location, even though originally only two stimuli induced the activation patterns in the one-dimensional fields.

This problem of knowing and maintaining the correct coupling between the features of multiple items in a distributed representation is one instance of the binding problem in cognitive science (more precisely, it is a form of the problem of feature binding; Treisman, 1996). Knowing the locations of specific surface features and, likewise, the conjunction between features is critical for many aspects of goal-directed action. If you want to reach for an apple, for instance, it is not sufficient to know that there is a red item in the visual scene along with a blue and a yellow one, and that one of them is to the right, one in the middle, and on to the left. To plan the correct reaching movement, you need to know at which position the red object is.

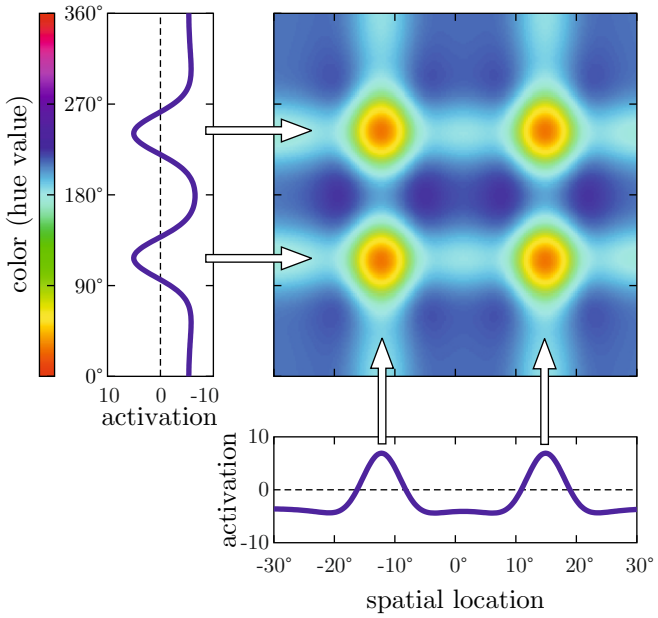


Figure 3.4: Intersections of multiple ridge inputs in a two-dimensional DNF. Peaks form at all intersection points, reflecting all possible combinations of peak locations in the one-dimensional DNFs.

On the other hand, many processes do not depend on feature conjunctions, and rely only on a specific subset of feature dimensions. For instance, in the neural systems that drive the planning and control of reaching movements, surface features such as color are no longer relevant once the correct target is selected. It is then in fact desirable to employ purely spatial representations in these systems that can generalize over the irrelevant features (so that one does not have to learn the reaching movement repeatedly for reach targets of all possible colors). Likewise, in processing feature information, invariance to spatial position is often advantageous. In object recognition, it is desirable that objects can be identified independent of their location in the visual field.

A specialization of cortical areas according to these principles appears to exist in the brain, as maintained by the Two-Paths Hypothesis (Mishkin et al., 1983). While early visual areas reflect both spatial and surface feature aspects of the visual input, later visual areas show a division into a ventral and a dorsal processing stream. Cortical visual areas in the ventral stream are predominantly involved with the processing of surface feature information and object identity, with increasing spatial receptive field size (thus decreasing spatial selectivity) along the pathway (Krüger et al., 2013).

Areas in the dorsal stream, in contrast, predominantly process spatial and motion-related aspects of the visual input, which are relevant in particular for the planning of movements.

Given these advantages of representations specialized for either surface features or spatial representations, and the evidence that such representations are common in the brain, the question is how to overcome the binding problem described above. One suggested solution, formulated in the Feature Integration Theory (Treisman and Gelade, 1980), is based on the simple notion that no mis-bindings can occur if only a single item is present in each separate low-dimensional representation. In such a situation, the separate low-dimensional representations are unambiguous even without the additional combined representation. It is obviously not reasonably to demand that only a single item can be present in the visual scene at any time. Therefore, to enforce this requirement, a mechanism is needed that selects a single item in the visual scene, and that ensures that the same item is reflected in each of the separate low-dimensional representations. This mechanism can be provided by visual attention (Reynolds and Desimone, 1999). The implementation of such a mechanism in the DNF architecture is described in the following section.

3.2.4 Visual search and attentional selection

The attentional selection of one visual item in both one-dimensional fields can be achieved in the DNF model by combining the mechanisms of ridge projection and read-out. It can be driven either by the bottom-up visual input alone, or can be dominated by top-down input specifying the color or location that attention should be directed at. Visual search is an example of the second case (compare Hamker, 2005b): A target feature is specified—here, a color—and the location of a matching visual item has to be selected. The basic DNF model of this task is shown in Figure 3.5. Visual stimuli provide localized inputs to the space-color field and induce activation peaks. The target feature is provided as an external input to the one-dimensional color field. This field now projects a ridge input into the space-color field. If the ridge overlaps with one of the localized peaks, this peak is strengthened. Its activation level and extent are increased, and it consequently produces a higher total output.

A read-out operation is now applied along the spatial dimension, providing input to the one-dimensional spatial field. A larger activation peak in the space-color field will produce a larger input to the spatial field. Note that the Gaussian convolution in the read-out operation is necessary to fully obtain this effect: The sigmoid function used to compute the output from the activation at each point in the field typically saturates at moderate activation values, so further increasing the activation level of a peak does not necessarily yield a higher output at a single point in the field. But the

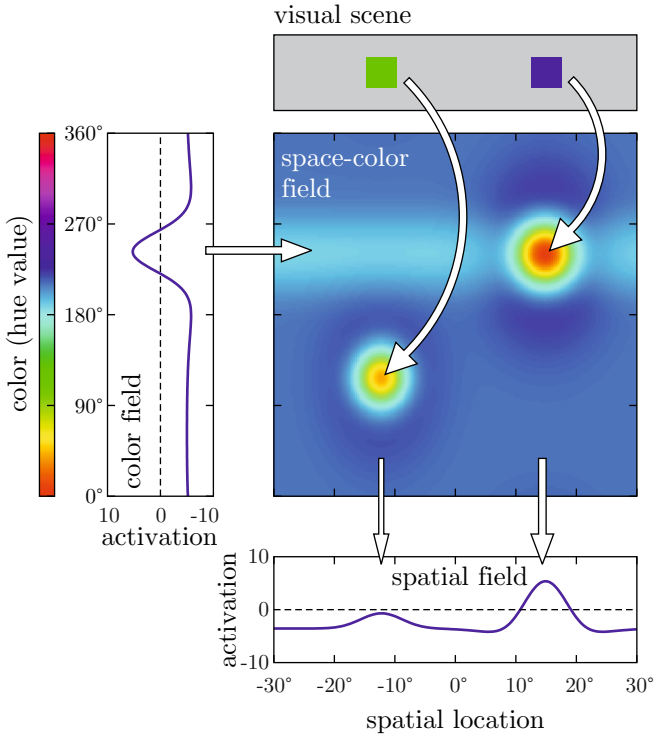


Figure 3.5: Implementation of a simple feature search mechanism in a DNF architecture. The search template is given in the color field, the location of a matching visual stimulus is read out into the spatial field.

stronger activation peaks also produce output over a larger area. If this wider output is smoothed by a Gaussian convolution, the resulting input to the target field is not only wider, but also reaches higher values.

Within the one-dimensional spatial field, a selection of a single location can be produced by implementing competitive lateral interactions (with local self-excitation and global inhibition). The field will then always form a single activation peak at the location of the strongest input it receives from the space-color field. In the case that a single visual stimulus matches the target color, this item is almost certain to be selected. If there are multiple matches, the selection is based on both the goodness of the color match and the size and saliency of the matching stimuli. In the case that no stimulus matches the target color, a random selection of one stimulus location may take place. This is problematic with respect to the binding problem because in this case, the representations in the two one-dimensional fields do not correspond to the features of a single visual item. I will return to this case below.

An analogous mechanism can be used for the inverse task: Given a spa-

tial location, determine the color of a visual stimulus at this location (in a larger system, this can serve as a first step to identify a spatially selected object). To solve this task, the target location is fed as external input into the spatial field, and projects a ridge input into the space color field (this would be a vertical ridge when plotted as in Figure 3.5). This ridge strengthens matching activation peaks induced by visual input, and the color of an item thus highlighted can be determined via a read-out along the color dimension. This read-out then provides an input into the one-dimensional color field, and competitive lateral interactions can ensure the formation of a single activation peak.

Of course, in a biological neural system, switching between these two tasks cannot be achieved by inverting the directions of the projections between the different neural representations. These projections reflect synaptic connections between neurons, and these do not switch directions or significantly alter their structure over brief durations. But behavioral flexibility can be achieved in an other way, namely by combining the connection patterns from these two tasks into a single architecture. The projections between the one-dimensional fields and the single two-dimensional fields then become bidirectional: Both the the spatial field and the color field receive input from the space-color field via a read-out operation, and project ridge inputs back to it. Competitive lateral interactions within the one-dimensional field drive selection decisions in both of them.

Let us first consider the behavior of such a system when only the space-color field receives external input, reflecting a visual scene with several stimuli. The spatial field and the color field do not receive any external input. This case is shown in Figure 3.6. As soon as peaks form in the space-color field, they provide inputs to both the spatial field and the color field, via the separate read-out projections (Figure 3.6a). When this input drives activation levels locally beyond the output threshold in each of these fields, two things happen: Through the lateral interactions, the emerging activation peaks compete with each other within each one-dimensional field. At the same time, these nascent peaks start to project ridge inputs back into the two-dimensional space-color field. When one of the peaks—say, in the spatial field—now grows a little stronger than its competitors, it not only gains an advantage in the selection process in its own field, but it also produces a stronger ridge input. This strengthens the corresponding peak in the space-color field, and in turn increases the input for the associated feature value in the color field. The competition in the color field is thereby biased to select the same visual item as is selected in the spatial field.

This effect acts in both directions and thereby reinforces itself. With indirect back-and-forth projections between them, a joint selection of the same visual stimulus in the spatial field and the color field is promoted (Figure 3.6b). This selection can furthermore be biased by additional inputs

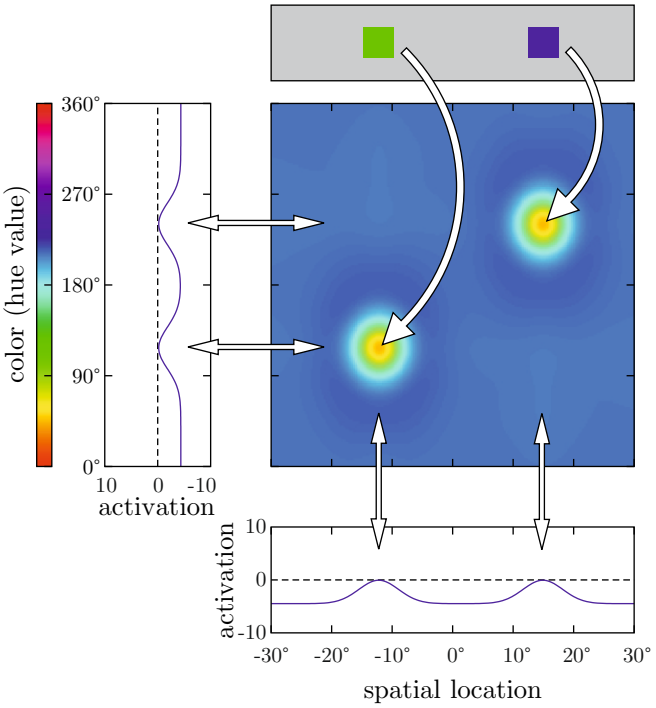
to the spatial or color field—or to both of them. A relatively small additional input can tip the scales to select preferentially items of a certain color or at a certain location. This also solves the problem raised above of having an inconsistent selection in a visual search task when no item matches the target feature: If the input for the search target is relatively weak, it can bias the outcome of the search when there is a matching item in the scene, but still allow the coherent selection of another stimulus if there is no match.

For the joint selection in spatial and color field to work reliably, certain constraints must be met by the interaction parameters within and between fields. In particular, there must be a certain balance between the competitive interactions in each one-dimensional field and the indirect biasing effects that these two fields exert on each other. The competitive interactions must be strong enough to push the system as a whole toward a selection decision, but they must not be so strong that they drive the fields to a selection independent of each other (e.g., by immediately forcing a selection of one input location in each field if it shows a slightly higher activation due to random noise). The strength of the indirect coupling between the fields is a result of the projection strengths into the visual sensory field, strengths of the read-out projections, and lateral interaction effects within the visual sensory field. Balancing the two effects can be difficult (and no analytical solution for determining appropriate parameter values exists to my knowledge).

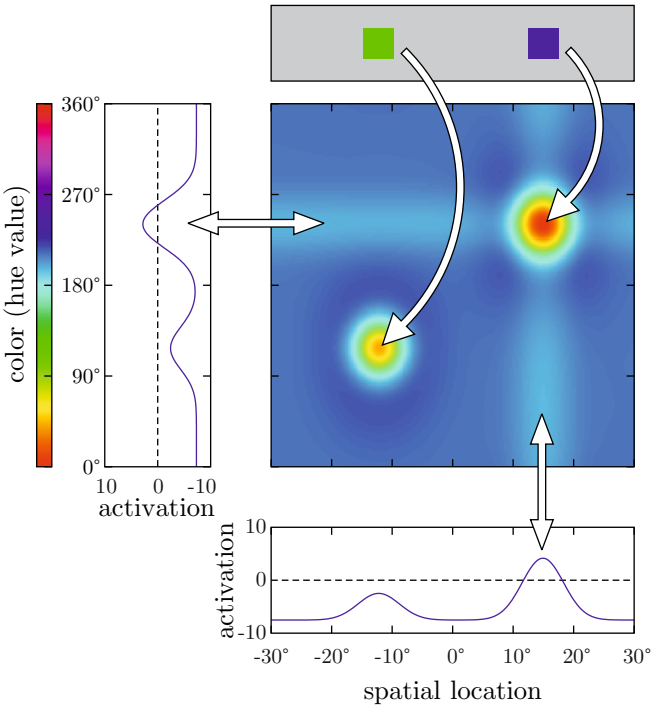
One way to ensure that the same visual item is selected in both one-dimensional fields is to make one of them dominant over the other. In the biased competition model described below, and also in the model of scene representation treated in Chapter 5, the selection behavior is stronger in the spatial field than in the color field (and other surface feature fields in the scene representation). Thus, even if a mismatched selection has occurred in both fields simultaneously, the stronger modulatory influence of the spatial field can override the selection in the color field and enforce a matched selection of a single visual stimulus. Such dominance of spatial representations in attentional selection is also consistent with experimental evidence (Vidyasagar, 1999). The flexibility of the architecture is still retained: Top-down inputs can be used to preshape the activation in the spatial field or the color field (or both of them), and in both cases the preshaping creates biases in the selection decision.

Another, more general issue when coupling fields together with bidirectional projections is that activation levels must be held in check. In the visual attention mechanism, once a single item has been consistently selected, the activation peaks in the individual fields mutually excite each other. This can potentially lead to a self-reinforcing growth of activation levels in all fields. It can be counteracted by adjusting the levels of lateral inhibition within each field. Since the range of lateral inhibition is higher than that of the excitatory interactions (both lateral and between fields), the inhibitory

(a)


360°
270°
180°
90°
0°
color (hue value)
10
0
-10
activation
10
0
-10
activation
-30°
-20°
-10°
0°
10°
20°
30°
spatial location

(b)


360°
270°
180°
90°
0°
color (hue value)
10
0
-10
activation
10
0
-10
activation
-30°
-20°
-10°
0°
10°
20°
30°
spatial location

interaction effects can become dominant once the activation peaks exceed a certain width, and an excessive growth of activation can be prevented.

3.3 DNF model of biased competition

In this section I will describe the extension of the basic association mechanism discussed so far to a new model architecture that can account for experimentally observed interactions between VWM and saccade planning. The model is used to emulate an experiment that was performed by Andrew Hollingworth and colleagues: Participants were facing a computer screen on which visual stimuli were displayed. They performed two tasks in an interleaved fashion. The first was a working memory task, in which participants had to memorize the color of a stimulus and were later tested on it. The second was a saccade task, in which participants had to make timed eye movements toward sudden-onset stimuli. The details of the experimental procedure, the different stimulus conditions, and the obtained results are given below after the description of the DNF model.

The stimuli used in the experiment were colored disks and squares for the saccade targets and memory cues, and a white cross as fixation point. Stimuli were always aligned on a horizontal axis at the center of the screen. It is therefore possible to retain the simplification that only a single spatial dimension is modeled (instead of the full two-dimensional visual space), while still capturing all behaviorally relevant aspects of the visual stimuli. I will first describe the DNF architecture verbally, the field equations and parameter values are then given at the end of this section.

3.3.1 Model architecture

The architecture for the biased competition model is shown in Figure 3.7. The core of the model is analogous to the space-feature association mechanism explained before, comprised of a single two-dimensional field and two one-dimensional fields. To be able to fully emulate the psychophysical experiment, the system is augmented by two additional fields: A new field over the surface feature dimension is introduced to serve as working memory representation. An additional spatial field reflects the saccade motor plan, and its output drives a simulated saccadic eye movement. Further-

---

Figure 3.6 *(preceding page)*: Coupled selection decision in interconnected DNFs. (a) Situation briefly after the onset of two symmetric visual stimuli. (b) Under the influence competitive interactions in the separate one-dimensional DNFs and mutual coupling via the two-dimensional DNF, one of the stimuli has been selected consistently in all fields.

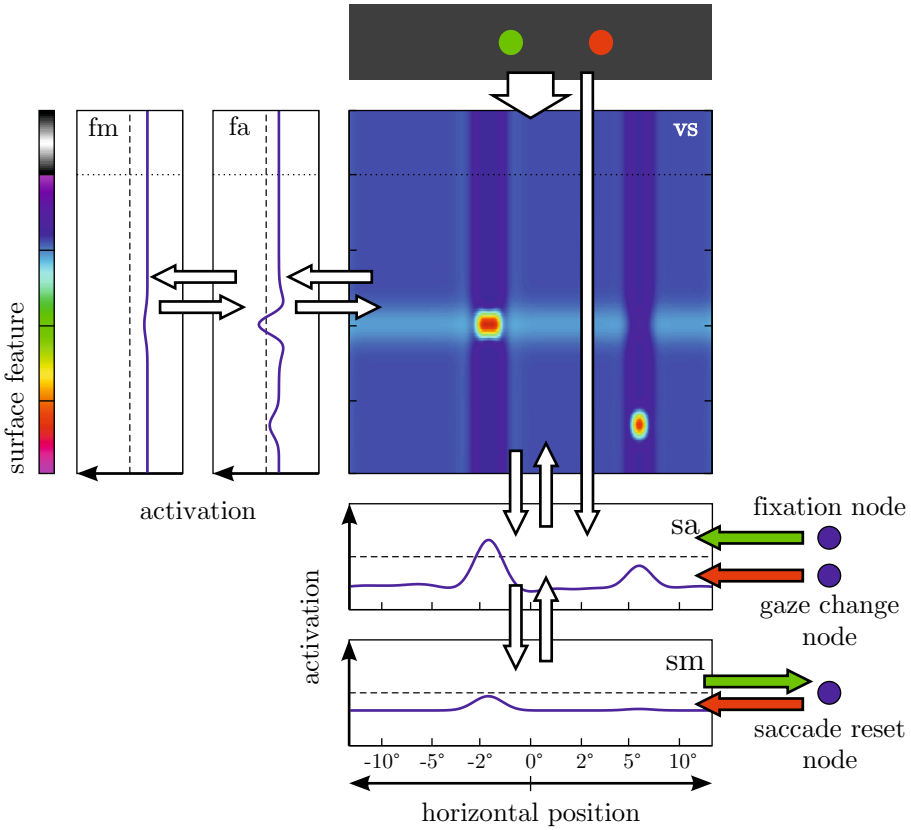


Figure 3.7: DNF model of biased competition. Fields are depicted as in the previous figures, dynamic nodes are shown as blue circles. White arrows indicate topological connections between fields, green arrows denote excitatory connections from or to a dynamic node, red arrows inhibitory connections. Note the logarithmic scaling of the spatial dimension in the model, leading to the different input strengths induced by the two stimuli in the exemplary visual scene (which are equal in size, but differ in eccentricity). Also note that due to this effect, activation peaks are typically not aligned with stimulus positions in the depiction of the visual scene. Abbreviations: vs – visual sensory field, fa – feature attention field, fm – feature WM field, sa – spatial attention field, sm – saccade motor field.

more, three discrete nodes are introduced to the architecture to modulate saccade behavior and terminate saccadic eye movements. These nodes are intended to model groups of neurons that do not encode metric information (and therefore cannot be described as an activation distribution over a feature space).

The spatial dimension in this model reflects horizontal stimulus position

in a retinocentric frame of reference, such that spatial representations shift when a simulated eye movement occurs. The scaling along this dimension is logarithmic, such that the resolution is highest at the center of the spatial fields (the foveal region in retinocentric space) and falls off in the periphery. In effect, a stimulus that appears near the fovea produces a stronger and wider activation peak than would be produced by the same stimulus at a more extrafoveal location. In the surface feature dimension, a color representation (over hue values) is augmented by a separate section that represents gray values. This is necessary to accommodate for the white fixation point in the experiment. This implementation reflects the minimal assumption that surface features of non-colored items are represented in an analogous fashion to color features, and are subject to the same effects of feature attention.

The simulated visual input drives activation in the two-dimensional *visual sensory field*. Individual stimuli induce localized activation peaks in this field, reflecting the combination of stimulus location and color value. Lateral interactions of moderate strength stabilize these peaks against random fluctuations in field activation. The lateral inhibition along the surface feature dimension is global, such that only a single feature value can be represented for each spatial position (see inhibitory grooves in the visual sensory field in Figure 3.7).

In the surface feature pathway, the visual sensory field projects into the *feature attention field*, via a read-out projection that integrates over the spatial dimension. The lateral interactions in this field include local excitation and global inhibition to create a competition effect. The feature attention field projects back to the visual sensory field, inducing ridge inputs along the spatial dimension. It also provides feed-forward input for the *feature working memory (WM) field*, which spans the same dimension of surface features. The feature WM field in turn projects back to the feature attention field. Lateral interactions in the feature WM field are set up with strong self-excitation and strong surround inhibition so that they support self-sustained activation peaks that serve as working memory representations.

In the spatial pathway, the *spatial attention field* receives the spatial read-out of the visual sensory field as input, and projects back to it. The field also receives direct visual input (only the spatial stimulus components) to reflect the neurophysiology of the saccade system in the brain (see below). During the saccade task, the field activation is furthermore preshaped to reflect the expectation of target locations and the task instructions, with activation level increased around the expected location of the saccade target and suppressed around the distractor locations. This preshaping corresponds to cognitive inputs reflecting intentions and expectations (compare Trappenberg et al., 2001), whose source lies outside the scope of the current model.

The spatial attention field is bidirectionally coupled to the *saccade motor field*, providing feed-forward input and receiving feedback from it. The feed-

forward input is suppressed in the foveal region (at the center of the two fields), such that stimuli that are already being fixated cannot induce a saccadic eye movement. Both fields feature lateral interactions with a global inhibitory component, creating a selection regime. The interaction strength is moderate in the spatial attention field, such that two peaks may coexist for some time if they are both supported by external input to the field. The interactions in the saccade motor field are strongly competitive, ensuring that only a single peak can persist at any time, and thus that a unique target location is selected for a saccadic eye movement.

Two dynamic nodes project with a fixed weight pattern to the spatial attention field to modulate the system's saccade behavior: The *fixation node* excites the central (foveal) region of the field. Since this region does not project to the saccade motor field, but does compete with other active regions in the spatial attention field through lateral interactions, its effect is to suppress saccade initiation and stabilize fixation. The *gaze change node* has the opposite effect, it suppresses activation in the foveal region and thereby facilitates shift of attention and saccade initiation to peripheral stimuli. Finally, the *saccade reset node* receives input from the whole saccade motor field and projects inhibition back to it. It has the role of terminating the eye movement signal, as detailed below.

3.3.2 Saccade generation

The DNF model of biased competition produces behavioral responses in the form of saccadic eye movements, which are generated in the model's spatial pathway. As stated above, this spatial pathway is comparable to previous two-layer DNF models of saccade planning in the SC, in particular to the model of Trappenberg et al. (2001; see also Marino et al., 2012; Wilimzig et al., 2006). The top layer (the build-up layer in the model of Trappenberg, corresponding to the spatial attention field here) receives external visual input as well as top-down input reflecting task instructions or expectations about stimulus positions. Different active locations compete in this layer through lateral interactions. When a sufficiently strong peak has formed at one location, it induces an activation peak in the bottom layer as well (burst layer in the model of Trappenberg, saccade motor field here). The formation of an activation peak in this layer is taken as a signal that a saccade is initiated, with the peak location specifying the selected saccade target.

What previous models lacked was a treatment of the further evolution of activation patterns after a saccade motor peak had formed. To produce the detailed saccade metrics required to fully account for psychophysical data, it is necessary to also consider the generation of the saccadic motor command from the activation peak and the termination of the saccade. Furthermore, the model should be able to perform multiple saccades within one trial.

Even though only the characteristics of the first saccade are analyzed in the saccade task of the experiment, the further behavior of the model is relevant for the results in the subsequent memory test.

To terminate the saccade signal, the saccade reset node is introduced. This node acts as a simple (lossy) neural integrator that receives as input the integrated output over the whole saccade motor field, scaled with a constant weight (see Equation 3.20 below). As long as there is no activation peak in that field, the node's activation remains at the resting level. Once a peak forms, node activation rises until it reaches the output threshold. The node's output then provides a strong global inhibitory input to the saccade motor field, extinguishing the present activation peak there. The node features moderate self-excitation, such that it remains active a brief time after the input from the saccade motor field has ceased. This ensures that the peak in the saccade motor field is fully extinguished by the reset node. Ultimately, however, the node activation returns to the resting level without the external input. The output of the saccade reset node is also used to determine the beginning and end of the simulated eye movement: The actual saccade is taken to start when the node output exceeds a threshold θ_{start}, and to end when the output falls back below θ_{end}.

Due to the strong interactions in the saccade motor field, the time course of peak formation and suppression in that field is highly stereotyped. As soon as an external input drives the field activation at some point to the output threshold, the lateral interactions will be dominant in determining the shape (but not the location) of the emerging peak. The activation peak drives the saccade reset node, independent of the location of the peak, which in turn suppresses the peak after a largely fixed time. The lateral interactions thus create a normalizing effect, decoupling the time course of activation patterns in the field from the strength and shape of the external input that it receives. This normalization is not complete, however; a stronger input signal may still create a slightly larger and more enduring saccade motor peak. This will be important in one of the experimental paradigms modeled with this architecture.

The method of determining saccade metrics from the activation peak is inspired by the interpretation of SC motor activity of Goossens and Van Opstal (2006). In their view, every spike of a saccadic burst neuron in the deep layers of the SC contributes to the motor command by adding a "mini-vector" to the saccade metrics. The size and direction of this vector is fixed for each neuron, reflecting the preferred saccadic end point of the neuron and correspondingly its position in the topographic map in the SC (of course, the mini-vector for each spike covers only a small fraction of the distance to the neuron's preferred saccade end point).

This idea is transferred to the DNF model as follows: Each field location is assigned a preferred saccade vector (actually a positive or negative scalar

value, since the model only covers gaze changes in one dimension). To determine the metrics of one saccade in the model, the field output at each location is scaled with this preferred saccade vector. The scaled output is then integrated over the whole field, and over the whole time that an activation peak is present in the field. The result, scaled with a constant conversion factor, is interpreted as a saccadic motor command to the ocular muscles, and yields the final saccade metrics generated by the model. This transformation from the population code in the field to a metric value (space-to-rate code), in combination with the reset mechanism, implements an instance of a *summation with saturation* model according to the classification of saccade models introduced by Groh (2001).

The implementation in the model deviates from the mechanism proposed by Goossens and van Opstal in the way that the saccade is terminated. In the explanation of these authors, a mechanism located downstream from the SC performs a comparison of the generated motor signal with a desired motor signal (both derived from the spiking activity in the SC) and terminates the saccade when the two match. How the termination of the neural activity in the SC is achieved is not specified in this explanation (although the theory would seem to require a precisely controlled termination of that activity, since it maintains that every spike in the SC's deep layers contributes to the saccade metrics). I opted instead for an explicit termination signal provided by the saccade reset node: In this model implementation, the saccade motor peak is always suppressed after an (approximately) fixed time, when a certain total output has been produced by this peak and has activated the reset node, while the metrics of the saccade are determined by the location of the peak in the field. A further downstream comparison process (of two signals derived from the same source) does not appear to me to be capable of improving the precision of the motor command, and thus is omitted in the DNF model.

During a simulated saccade (with start and end time as defined above), all visual input is suppressed in the model. At the end of the saccade, the saccade amplitude is determined, and the system's fixation point within the visual scene is shifted by this value. Stimuli located at the new fixation point now project to the central (foveal) region along the model's spatial dimension, and are represented at higher resolution due to the logarithmic scaling of the spatial dimension in the model.

3.3.3 Biological basis of the DNF model

The DNF architecture for the biased competition task is not intended as a strictly neurophysiological model. The individual DNFs in the model are defined based on functional considerations, and as such, there is no one-to-one mapping of each DNF to a specific brain area. Nonetheless, the model aims to preserve the general structure of the visual processing pathways. In particular, it reflects the division into ventral pathway (focused on visual

surface features and object identity) and a dorsal pathway (focused on spatial aspects of visual perception and movement planning, see Mishkin et al., 1983).

The visual sensory field, which represents both spatial position and one surface feature of visual stimuli, can be equated with the early areas of the visual cortex (V1, V2, and especially V4), before the division into separate pathways. Neurons in these areas have still relatively localized spatial receptive fields, combined with different feature selectivities. In particular in V4, color representations have been found that are consistent with a space code representation of hue values as in the model (Wachtler et al., 2003; Krüger et al., 2013). In addition, neural responses in V4 show pronounced modulation by spatial and feature attention, consistent with the behavior of the field in the model (McAdams and Maunsell, 2000; Reynolds and Chelazzi, 2004). Some experimental support for attentional modulation has also been found for the earlier areas V1 and V2 (Motter, 1993), but the evidence is less clear for these.

The two fields of the spatial pathways in the model are concerned with spatial attention and saccade target selection. In the human brain, there are both sub-cortical and cortical structures involved with these functions, which appear to be partly redundant. The key sub-cortical structure here is the SC. This midbrain structure consists of multiple layers of neurons, each forming a topographically organized map of the visual space in a retinocentric reference frame (Sparks and Nelson, 1987). It receives visual input from the lateral geniculate nucleus (a thalamic structure that conveys the neural signals directly from the eyes) as well as from visual cortical areas such as V1, and also integrates input from other sensory modalities. The superficial layers of the SC represent locations of salient visual stimuli and are involved in attentional selection. The activation patterns in deeper layers are strongly coupled to saccadic motor behavior (and gaze control in general), and localized peaks of activation in these layers have been found to be temporally aligned and causally involved in each saccadic gaze change (Lee et al., 1988; Dorris et al., 1997).

Cortical areas involved in spatial attention and saccade planning are in particular the posterior parietal cortex (PPC) and the frontal eye field (FEF). The PPC is part of the dorsal stream, receiving input from the visual cortical areas. It contains spatial representations with activation patterns that reflect attentional selection (Colby and Goldberg, 1999). Projections from the PPC are on the one hand directed back to the visual cortex, allowing attentional modulation of feed-forward activation patterns, and on the other hand provide input to the FEF. The FEF itself is involved in the control of saccadic eye movements and projects, among others, to the deep layers of the SC (Schall, 2004).

The two spatial fields in the model can be most closely related to the

different layers of the SC (the spatial attention field corresponding to the superficial layers, the saccade motor field to the deep layers). This is also reflected in the direct visual input that the spatial attention field receives in the model (in addition to the input from the visual sensory field), matching the direct thalamic input to the SC. This part of the model architecture is also consistent with previous neurodynamic models of saccade generation (Trappenberg et al., 2001; Marino et al., 2012), which explicitly aimed to reproduce activation patterns in the SC. As a functional model, however, the two DNFs are also meant to incorporate the cortical contributions to attentional selection and saccade initiation, which are functionally similar to the roles of the different layers in the SC.

The two surface feature fields in the DNF architecture model different functional aspects of the ventral processing stream. The feature attention field can be equated with temporal areas such as the infero-temporal cortex (IT). Neurons in this region have large spatial receptive fields that often cover large part of the whole visual field, and thus are spatially unspecific (Krüger et al., 2013). They do show, however, high specificity for certain surface features (such as forms or colors) or combinations of such features.

The representations that form the basis for working memory are believed to be distributed in the brain. To memorize a certain feature, neural populations involved in the perception of that feature are recruited, such as in areas IT and V4 (as evidenced by sustained firing of such neurons in experimental tasks involving VWM; Fuster and Jervey, 1981; Pasternak and Greenlee, 2005). There is also an involvement of the prefrontal cortex, which may act to control what is retained in the sensory areas (Miller et al., 1996). The separate representations of feature attention and feature WM in the model are also consistent with behavioral data showing that VWM content and currently attended features can be dissociated when required by a task (Hollingworth and Hwang, 2013; Houtkamp and Roelfsema, 2006; Olivers et al., 2011).

Several aspects of the the visual processing in the brain are not addressed at all in the model. In particular, the model only deals with a single selected surface feature, and ignores the hierarchical progression from simple to complex features that is found in the ventral pathway. This reflects an intentional simplification on the model, since the focus is on interactions between spatial and feature attention.

3.3.4 Formal description of the DNF model

The full model can be described as a coupled dynamical system governed by a set of differential equations. I will use a unique index to identify each DNF and each dynamic node in these equations: *vs* for visual sensory field, *sa* for spatial attention field, *sm* for saccade motor field, *fa* for feature attention field, *fm* for feature memory field, *fix* for fixation node, *gc* for gaze

change node, and r for saccade reset node. The kernels and parameters of projections between fields are identified by two indices, the first signifying the target, the second the source of the projection. The parameter values of all fields and their lateral interactions are given in Table 3.1, the parameters of interactions between fields are given in Table 3.2. All interactions within and between fields are mediated by generalized difference-of-Gaussians kernels (see Equation 2.4 in the previous chapter) unless specified otherwise. The time constant τ for all field equations is 20 ms. Note that dependence of field activation on time is omitted in the field equations for brevity.

Field equations

field index	h	β	q	c^{exc}	σ^{exc}	c^{inh}	σ^{inh}	c^{gi}
vs (ftr/spt)	-5	1	0.25	10	5 / 2.5	1	- / 6.25	0
fa	-3.5	4	0.25	10	4	18	8	0.1
fm	-5	4	0.5	30	3	37.5	9	0.1
sa	-2	1	0.25	15	12	0	-	0.3
sm	-5	4	0.5	42	8	0	-	0.95
fix	-5	1	0.2	0	-	0	-	0
gc	-5	1	0.2	0	-	0	-	0
r	-5	4	0.2	3	-	0	-	0

Table 3.1: Field parameters and parameters of lateral interactions.

projection index	c^{exc}	σ^{exc}	c^{inh}	σ^{inh}	c^{gi}
fa, vs	0.4	4	0	-	0
vs, fa	3.75	6	0	-	0
fm, fa	2.5	6	0	-	0
fa, fm	8.5	8	0	-	0
sa, vs	1.5	10	1	25	0
vs, sa	2.5	12	0	-	0
sm, sa	7.25	10	0	-	0
sa, sm	7.25	10	0	-	0.1
r, sm	0.4	-	0	-	0
sa, r	0	-	0	-	12
sm, r	0	-	0	-	12
fix, r	0	-	0	-	5
sa, in	1.25	10	0.5	25	0.015

Table 3.2: Parameters of interactions between fields.

The field equation for the visual sensory field can be given as:

$$\begin{aligned}\tau \dot{u}_{\mathrm{vs}}(x,y) = &- u_{\mathrm{vs}}(x,y) + h_{\mathrm{vs}} + i_{\mathrm{vs}}(x,y) + [k_{\mathrm{vs,vs}} * f(u_{\mathrm{vs}})](x,y) \\ &+ [k_{\mathrm{vs,sa}} * f(u_{\mathrm{sa}})](x) + [k_{\mathrm{vs,fa}} * f(u_{\mathrm{fa}})](y) + q_{\mathrm{vs}}\xi(x,y).\end{aligned} \tag{3.10}$$

Here, i_{vs} is the external visual input. The two-dimensional lateral interaction kernel, $k_{\mathrm{vs,vs}}$, featuring local surround inhibition along the spatial dimension and global inhibition along the surface feature dimension:

$$\begin{aligned}k_{\mathrm{vs,vs}}(x,y) = &\frac{c^{\mathrm{exc}}_{\mathrm{vs,vs}}}{2\pi\sigma^{\mathrm{exc,spt}}_{\mathrm{vs,vs}}\sigma^{\mathrm{exc,ftr}}_{\mathrm{vs,vs}}} \exp\left(-\frac{x^2}{2(\sigma^{\mathrm{exc,spt}}_{\mathrm{vs,vs}})^2} - \frac{y^2}{2(\sigma^{\mathrm{exc,ftr}}_{\mathrm{vs,vs}})^2}\right) \\ &- \frac{c^{\mathrm{inh}}_{\mathrm{vs,vs}}}{\sqrt{2\pi}\sigma^{\mathrm{inh,spt}}_{\mathrm{vs,vs}}} \exp\left(-\frac{x^2}{2(\sigma^{\mathrm{inh,spt}}_{\mathrm{vs,vs}})^2}\right).\end{aligned} \tag{3.11}$$

The interaction kernels $k_{\mathrm{vs,sa}}$ and $k_{\mathrm{vs,fa}}$, mediating the ridge inputs from spatial and feature attention fields, are simple Gaussian kernels.

The feature attention field is governed by the differential equation:

$$\begin{aligned}\tau \dot{u}_{\mathrm{fa}}(y) = &- u_{\mathrm{fa}}(y) + h_{\mathrm{fa}} + i_{\mathrm{fa}} + [k_{\mathrm{fa,fa}} * f(u_{\mathrm{fa}})](y) \\ &+ [k_{\mathrm{fa,vs}} * O^{\mathrm{ftr}}_{\mathrm{vs}}](y) + [k_{\mathrm{fa,fm}} * f(u_{\mathrm{fm}})](y) + q_{\mathrm{fa}}\xi(y)\end{aligned} \tag{3.12}$$

It receives input from the visual sensory field, computed by integrating the field output over the spatial dimension, $O^{\mathrm{ftr}}_{\mathrm{vs}}(y) = \int f(u_{\mathrm{vs}}(x,y))dx$.

The field equation for the feature WM field is:

$$\begin{aligned}\tau \dot{u}_{\mathrm{fm}}(y) = &- u_{\mathrm{fm}}(y) + h_{\mathrm{fm}} + i_{\mathrm{fm}} + [k_{\mathrm{fm,fm}} * f(u_{\mathrm{fm}})](y) \\ &+ [k_{\mathrm{fm,fa}} * f(u_{\mathrm{fa}})](y) + q_{\mathrm{fm}}\xi(y).\end{aligned} \tag{3.13}$$

Here, i_{fm} is a global excitatory control input that determines when new activation peaks can form in the field.

In the spatial pathway, the spatial attention field is governed by the field equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{sa}}(x) = &- u_{\mathrm{sa}}(x) + h_{\mathrm{sa}} + i_{\mathrm{sa}}(x) + p_{\mathrm{sa}}(x) + [k_{\mathrm{sa,sa}} * f(u_{\mathrm{sa}})](x) \\ &+ [k_{\mathrm{sa,vs}} * O^{\mathrm{spt}}_{\mathrm{vs}}](x) + [k_{\mathrm{sa,sm}} * f(u_{\mathrm{sm}})](x) \\ &+ W_{\mathrm{sa,fix}}(x) f(u_{\mathrm{fix}}) - W_{\mathrm{sa,gc}}(x) f(u_{\mathrm{gc}}) - c^{\mathrm{gi}}_{\mathrm{sa,r}} f(u_{\mathrm{r}}) + q_{\mathrm{sa}}\xi(x).\end{aligned} \tag{3.14}$$

The field receives direct visual input i_{sa} (purely spatial) and is modulated during the saccade task and memory test task by constant preshape p_{sa} reflecting task instructions and prior knowledge. It also receives spatial input from the visual sensory field, computed by integrating over the surface feature dimension, $O^{\mathrm{spt}}_{\mathrm{vs}}(x) = \int f(u_{\mathrm{vs}}(x,y))dy$. Inputs from the fixation node

and gaze change node modulate activation in the foveal region of the field (around zero) through weight patterns

$$W_{\text{sa,fix}}(x) = 2.25 \cdot \exp\left(-\frac{x^2}{2(\sigma^{\text{exc}}_{\text{sa,sa}})^2}\right) \tag{3.15}$$

and $W_{\text{sa,gc}} = -W_{\text{sa,fix}}$.

These two nodes are driven only by external control inputs reflecting task instructions, yielding the simple dynamic equations:

$$\tau \dot{u}_{\text{fix}} = -u_{\text{fix}} + h_{\text{fix}} + i_{\text{fix}} + c^{\text{gi}}_{\text{fix,r}} f(u_{\text{r}}) + q_{\text{fix}}\xi \tag{3.16}$$

$$\tau \dot{u}_{\text{gc}} = -u_{\text{gc}} + h_{\text{gc}} + i_{\text{gc}} + c^{\text{gi}}_{\text{fix,r}} f(u_{\text{r}}) + q_{\text{gc}}\xi \tag{3.17}$$

Both the spatial attention field and the nodes are suppressed by inhibitory input from the saccade reset node during a gaze change, expressed through the terms $c^{\text{gi}}_{\text{fix,r}} f(u_{\text{r}})$ with inhibitory connection weight $c^{\text{gi}}_{\text{fix,r}}$.

The field equation for the saccade motor field is

$$\begin{aligned}\tau \dot{u}_{\text{sm}}(x) = &- u_{\text{sm}}(x) + h_{\text{sm}} + [k_{\text{sm,sm}} * f(u_{\text{sm}})](x) \\ &+ [k_{\text{sm,sa}} * o^{\text{fov}}_{\text{sa}}](x) - c^{\text{gi}}_{\text{sm,r}} f(u_{\text{r}}) + q_{\text{sm}}\xi(x).\end{aligned} \tag{3.18}$$

In the input from the spatial attention field, the foveal region is suppressed (so no saccade signal will be created for already fixated stimuli), yielding

$$o^{\text{fov}}_{\text{sa}}(x) = \left(1 - \exp\left(-\frac{x^2}{2(\sigma^{\text{exc}}_{\text{sm,sa}})^2}\right)\right) f(u_{\text{sa}}(x)). \tag{3.19}$$

The dynamics of the saccade reset node are described by the equation

$$\tau \dot{u}_{\text{r}} = -u_{\text{r}} + h_{\text{r}} + c^{\text{exc}}_{\text{r,r}} f(u_{\text{r}}) + c^{\text{exc}}_{\text{r,sm}} \int f(u_{\text{sm}}(x))dx + q_{\text{r}}\xi. \tag{3.20}$$

Visual stimuli

For each visual stimulus j with screen position p_j and size l_j, the spatial pattern on the screen is reproduced as a step function

$$h_j(x) = \begin{cases} 1, & \text{if } |x - p_j| \leq \frac{1}{2} l_j \\ 0, & \text{otherwise} \end{cases} \tag{3.21}$$

This pattern is then transformed into a logarithmically scaled retinocentric pattern m_j (with current fixation point x_{fix}) as

$$m_j(x) = h_j\left(\text{sign}(x)\zeta\left(\exp\left(\chi|x|\right) - 1\right) - x_{\text{fix}}\right), \tag{3.22}$$

with scaling parameters $\zeta = 100\,\text{px}$ and $\chi = \frac{\ln\left(\frac{450\,\text{px}}{\zeta}+1\right)}{150}$. This spatial pattern is smoothed with a normalized Gaussian kernel $k_{\text{vs,in}}$ with width $\sigma_{\text{vs,in}} = 2.5$.

It is then expanded to a two-dimensional pattern by multiplying it with a Gaussian pattern over the space of color hue values, centered on the stimulus color c_j and with width $\sigma_c = 4$. The temporal pattern for each stimulus is phasic-tonic, with the phasic component dependent on the stimulus start time $t_{j,\text{start}}$. The complete visual input for the visual sensory field is the sum of all stimulus patterns:

$$\begin{aligned} i_{\text{vs}}(x, y, t) = \sum_j \left(5 \cdot \exp\left(-\frac{t - t_{j,\text{start}}}{100\,\text{ms}}\right) + 10\right) \\ [k_{\text{vs,in}} * m_j](x) \cdot \exp\left(-\frac{(y - c_j)^2}{2\sigma_c^2}\right) \end{aligned} \tag{3.23}$$

The spatial visual input to the spatial attention field is purely phasic. It is based on the same pattern m_j used above, now smoothed with difference-of-Gaussians kernel $k_{\text{sa,in}}$ with a global inhibitory component that reduces input strength when multiple stimuli are present:

$$i_{\text{sa}}(x, t) = \sum_j 7.5 \cdot \exp\left(-\frac{t - t_{j,\text{start}}}{100\,\text{ms}}\right)[k_{\text{sa,in}} * m_j](x) \tag{3.24}$$

While a saccade is in progress, all visual input is set to zero.

Preshape

The preshape for the saccade task pre-activates the spatial attention field in those regions where the target stimulus may appear, and suppresses it at the possible remote distractor locations. To compute the excitatory preshape pattern, the average over the target stimulus patterns $m_{t_1}, \ldots, m_{t_n}$ for all possible eccentricities of the target stimulus (in steps of one pixel) is computed, and smoothed with the kernel $k_{\text{sa,in}}$ specified above. For blocks of trials with a remote distractor stimulus, the stimulus pattern m_d for the distractor is subtracted, otherwise this is omitted. The patterns for the two possible directions of target and distractor from the fixation point (left or right) are added up:

$$p_{\text{sacc}}(x) = \sum_{\text{dir}=\{l,r\}} \left(\frac{2.6}{n} \sum_{i=1}^{n} [k_{\text{sa,in}} * m_{t_i}^{\text{dir}}](x) - 1.2[k_{\text{sa,in}} * m_d^{\text{dir}}](x)\right) \tag{3.25}$$

The preshape for the memory test simply pre-activates the locations of the two memory test stimuli (left and right), based on their stimulus patterns m_{mt}^l and m_{mt}^r:

$$p_{\text{mt}}(x) = 1.25\left([k_{\text{sa,in}} * m_{\text{mt}}^l](x) + [k_{\text{sa,in}} * m_{\text{mt}}^r](x)\right) \tag{3.26}$$

Saccade metrics

A simulated saccade is assumed to start at the time t_{start} at which the output of the saccade reset node first exceeds a threshold $\theta_{\text{start}} = 0.25$, and ends at the time t_{end} at which the node's output falls below $\theta_{\text{end}} = 0.05$. The saccade amplitude s (in pixels on the screen) is determined by integrating the output of the saccade motor field over the whole time that a supra-threshold activation peak is present in that field (the integration thus begins before the saccade start time t_{start}). The output signal from each field location is scaled in this integration to reflect the stimulus eccentricity it represents, using the same mapping from field positions (retinocentric with logarithmic scaling) to screen positions as used in computing the visual input:

$$s = 0.0025\,\text{px} \iint f(u_{\text{sm}}(x,t))\,(\text{sign}(x)\zeta\,(\exp\,(\chi|x|) - 1))\,dxdt \tag{3.27}$$

Numerical simulations

Numeric simulations of the dynamical system are performed using the Euler method with a step size of 2 ms. The surface feature dimension is sampled with 174 units, with separate regions for color hue values (144 units) and gray values (30 units). The feature space in each of these regions is defined in a circular manner, without any local interactions between the regions. The spatial dimension is sampled with 301 units, covering a range from approximately -15° to 15° in retinocentric space (with logarithmic mapping of stimulus positions onto this spatial dimension, as described above). All parameter values for the interaction widths are given in these field units.

3.4 Test of the DNF model

3.4.1 Experimental procedure and task conditions

In this section I will describe the procedures of the psychophysical experiment that was used to investigate the effects of VWM content on saccade target selection. The experimental design was a slight variation of two previous studies (Hollingworth et al., 2013a,b), devised in collaboration with Andrew Hollingworth and executed by him at the University of Iowa specifically to test the DNF model. The psychophysical experiment combines a color working memory task and timed saccade task in an interleaved fashion.

Participants faced a monitor on which stimuli were presented, and their eye movements were recorded via an eye tracking system. Participants then performed trials that can be divided into three phases (Figure 3.8). In the first phase, a colored square stimulus was presented in the center of the screen, visible for 300 ms, and participants were instructed to memorize its

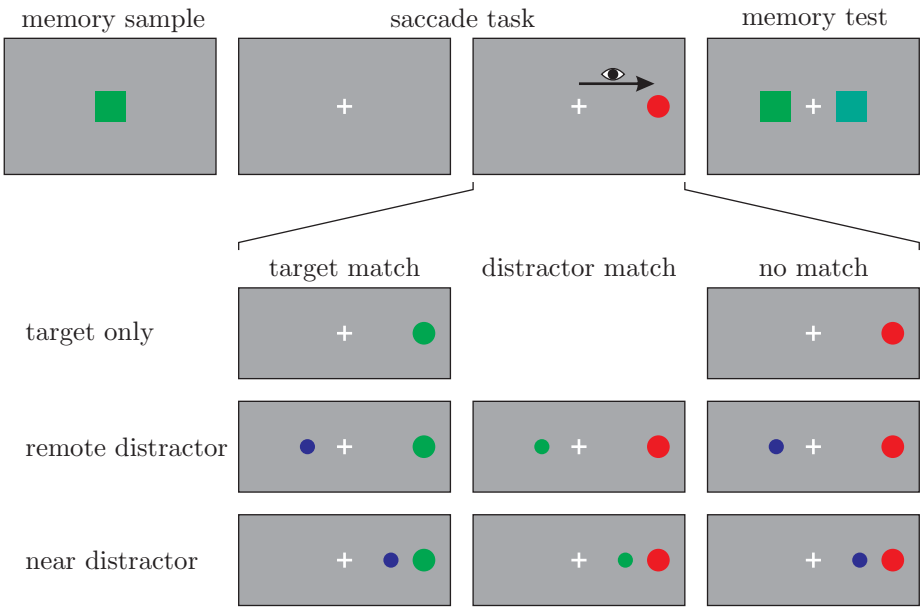


Figure 3.8: Psychophysical task to test interactions between VWM and saccade behavior. The top row shows the sequence of displays presented to participants in the course of one trial. The arrow with the eye symbol indicates the instructed eye movement to be performed in this phase of the trial (this is not part of the stimulus display). In the bottom, the different paradigms and color match conditions for the saccade task are depicted.

color. In the second phase, a saccade task was performed: The colored stimulus was replaced by a fixation cross, and after a fixed delay period (700 ms) a saccade target stimulus appeared, in part of the trials accompanied by a distractor stimulus. The saccade target could be identified unambiguously by its location (it was always the outer object, as detailed below), and participants were instructed to fixate it as quickly as possible after it appeared. The target and distractor stimuli were removed after a fixed delay after the participant had made a saccade, and only the fixation stimulus remained. The third phase of the trial was a memory test: Two colored squares appeared on either side of the fixation cue, one of them matching the color of the memory cue from the first phase of the trial, the other of a similar color chosen from the same color category, referred to hereafter as the foil color. Participants had to indicate by a button press which of these stimuli (left or right) matched the color they had memorized in the first phase of that trial. Participants received no feedback of their performance during the trials.

In the saccade task, three different stimulus arrangements were employed (Figure 3.8): In the *target only* paradigm, only the saccade target stimulus

was presented. It was a colored disk with a diameter of 1° of visual angle and was located either to the left or the right of the fixation point, with distance varied between 4.6° and 7°. In the *remote distractor* paradigm, another, smaller colored disk (0.66° diameter) was presented simultaneously with the saccade target. It was always located on the side opposite to the target stimulus, with a distance of 1.3° from the fixation cue. Participants were instructed to ignore this distractor. Note that the distance for the distractor does not overlap with distance range of the target stimulus, such that the two can be clearly distinguished based on their position (and also based on their size). The third paradigm used was the *near distractor* paradigm. Here, the distractor stimulus was located on the same side as the target, 2.3° closer to the fixation point. Target only trials and remote distractor trials were randomly interspersed within blocks of trials (25% remote distractor trials), near distractor trials were performed in separate blocks.

The key manipulation in the experiment pertained to the stimulus color in the saccade task. Three conditions were distinguished: In the *target match* condition, the saccade target stimulus was the same color as the memory stimulus in that trial (and the distractor was a different color); in the *distractor match* condition, the color of the distractor (but not the target stimulus) matched the memory cue; and in the *no match* condition, neither the saccade target nor the distractor were the same color as the memory cue. The memory cue color and match condition for each trial was varied in a pseudo-random manner, such that the stimulus colors in the saccade task were not informative regarding the role of each stimulus for the task (saccade target or distractor). Nonetheless, the color match condition in the saccade task had significant effects on participants' saccade behavior, as detailed below.

As an additional manipulation, the color match for both target and distractor could be either exact or inexact. For an inexact color match, the target or distractor was the same as the foil color in the following memory test (belonging to the same color category as the memory color, but slightly different from it). These inexact match trials were used to investigate metric properties of the interaction effects, and in particular revealed effects of the stimuli in the saccade task onto the working memory representation. Additional details about the experimental method and the data analysis can be found in Hollingworth et al. (2013a,b).

3.4.2 Empirical and simulation results for the saccade task

The results of the study revealed significant effects of color match condition on saccade target selection, saccade amplitudes, and saccade latencies. I will present them here together with the simulation results, and describe how the observed effects arise in the DNF model in the following section. Parameter values in the DNF model (such as interaction strengths within and between

fields) were manually adjusted to obtain a fit of the experimental data. The same parameters were used in the simulations of all experimental conditions. The model then generated novel predictions that were tested in a separate experiment, described below.

Mean values for saccade metrics and latencies in the experiments were obtained by averaging over the results of the trials of all participants. Trials with leftward and rightward target directions were pooled and results are presented as in rightward trials. Saccade amplitudes were measured as the size of the horizontal displacement of the fixation point during the first saccade (in visual angle), saccade latencies as time between target stimulus onset and first saccade initiation. For the target only and remote distractor paradigms, twelve participants each performed a total of 384 trials (288 in target only paradigm, 96 for remote distractor). For the near distractor condition, eight participants completed 400 trials each. For the simulation results, a total of 7296 trials were run to approximate the total number of trials in the empirical study (3040 target only, 1824 remote distractor, and 2432 near distractor trials). Random noise was added to the field activations to produce a stochastic distribution of results in the simulations. Trials with saccade latency below 60 ms or above 500 ms were excluded from the analysis for both experimental and simulation results. The results for exact and inexact match conditions were pooled for the saccade data since they are qualitatively equal, as are the results for leftward and rightward saccades.

The empirical results for the target only paradigm are shown in Figure 3.9a-b. Saccades to the target stimulus generally fell slightly short of the target. This undershoot was significantly reduced in the target match condition. Mean saccade landing point relative to target location was -0.31° for target match vs. -0.40° for no match ($t(11) = 4.38$, $p = 0.001$). Furthermore, saccades were initiated significantly faster if the target matched the memory color, with mean saccade latencies of 140 ms for the match condition compared to 146 ms in the no match condition ($t(11) = 3.20$, $p = 0.008$).

Both of these effects were reproduced in the model simulations (Figure 3.9c-d). Mean relative landing position was -0.42° in the target match condition compared to -0.47° in the no match condition ($p = 0.01$). Mean saccade latencies were likewise reduced in the target match condition (149 ms) vs. the no match condition (160 ms, $p < 0.001$). The model moreover provided a good qualitative fit of the overall pattern of the saccade landing point distribution.

For the remote distractor paradigm, the key measure for color WM effects is the proportion of first saccades that were directed toward the target stimulus (landed within 1.5° from the target center) rather than toward the distractor. The distribution of saccade landing positions and saccade latencies for the empirical study are shown in Figure 3.10a-b. As can be seen, the color match condition has a significant influence on the effectiveness of

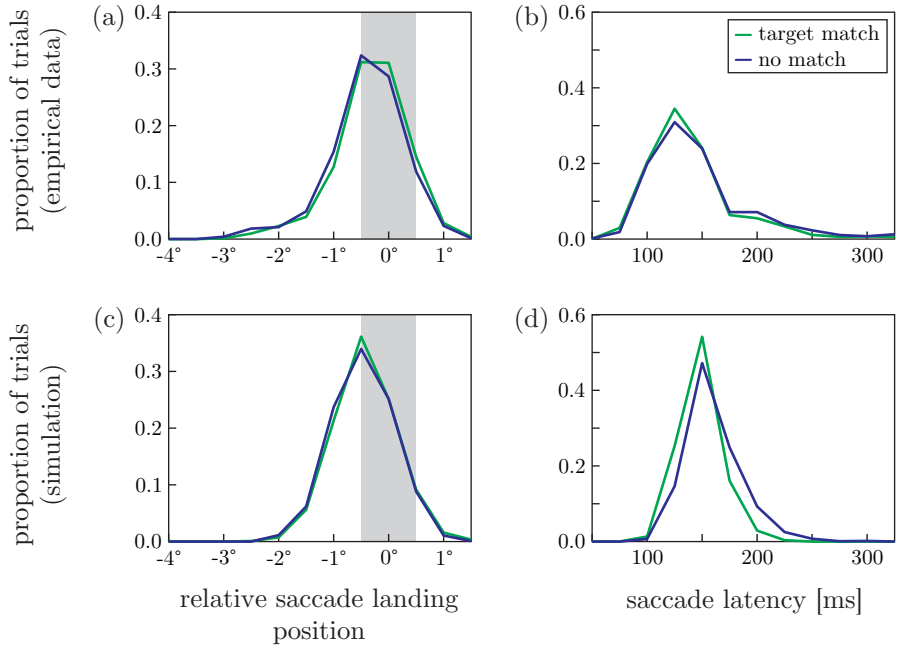


Figure 3.9: Empirical (top) and simulation results (bottom) in the target only paradigm of the saccade task. (a, c) Histogram of saccade landing positions relative to the target location. Negative position values indicate saccades that fell short of the target location, positive values indicated overshoot. The gray bar shows the extent of the target stimulus. (b, d) Histogram of saccade latencies.

the distractor to capture saccades. In the target match condition, 93.6% of saccades landed near the target. This proportion was significantly reduced in the no match condition (79.6%, $t(11) = 3.62$, $p = 0.004$), and further significantly reduced in the distractor match condition (40.1%, $t(11) = 12.7$, $p < 0.001$). This means that the stimulus that matched the color held in VWM was generally more likely to be selected as saccade target. Mean saccade latency (averaged over saccades to the target only) was lowest in the target match condition (172 ms, $t(11) = 5.85$, $p < 0.001$), but did not differ significantly between no match (199 ms) and distractor match (202 ms) conditions.

The simulation results shown in Figure 3.10c-d again reproduce this pattern of results. In the simulation, the proportion of saccades directed at the target was 89.3% in the target match condition, significantly reduced to 69.4% in the no match condition ($\chi^2 = 80.3$, $p < 0.001$) and further reduced to 39.1% in the distractor match condition ($\chi^2 = 125.2$, $p < 0.001$). Mean latency of simulated saccades to the target was 168 ms in the target match

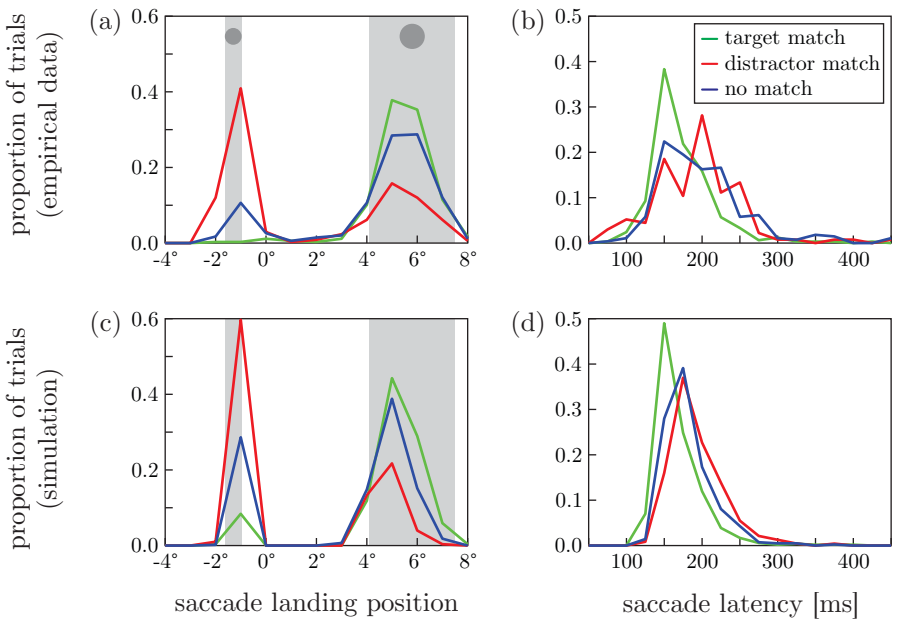


Figure 3.10: Empirical and simulation results for the remote distractor paradigm. (a, c) Histogram of saccade landing position, with initial fixation point at 0° and saccade target always on the right. Grey circles show the sizes of target and distractor stimuli, gray bars the range of stimulus positions. (b, d) Histogram of saccade latencies.

condition, significantly lower that in both the no match condition with 192 ms ($p < 0.001$) and the distractor match condition (181 ms). In the simulation, the smaller latency difference between these two latter conditions also reached significance, unlike in the experimental results ($p < 0.001$).

In the near distractor paradigm, the landing point of the first saccade after stimulus onset was typically located between the target and the distractor stimulus in all conditions (see Figure 3.11a). This constitutes an instance of averaging saccades, a well-known phenomenon in saccade target selection that occurs when multiple stimuli are located close to each other (Van der Stigchel and Nijboer, 2011). The mean location of the saccade landing point varied with match condition: In the target match condition, the mean saccade landing position relative to the target was -1.07°. In the distractor match condition, it was $-1.48°$ on average, meaning the saccade landed closer to the distractor stimulus. Mean relative landing position in the no match condition was $-1.27°$, lying between the two other conditions (significantly different from target match, $t(7) = 4.98$, $p = 0.002$, and from distractor match, $t(7) = 4.28$, $p = 0.004$). Again, mean saccade latency was

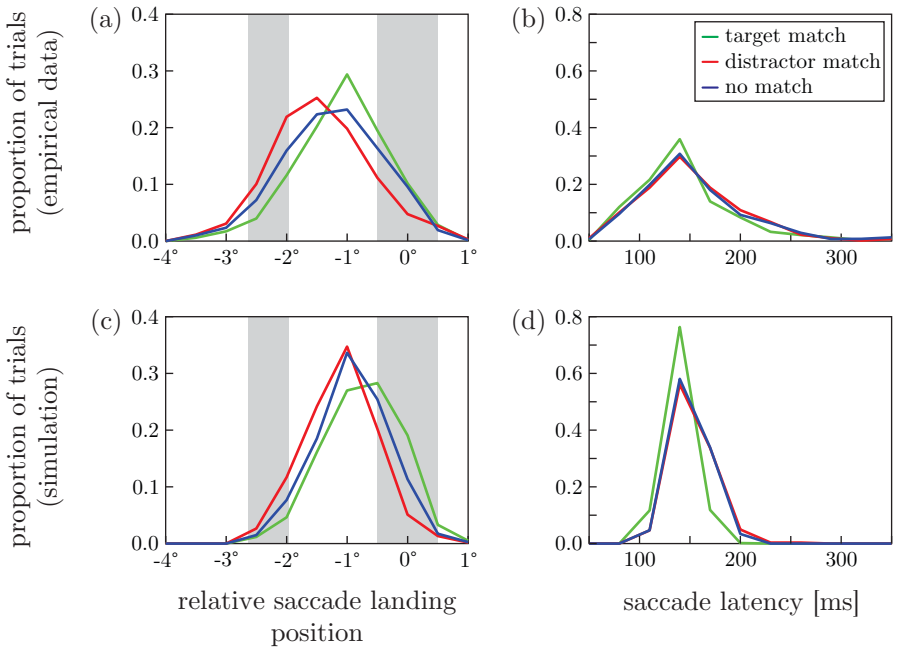


Figure 3.11: Empirical and simulation results for the near distractor paradigm. (a, c) Histogram of saccade landing position relative to the location of the target stimulus. Gray bars show the locations and sizes of target and distractor stimuli. (b, d) Histogram of saccade latencies.

shortest in the target match condition (155 ms, significantly shorter than no match, $t(7) = 2.73$, $p = 0.03$). Mean latencies in the distractor match condition (160 ms) and no match condition (161 ms) did not show significant differences (Figure 3.11b).

The model succeeded in reproducing these results as well (Figure 3.11c-d). It produced averaging saccades in response to the two nearby stimuli, and the saccade landing point dependent on the color match. In the target match condition, mean saccade landing point relative to the target position was $-0.75°$. It shifted to $-0.92°$ in the no match condition (significantly different, $p < 0.001$), and further away from the target to $-1.10°$ in the distractor match condition (significantly different from no match, $p < 0.001$). As in the experimental results, mean saccade latency was shortest in the target match condition (149 ms, significantly different from both other conditions, $p < 0.001$), and not significantly different between distractor match (159 ms) and no match conditions (161 ms).

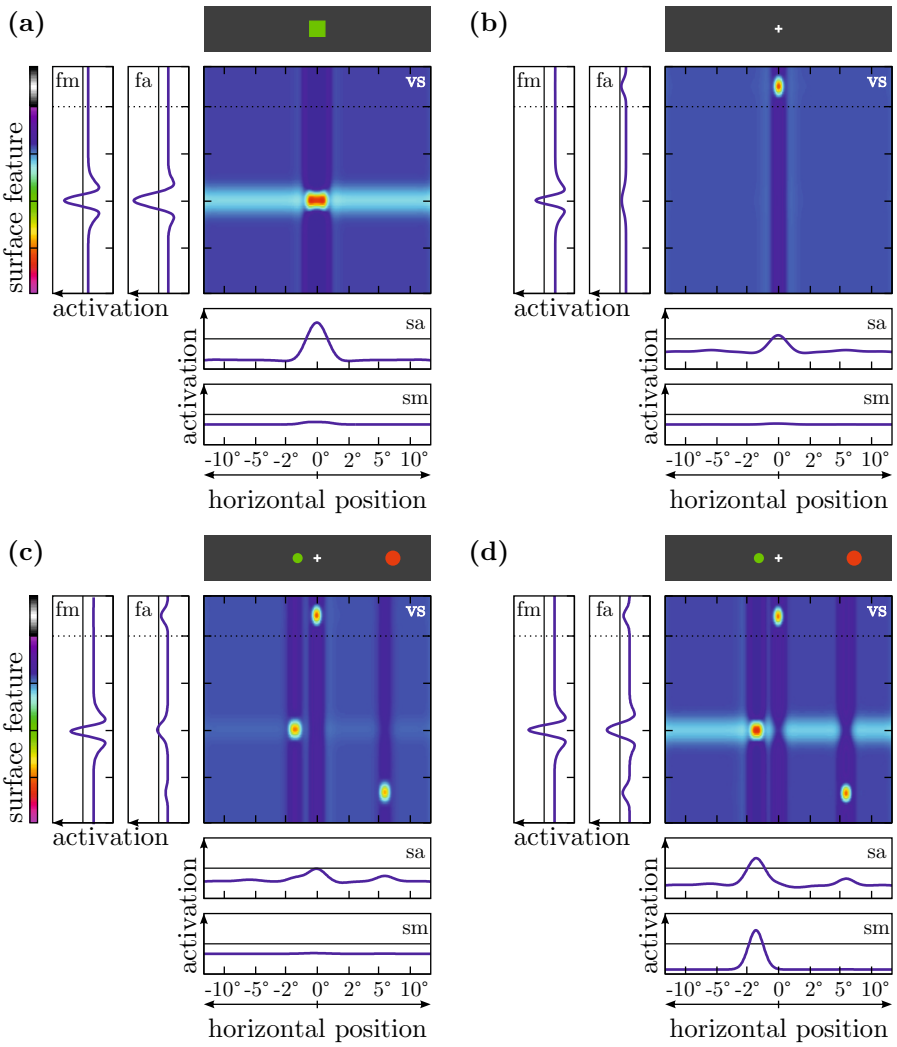


Figure 3.12: Evolution of activation patterns in the DNF model during one trial of the saccade task. (a) Presentation of the memory sample stimulus. (b) Delay period. Visible are the pre-activation for the memorized color in the feature attention field, and the weak preshaping of the activation in the spatial attention field. (c) Situation briefly after the onset of target and distractor stimulus. (d) Initiation of a saccade to the distractor stimulus.

3.4.3 Time course of activation during a simulated trial

To explain how the biased competition model integrates VWM, perceptual processing, and saccade behavior to produce the results described above, I will go through the time course of one trial in the model and describe the

evolution of activation patterns (Figure 3.12). I will do this in detail for a trial of the remote distractor paradigm, in which most of the key effects that occur in the model are covered, and then more briefly discuss effects that are specific for the target only and near distractor paradigms.

At the beginning of each trial, all fields in the model are at their respective resting levels, and the system is fixating the center of the screen. The memory stimulus is now presented at this location, and produces a strong activation peak in the visual sensory field (Figure 3.12a). The peak is located centrally along the spatial dimension, and its location along the surface feature dimension reflects the stimulus color (here, green). Along the spatial pathway, the single salient stimulus induces a strong activation peak in the spatial attention field. Since it is located at the central (foveal) position in the field, activation is not projected further to the saccade motor field (the forward connection is suppressed in this region).

In the surface feature pathway, the input from the visual sensory field first induces an activation peak for the stimulus color in the feature attention field. The output of this field then drives activation locally in the feature working memory field. During the memorization period, the feature working memory field receives and additional control input that globally increases the activation levels. Through the combination of these inputs, an activation peak forms for the color of the presented stimulus.

The memory stimulus is now removed and replaced by a small white fixation stimulus. The activation patterns in the architecture during this fixation period are shown in Figure 3.12b. The activation peak in the feature working memory field remains self-sustained after both the stimulus that induced it and the global control input are turned off. It provides feedback input to the feature attention field, which produces a localized hill of activation in that field. The activation level remains below the field's output threshold, and therefore no feedback input for this feature value is projected further back to the visual sensory field. At the same time, the smaller fixation cue induces an activation peak in the visual sensory field, which in turn creates peaks in the feature and spatial attention fields.

After the delay period with only the fixation point visible, the saccade target and distractor stimuli are activated. In this example, I will describe a distractor match trial, as shown in Figure 3.12c. A green distractor stimulus is presented to the left of the fixation point, relatively close to it, while a larger red saccade target stimulus appears further from the fixation point on the right side. Both of these new stimuli induce activation peaks in the visual sensory field, in addition to the still present fixation stimulus.

In the spatial attention field, the inputs from the three stimuli compete with each other by means of the lateral interactions in the field. This competition is modulated by additional inputs to the field that reflect the task instructions for the saccade task: The gaze change node in the model is ac-

tivated, and suppresses the foveal region of the spatial attention field, thus weakening the activation peak induced by the fixation stimulus. In addition, the instruction to make a saccade to the outer stimulus and ignore the distractor is reflected by a constant preshaping input: The regions on both sides of the field where the target stimulus can appear are pre-activated, the possible distractor locations are suppressed. With these modulatory inputs, the saccade target would typically prevail in the competition based only on its spatial characteristics.

The competition for spatial attention is however biased by the interactions in the surface feature pathway. The peaks that the two colored stimuli induce in the visual sensory field project to the feature attention field and create hills of activation there. In the case of a color match, one of these inputs will coincide with the region that is already pre-activated by the peak in the feature WM field—in the example, the green distractor item matches the memory color (Figure 3.12c). Consequently, a supra-threshold activation peak for this color value forms much more quickly, and due to the extra input from the feature WM field becomes larger than a purely stimulus-driven peak would be. The feature working memory field projects back to the visual sensory field, and consequently the peak for the matching stimulus in that field is strengthened compared to other peaks.

The biasing effect of the feature match on the representation in the visual sensory field is also transmitted further to the spatial attention field, via the spatial read-out projection. It influences the competition process that is in progress in that field and creates a bias to select the location of the item that matches the memorized color—here the distractor stimulus. In this scenario, the resulting total inputs to the spatial attention field are of approximately equal strength for the distractor and target stimulus. Nonetheless, the competitive interactions will enforce a selection decision in which one peak prevails while the other one is suppressed. Here, the distractor location is selected, and the resulting single strong activation peak is enough to drive the activation in the saccade motor field above the threshold (Figure 3.12d).

This triggers the saccade behavior as described above: A saccade motor peak forms, drives an eye movement, activates the saccade reset node, and is extinguished again. While the saccade is in progress, the external visual input is suppressed. When it comes up again, it is shifted and centered on the new fixation point. If one of the stimuli is fixated after the saccade, the corresponding input to the central region of the spatial attention field acts to stabilize this fixation and suppresses further saccades. However, the saccade may fail to bring a stimulus into the foveal region for several reasons: Random noise in the fields leads to variations in saccade amplitude, and saccades to distant targets have a tendency to undershoot in the model (consistent with human saccade behavior). Furthermore, if two stimuli are

close to each other (as in the near distractor paradigm, see below), the system will generally perform an averaging saccade that lands between the stimuli. In all of these cases, the system is in a state with no stimulus in the foveal region after the saccade, but with one (or more) close to the fovea. By the very same mechanism that drove the initial saccade, the system will then autonomously perform a correction saccade to fixate that stimulus. This is again consistent with human behavior in such situations.

3.4.4 Different interaction effects in the model

The biasing effect of the color match on the spatial selection process directly explains the saccade target selection results in the remote distractor paradigm. Without any color bias (the no match condition), the saccade target has a clear advantage due to the different spatial effects (stimulus size and preshaping of the field activation), and prevails in the competition in the majority of trials. The distractor stimulus is still selected in part of the trials due to effects of random noise on the competition. If the distractor color matches the memory color, the color bias largely cancels out the advantages of the target stimulus, leading to approximately equal proportions of saccades to distractor and target. In the target match conditions, all biasing effects favor the designated saccade target, and only very few saccades to the distractor are made.

The differences in saccade latency in the remote distractor paradigm result primarily from the selection process in the spatial attention field. The competition between two locations is resolved very quickly if one receives significantly more input than the other, but can take a long time to resolve if activation levels are very similar. In the latter case, the small differences in activation value induce only a low rate of change for the activation at both locations, and the balance between two locations may change back and forth due to random noise. This is reflected in the longer mean saccade latencies and high proportion of very slow saccades in the distractor match condition (Figure 3.10b and d). In the target match condition, the competition is the most lopsided, and consequently mean saccade latency is lowest.

In the near distractor paradigm (target and distractor on the same side), the model typically generates an averaging saccade that lands between the two stimuli, consistent with the experimental results. This averaging is produced by the relatively broad lateral and feed-forward interaction kernels in the spatial pathway: Through broad feed-forward projections from the visual sensory field to the spatial attention field, the inputs created by proximate stimuli are partly joined. Broad lateral excitatory interactions then further act to merge these adjacent inputs into a single activation peak, centered between the original input locations. This averaged peak in the spatial attention field then projects to the saccade motor field and produces a corresponding saccade signal.

The exact location of the activation peak between the two input locations is influenced by the strengths of the inputs. In the target match condition, the input for the saccade target stimulus is strengthened by the color bias, and the averaged peak is consequently centered closer to the target location as compared to the no match condition. For the same reason, in the distractor match condition, the peak forms on average closer to the distractor location. This reproduces and explains the saccade amplitude effects for this experimental paradigm. The lower mean saccade latency for the target match condition can be explained in a similar way as for the remote distractor paradigm: In the target match conditions, all positive biasing effects converge on the target location, so a peak can form here particularly quickly.

In the target only paradigm, there is only the single saccade target. The corresponding activation peak in the visual sensory field is strengthened if it matches the memory color, and is not modulated otherwise. This directly explains the shorter saccade latencies in the target match condition: With stronger input, the peaks in the spatial attention field and subsequently in the saccade motor field can form more quickly. The small (but significant) differences in mean saccade amplitude are a result of the incomplete normalization in the saccade generation: As stated above, the strong interactions in the saccade motor field make the time course of peak formation and decay largely independent of the input strength, but this normalization is not complete. Since the input signal is strengthened in the target match condition, it takes slightly longer for the saccade reset node to suppress the saccade motor peak, and the resulting saccades are slightly longer than they would be without the color match.

The fact that these interaction effects arise even for an inexact color match (albeit slightly weaker) can be explained in the model by the extent of the activation peaks along the feature dimension and the spread of activation in the projections between fields. The activation peaks in the surface feature dimension (as in the spatial dimension) are not localized to a single feature value, but extend over a range of values. In the projection to another field—for instance, in the feedback projection from a working memory peak to the feature attention field—activation is spread out over an even larger range. Consequently, the working memory peak for a certain color can provide additional excitation not only for stimuli of the exact same color, but also for similar colors (that are metrically close in the surface feature dimension). The additional excitation is weaker than for an exact color match, and decreases with further distance between the color values, but its effect on the saccade system is qualitatively the same as for an exact match.

3.4.5 Effects of color matches on memory performance

The experimental and simulation results discussed so far have been explained by uni-directional biasing effects from the VWM for surface features on the

spatial selection for saccadic eye movement. If the idea of a dynamically and bidirectionally coupled system as introduced in the beginning of this chapter is an accurate description of the visual system, we should expect that interaction effects can also be observed in the opposite direction. In fact, the results of the memory test in the experiment provide evidence for such interactions, in that they show an effect of perceptual processing on the working memory representation.

In the color memory test, participants had to indicate via a manual response which of two stimuli matched the color of the memory cue from the beginning of the trial. Analysis of the experimental results yielded a significant effect of the color match type—exact, inexact, or no color match—on memory test performance. In the no match condition, pooled over the three experimental paradigms, participants reached a performance of 77.2% correct choices. For the exact match trials, pooled over paradigms and target/distractor match, the performance increased to 82.5% correct. In contrast, for the inexact match trials pooled in the same way, performance was decreased to 74.1% correct. Recall that in these trials, the color of one item in the saccade task matches the foil color in the memory test.

The memory test is emulated in the model as a forced-choice saccade task, in which an eye movement from the central fixation point to either of the two equidistant stimuli is taken as the response for the memory test. This is not intended to be a direct model of the experimental procedure (in which participants may freely fixate either or both of the stimuli during their decision making), but as an alternative probe of the working memory content. The underlying assumption is that the differences in performance arise from a change in the working memory representation itself, and not the procedure of the memory test.

The model mechanism for the memory test works as follows (Figure 3.13): After the saccade task, only the fixation stimulus remains, and the model's gaze direction is externally reset to be centered on that fixation stimulus. The artificial gaze reset is used instead of a saccade generated by the model itself to ensure that there are no variations in the initial gaze direction, which could bias the subsequent memory test saccade. Then, in preparation of the memory test, the activation level of the feature attention field is globally increased. This allows the feedback input from the feature working memory field to form a supra-threshold activation peak (Figure 3.13a), and thereby increases the effect that the working memory representation has on the visual sensory field. Effectively, the system is brought into an explicit visual search mode, with the search target defined by the working memory peak. When the two memory test stimuli are presented and corresponding peaks form in the visual sensory field, the peak that matches the feature value in VWM is immediately supported by the already present input ridge from the feature attention field.

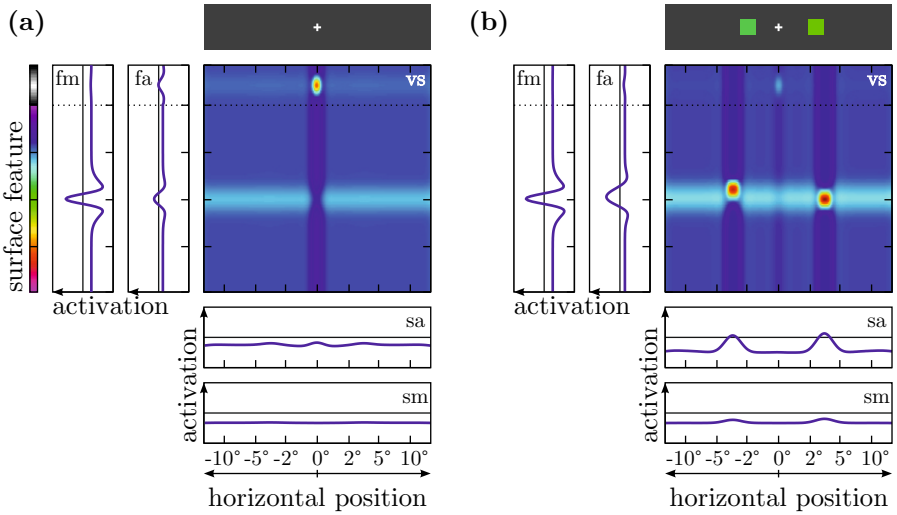


Figure 3.13: Evolution of activation patterns during the memory test period of one trial. (a) Preparatory period directly before the memory test, with activation level in the feature attention field globally increased. (b) Situation briefly after onset of memory test stimuli.

The saccade target selection then works in the same way as during the saccade task. Since the two stimuli are equivalent in their spatial features, the outcome is determined primarily by the bias from the feature match. However, since the two stimuli are similar in the color hue value, even the non-matching peak in the visual sensory field is somewhat strengthened by the feedback from the feature attention field, and the biasing effect is not as strong as it would be otherwise. The resulting performance of the model in the no-match condition is similar to participants' performance in this task: The correct memory color is selected as saccade target in 87.4% of all trials. Moreover, the model also reproduces the effects of exact and inexact color match observed in the experiment: An exact color match in the saccade trial raises the performance to 90.0%, an inexact match decreases it to 78.2%.

In the model, the effect of the color match type in the saccade task on the memory test performance is a result of the continuous coupling between the feature working memory field and the visual processing in other fields. In the same way as the working memory representation influences the saccade behavior even if the task does not demand this, the earlier visual fields affect the activation patterns in the working memory field even if there is no instruction to memorize something. The detailed effects are as follows: In all conditions, the working memory peak shows a certain amount of random drift along the feature dimension, due to noise in the field activation. This

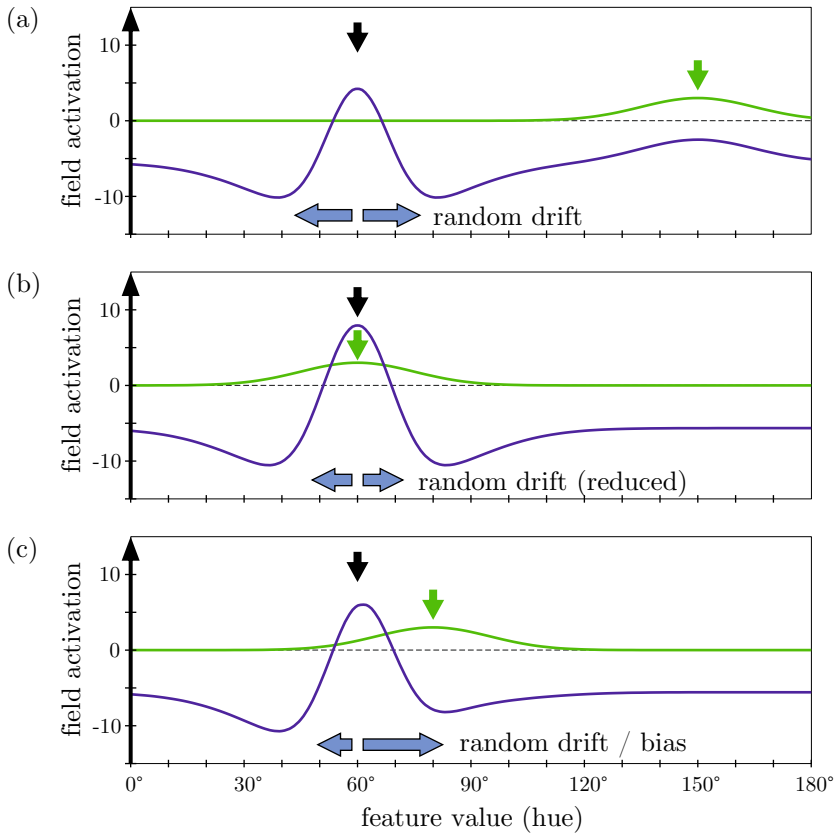


Figure 3.14: Effects of saccade task stimuli on activation peak in feature WM field. The blue plot shows the activation distribution in a part of the feature WM field, the green line shows the input induced by a visual stimulus (via the feature attention field). The originally memorized hue value is indicated by a black arrow, the green arrow indicates the hue value of the current stimulus. (a) Situation for dissimilar hue values. (b) Situation for exactly matching hue values. (c) Situation for an inexact match in hue values.

drift makes the memory representation imprecise and reduces performance in the memory test. It decreases the biasing input for the stimulus that matches the originally memorized color, and can furthermore increase the input for the incorrect stimulus if the drift brings the peak closer to the foil color value.

During the saccade task, additional visual stimuli appear that provide input to the feature WM field. When all stimulus colors are dissimilar to the memorized color (no match condition), these inputs do not interact with the working memory peak, and random drift occurs in the same way as if

no additional inputs were present (Figure 3.14a). However, when one of the stimuli has the exact same color as is held in memory, this input stabilizes the memory peak at its original position and thereby reduces random drift (Figure 3.14b). This explains the improvement in performance for the exact color match trials compared to the no match trials. For an inexact color match, the input from the saccade task stimulus is slightly offset from the position of the working memory peak. It creates an activation gradient that pulls the working memory peak toward the feature value of the input (Figure 3.14c). As a result, the peak in the feature WM field is located at an intermediate position between the originally memorized color and the foil color when the memory test is performed. It is then less effective to bias the decision toward the correct color, which explains the decreased memory performance for the inexact color match.

The effect in the model was quantified by determining the position of the activation peak in the feature WM field at the beginning of the memory test, and computing its deviation from the memorized color. In inexact match trials, the mean deviation in hue space was 3.4° toward the foil color (significantly different from zero, $p < 0.001$), confirming the biasing effect qualitatively described above. In the exact match and no match conditions, no significant deviation in mean peak position was found (0.19° and 0.10°, respectively). The analysis also confirmed the stabilizing effect of an exactly matching stimulus on the peak position: The standard deviation in peak position was lower in exact match trials (3.9°) compared to no match trials (4.8°) and inexact match trials (5.1°).

This explanation in the DNF model provides an experimentally testable prediction for the working memory representation in human participants: The encoded feature value after a saccade task with an inexact color match should show a metric bias toward the foil color. This prediction was tested experimentally in a variant of the experiment, using only the target-only paradigm: Instead of a two-alternative forced choice task, the participants were asked to indicate the color they had memorized by setting a slider on a color wheel (Figure 3.15; see Schneegans et al., 2014, for details of the experimental procedure).

The results confirmed the expected bias toward the foil color in the inexact color match condition. The mean deviation in participants' responses for this condition was 1.9° toward the foil color (significantly different from zero, $t(15) = 4.1$, $p < 0.001$). This is comparable to the deviation predicted by the model. In the other match conditions, no systematic bias in color memory was found. The predicted stabilizing effect of an exact color match was also confirmed by the experiment. In this condition, the mean standard deviation in participants' color response (12.4°) was significantly smaller than in the inexact match condition (14.2°) and the no match condition (15.9°). It is noticeable that the standard deviations are overall higher in the experiment

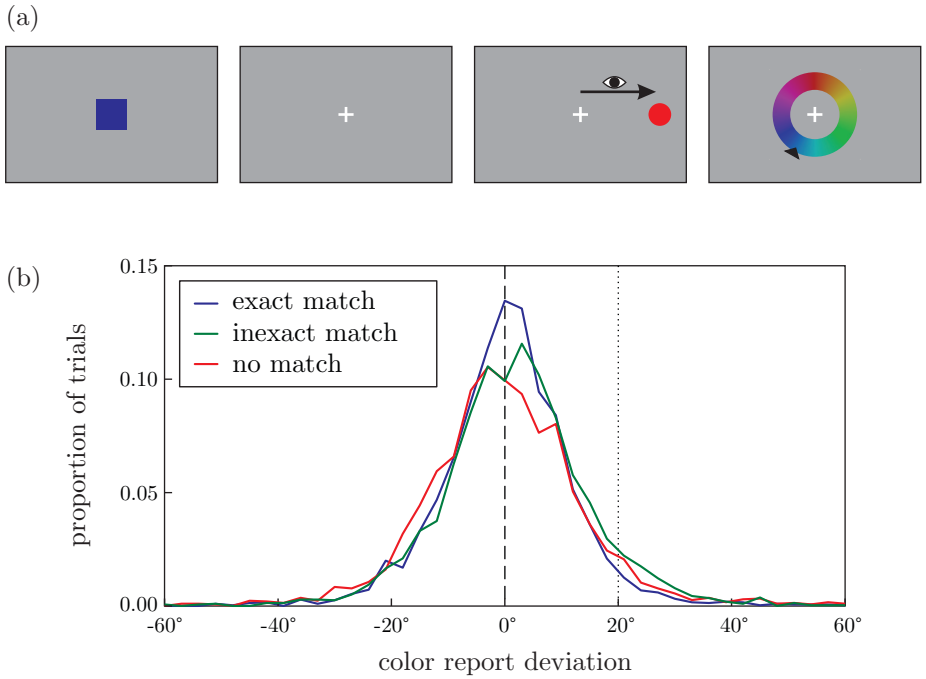


Figure 3.15: Experiment to test model prediction of WM biases. (a) Experimental procedure. The two-alternative forced choice test at the end of the trial is replaced with a continuous response test, in which participants have to indicate the memorized color on a color wheel. (b) Distribution of response errors in the memory test.

than those predicted by the model. This may be explained, however, by the fact that the memorized color in the model was read out directly from the peak location. The generation of a manual response to indicate the memorized color in the experiment is likely to introduce additional variability.

3.5 Discussion

In the present chapter, I have introduced multi-dimensional DNFs, and have shown how they can be used to mediate associations between different feature dimensions. The proposed mechanism assumes that localized activation is present in the multi-dimensional DNF, reflecting locations and colors of visual stimuli in the examples of this chapter. Through bi-directional connections, the system then achieves a coupled selection of one item in separate one-dimensional fields over different feature dimensions.

This association mechanism can be used to deal with the problem of fea-

ture binding in distributed representations, by selecting one visual item at a time through focused attention. The approach is consistent with the Feature Integration Theory (Treisman and Gelade, 1980), a prominent psychological theory of visual perception. Ample evidence from different experimental paradigms supports the core claim of this theory, namely that focused attention is necessary to form and use conjunctions of different visual features. The reasons why the separation into multiple distinct representation in visual processing is employed in the first place, instead of using the type of joined representation as in the two-dimensional field throughout, will be addressed in more detail in the following chapters.

The psychophysical experiment on interactions between VWM and saccade behavior that I have presented here provides three key constraints for the modeling of attentional selection in the visual system. First, interactions between feature and spatial representations occur at an early level of visual processing. This can be derived from the finding that even fast reactive saccades to sudden onset targets, which have generally been thought to be purely stimulus driven, are affected by VWM content. This observation speaks against separate processing steps in visual perception, with top-down effects only in the later stages, as proposed by some models of visual search (Wolfe, 1994; Bundesen et al., 2005).

Second, interaction effects between VWM and saccade planning are not strictly strategic. They take place even when they are not relevant for the current task, or are even interfering with it. This indicates that they are not under cognitive control, but instead reflect an inherent property of the neural architecture. And third, the interaction effects are bidirectional, as is evident here in the VWM biases induced by visual processing during the saccade task. This highlights that VWM is not stored away in a passive fashion, disconnected from visual processing, but that VWM maintenance is an active process that not only affects the visual processing, but is in turn also affected by it.

Taken together, these findings provide strong support for the kind of mechanism that is implemented in the DNF model to achieve an attentional selection of a stimulus. This mechanism is characterized by continuous coupling between different representations, mediated by bi-directional projections between them. The selection of an item in the focus of attention is not a discrete operation, but rather emerges from the competitive interactions that operate continuously in time, as previously proposed in models of Deco and Lee (2002) and Hamker (2005a). The active maintenance of VWM and the planning of saccadic eye movements are directly coupled to this selection mechanism, and employ the same type of representation that is also the basis for perceptual processing.

The modeling results demonstrate how measured behavioral variables can be used as signatures of the underlying neural processing. The DNF model

generates distributions of saccade landing positions and saccade latencies by actually executing a large number of simulated trials, directly emulating the experimental task. It reproduces the influence of VWM on saccade target target selection and saccade amplitudes. Saccade latencies in the model directly reflect the time it takes for the competitive interactions to produce a selection decision, and they successfully recreate the patterns found in the experiment. The model provides a concrete functional explanation for the effects of visual processing on memory performance, and the proposed bias effect was confirmed in a novel experiment study.

To produce these results, the approach presented here combines two previous lines of DNF models, one focusing on saccade planning, the other on VWM and change detection. Models of saccade planning have explained effects of distractors and presence of a sustained fixations point on saccade latency (Trappenberg et al., 2001; Wilimzig et al., 2006), based on the same competitive mechanisms as employed in the present model. They have also described the transition from selection to averaging for multiple nearby stimuli, which was used here to capture the metric effects in the near distractor paradigm (Wilimzig et al., 2006; Marino et al., 2012). The present work goes beyond these previous models in that it does not only capture the activation time course leading up to the selection of a saccade target, but also the full process of space-to-rate-code transformation for the generation of the motor signal, and the termination of the saccade.

Previous DNF models of VWM have addressed questions of working memory capacity and effects of feature similarity on change detection performance (Johnson et al., 2009a,b). They have also explained biasing effects found in spatial VWM due to perceptual inputs. These explanations are analogous to the mechanism proposed here to account for effects of color match on memory performance (Schutte and Spencer, 2009, 2010). This consistency between the different models supports the notion that close coupling to sensory processing is a general property of VWM. I will return to the task of modeling VWM in Chapter 5, where I will combine the attentional selection mechanism introduced here with previous models of change detection to provide an account of human scene representation.

The DNF model presented here shows strong parallels to previous models of visual search, in particular to the neurodynamic models of Hamker 2005b; 2006. Like the present approach, these models propose separate spatial and surface feature pathways that are coupled to a shared low-level sensory representation. These models differ somewhat in formulation of the differential equations governing the neural dynamics, but they employ the same population code representations and the same basic mechanism for attentional selection. These models have not previously been employed, however, to explain VWM effects on saccade planning outside of explicit visual search tasks, and they lack the detailed mechanism for saccade generation presented here

to account for the metric details in saccade behavior. These previous models also have not addressed memory biases induced by perceptual processing.

One aspect that is lacking in the DNF model compared to many visual search models is the treatment of different surface feature dimensions and their interactions. These are critical for explaining key effects in visual search tasks (such as the difference between parallel search for single features conjunctions and serial search for feature conjunctions), but were not required to capture all relevant stimulus parameters in the experiment treated here, so they were omitted in the model for simplicity. An extension of the mechanism that includes multiple surface feature dimensions will be treated in Chapter 5, using multiple feature maps coupled via their shared spatial dimension. The same kind of extension has also been used to address the effect of illusory conjunction, which represent failures in feature binding that can occur under certain conditions when spatial attention cannot be sufficiently focused onto an individual stimulus (publication in preparation).

With respect to the more general goal of this thesis, as outlined in the introduction, the present chapter demonstrates the principle of an autonomous neurodynamic model. It operates continuously and generates behavior in response to external stimuli, based only on the continuous evolution of activation patterns in the DNFs, and without any algorithmic structure controlling its operation. However, the system is very limited in the sense that its behavior is almost purely reactive and determined by the current input. While it does contain working memory (which is sometimes taken as one signature of what constitutes cognitive behaviors), the effects of this working memory in the context of the tasks treated here is quite subtle.

Nonetheless, one basic mechanism for behavioral flexibility are already foreshadowed here: Different behavioral regimes can be created in a fixed architecture with bidirectional connections by global modulations of fields activation level. This is used here to signal when working memory representation should be formed (by globally exciting the feature WM field), and to bring the system into an explicit visual search mode during the memory test phase of the trial (by globally exciting the feature attention field and thereby making surface feature match more dominant in coupled attentional selection). The later chapters of this thesis will build on this mechanism to generate more complex and flexible behaviors in larger DNF architectures.

Chapter 4

Spatial Transformations

4.1 Overview and motivation

After the previous chapter treated issues of separation and integration between spatial and surface feature representations, this chapter focuses on operations in spatial representations alone. In particular, I will address the issue of spatial reference frames. When describing a spatial location, it is always necessary to describe it relative to some reference frame: Relative to the own body, to another object, or to the world. When positions are given in different reference frames, it is necessary to map them into a common frame of reference before they can be combined or compared. The neural system continuously faces this problem, since it has to deal with different reference frames given by the different sensory modalities (in particular vision, hearing, and touch) and by the motor system.

Here, I will describe how spatial transformations, such as mappings between representations in different reference frames, can be realized within the framework of Dynamic Field Theory. Specific challenges arise here from the fact that spatial locations in DNFs are represented in the form of population codes, which are not compatible with standard algorithmic procedures for reference frame transformations. Moreover, to meet the requirements of autonomous process models, the transformation process has to be implemented not as a discrete operation performed at a specific point in time, but rather as a continuous coupling between different representations.

The concrete problem that I will address in this chapter is again taken from the field of active vision: How does the visual system deal with the shift of the visual image induced by every gaze change? To illustrate this problem, let us first take another look at the saccade mechanism described in the previous chapter. In this mechanism, the location of a visually perceived stimulus is mapped directly onto a saccade motor command. This simple direct mapping is possible because of special conditions in the oculomotor system. In particular, the motor system and the sensory input that

drives it use the same spatial reference frame. The locations of objects in visual space are initially perceived via the retina, so they are given relative to the current fixation point. The motor command for an eye movement is likewise given as a shift relative to the current fixation point (although downstream processes also have to take into account the initial state of the eye to generate the final signal to the muscles; Sparks, 2002). In addition, the kinetics of eye movements are particularly simple since the eyes can in good approximation be described as rotating spheres. Due to these factors, the planning of eye movements to visual stimuli can be achieved through a one-to-one mapping from each location on the retina to a certain movement command. As discussed before, this type of mapping can indeed be found in the brain.

However, this simple mechanism very quickly reaches its limits. Consider the case that you want to make an eye movement to a memorized location rather than to a salient stimulus in the current visual scene. For instance, you may have a cup of coffee standing on your desk while you are reading this text, and you want to look at the cup in preparation to grasping it. You saw the location of the cup before, likely you looked at it when you placed it on the desk. But while you are looking at the text, the cup may be outside of your field of view, or is perhaps only perceivable as a weak and unspecific stimulus in the visual periphery. So how does you visual system initiate a saccadic eye movement to the cup? The position of the cup's image on the retina while you were looking at it is not informative by itself, since you have made many eye movements in between, each of which shifted the whole visual scene. It is therefore necessary to take into account these different gaze directions when planning such saccades.

A common experimental task to test this ability under laboratory conditions is the double step saccade task (Hallett and Lightstone, 1976). In this task, subjects are fixating a point on a monitor, while two saccade target stimuli are briefly displayed in sequence. The subjects are then required to make a saccade to the memorized location of the first stimulus, and then another saccade from there to the second stimulus location. Again, to perform the second saccade accurately to the memorized location of the second stimulus, subjects have to take into account the gaze change that occurred between the stimulus presentation and the initiation of the second saccade. If subjects were simply using the retinocentric stimulus position to plan the second saccade, they would miss its actual location on the screen by the metrics of the first saccade. Experiments show, however, that subjects can perform both saccades reliably and with good accuracy under normal conditions (Heide et al., 1995).

One way how this ability may be achieved in the neural system is to use gaze-invariant representations. When the gaze direction is known at all times, it is possible to transform the retinocentric representation of a

stimulus position into a body-centered representation (and, if the body position in space is given, further into a fully allocentric spatial representation). However, this alone is still not enough to plan the second saccade in the experiment, whose amplitude and direction has to be specified relative to the current gaze direction. This can be achieved by transforming the location of the stimulus into a body-centered reference frame when it is first shown, based on the gaze direction at that time, and then transforming it back into the retinocentric frame of reference when planning the second saccade, using the new gaze direction after the first saccade has been completed. I will present a DNF model in this chapter that implements this general idea. Notably, however, the transformation process in the model is not realized as an operation that that is performed at discrete times, but rather as a continuous coupling between representations in different reference frames. Thus, the stimulus position is not transformed into the body-centered reference frame when it is seen and then back at a later time, but it is continuously represented in both reference frames during the period it is held in memory, with a flexible and bidirectional transformation mechanism maintaining the alignment between these representations.

Similar to what was presented in the previous chapter, the DNF model of reference frame transformation employs higher-dimensional neural fields to provide a combined representation of different variables, in this case retinocentric stimulus position and gaze direction. Unlike in the biased competition model, the association of these variables is not provided by an external input (such as a visual stimulus with a spatial position and a specific color), but it is generated by combining inputs from different sources. A third variable—namely, the stimulus position in a different frame of reference—can then be read out from the combined representation.

The general transformation mechanism that is implemented in the model can also be employed for other processes in which reference frame transformations are necessary. These include in particular the planning of limb movements and fusion of different sensory modalities. For instance, for a visually guided reaching movement, it is necessary to transform the retinocentric representation of the target location into a body-centered representation in order to determine the required arm configuration for the reach. It may moreover be necessary to perform additional transformations, for instance to determine the target location relative to the current hand position for trajectory planning. In sensor fusion, spatial transformations are needed to meaningfully integrate information from different sensory surfaces, such as visual (retinocentric) and auditory (head-centered) spatial information.

Below, I will first review neurobiological findings on spatial reference frames and transformations between them in neural systems, as well as existing theoretical models of these processes. I will then describe a basic mechanism for reference frame transformations in the framework of Dynamic Field

Theory, and show how this mechanism can be extended to a model of flexible coupling between spatial representations in different reference frames. This mechanism forms the basis for a concrete model of gaze-invariant spatial working memory and spatial updating during saccadic gaze shifts. I will show how this model can perform a memory-based saccade task and how activation patterns in the model can account for electrophysiological data from experiments on macaque monkeys. In doing so, the model reconciles two seemingly contradictory accounts of spatial working memory in the primate cortex. The concrete model and the results that I will show here have been published in Schneegans and Schöner (2012), and the underlying mechanism has been described in Schneegans (in press). This mechanism also provides an important building block for the models of scene representation and spatial language presented in the following chapters.

4.2 Biological basis and previous modeling work

4.2.1 Reference frames in neurophysiology

Numerous neural population representations of space have been identified in the cortical and subcortical areas that are involved in sensory processing, spatial attention, spatial memory, and movement planning, in particular for the planning of saccadic eye movements (Colby and Goldberg, 1999). In these population representations, each individual neuron has a spatial receptive field, that is, a spatial region within the visual field for which the neuron responds with strong activity if a stimulus is presented in it or a movement is planned to it. The receptive fields of the population as a whole cover the complete represented space. The reference frame of such a neural representation is the frame in which the receptive fields remain stable. For instance, for a retinocentric representation as it is found in most areas throughout the visual cortex (Gardner et al., 2008), the receptive fields are fixed when described in retinal coordinates (relative to the fovea). This means that a neuron from such a representation will always respond most strongly if a stimulus is presented in a certain direction and at a certain distance (expressed in visual angle) from the current fixation point, irrespective of the gaze direction. This necessarily means that with every gaze change, the receptive fields of these neurons are shifting with respect to every other frame of reference (such as head-centered, body-centered, or allocentric).

Note that when discussing the reference frames of neural populations, I deliberately avoid the terms 'coordinate system' or 'coordinate frame'. A coordinate system is formally defined by an origin and a set of axes or angles, such that a position in space can be described through a vector of real numbers. Neural population representations of space do not have a specified origin, nor do they define any axes, and therefore they cannot strictly be said

to have a coordinate system. Spatial locations are represented by activation distributions over the neural population and not by a set of real numbers. This also means that approaches to coordinate transformations that rely on arithmetic operations on the coordinates cannot be directly adopted for neural reference frame transformations.

In order to determine the reference frame of a neural spatial representation, it is necessary to repeatedly measure neural activity under different conditions (such as different gaze directions) that dissociate the candidate reference frames from each other. Such experimental work has shown that the spatial reference frames of most sensory representations in the brain reflect the corresponding sensory surface. The early stages of visual processing in subcortical and cortical structures show a clear retinocentric response pattern (as well as a retinotopic spatial organization). This retinocentricity is retained over large parts of the visual processing pathway, even in representations far removed from the sensory surface (Gardner et al., 2008). There are however also visual spatial representations that show coding in other reference frames, or in a mixture of different frames (Andersen et al., 1997; Snyder et al., 1998).

The tactile representations in the somatosensory cortex are aligned with the skin on the body surface, and they also show a somatotopic spatial organization (Kaas et al., 1979). The situation is somewhat more complex for auditory spatial perception. Unlike in the visual and somatosensory domain, where different locations map directly to different points on the sensory surface, the spatial information from auditory signals has to be extracted in a relatively complex manner from interaural time difference, interaural intensity differences and other cues. Auditory spatial representations can be found in the inferior colliculus and the auditory cortex. While these show at least in part a head-centered frame of reference, one often also finds significant influences from gaze direction (Groh et al., 2001; O'Dhaniel et al., 2005). This can be seen as a first step of a mapping to a retinocentric frame of reference.

Certain cortical areas, for instance parts of the parietal cortex, show activity in response to multiple different sensory modalities (e.g., both visual and auditory). For these multimodal areas, some neural representations with mixed reference frames have been described. Stricanne et al. (1996) investigated memory activity for auditory stimuli in monkeys, using different visual fixation points to separate the retinal and the head-centered frames of reference. They found that among the neurons with auditory memory activity in the lateral intraparietal area (LIP), the largest group actually showed activity in a retinal reference frame. Only a smaller number of neurons responded consistently with the head-centered reference frame that one would expect for auditory perception, and a third group appeared to employ a reference frame that was intermediate between the retinal and

head-centered frame, shifting to some degree with changing fixation points.

A similar observation was made by Avillac et al. (2005) for the nearby ventral intraparietal area (VIP). This cortical region receives both visual input and tactile input for the facial region, and many neurons are bimodal. While the tactile receptive fields for the face were found to be consistently in a head-centered reference frame, the reference frame for visual receptive fields was more varied in the bimodal neurons: Both retinocentric and head-centered receptive fields were identified, and some cells appeared to respond in an intermediate frame of reference: Their retinocentric receptive field centers shifted as the gaze direction changed, but they did not shift so much as to be stable with respect to the head.

4.2.2 Reference frame transformations through gain-modulated neurons

The presence of sensory spatial representations that are not in the reference frame of their associated sensory surface clearly indicates that mechanisms exist in the brain that allow a transformation between different spatial reference frames. A conjectured neural substrate for this transformation process has been identified in the form of gain-modulated neurons in the parietal cortex by Andersen and colleagues (Andersen and Mountcastle, 1983; Andersen et al., 1985). Such gain modulated neurons have later also been found in the frontal eye field, a region in the prefrontal cortex involved in the generation of saccadic eye movements (Cassanello and Ferrera, 2007). These neurons are visually responsive and have localized (although often broad) receptive fields in a retinocentric reference frame. But their overall response strength, or gain, is significantly modulated by the current gaze direction.

Neural populations in the parietal cortex comprise a large number of such neurons, with varied receptive fields and different gain modulations. Any visual stimulus excites many neurons at the same time, and their respective responses may be stronger or weaker depending on the current gaze direction. The activity of a single neuron in such a population is highly ambiguous regarding the location of a stimulus: A wide range of different combinations of retinal stimulus position, stimulus intensity, and current gaze direction will lead to the same activity level. The pattern of activity in the population, however, is unique for each combination of retinal stimulus position and gaze direction. It can therefore unambiguously be mapped onto the corresponding location in a head-centered representation. This has has been demonstrated in a number of neural network models (Zipser and Andersen, 1988; Pouget and Sejnowski, 1997).

The exact form of gain-modulation appears to be quite varied among neurons (Andersen et al., 1985). For some neurons, the modulation is best described by a linear (or, in two dimensions, planar) dependence on gaze direction. This means that the overall excitation evoked by a stimulus within

the neuron's retinocentric receptive field increases roughly linearly if the fixation point is shifted in certain direction. Other neurons have been described to have more localized gain fields, that is, their overall activity is maximal for a certain preferred gaze direction, and decreases with deviations in any direction. Some neurons cannot be clearly assigned to either of these categories (and often the classification remains ambiguous due to the limited range of tested gaze directions).

This variability is also reflected in computational models of the reference frame transformation process. Most use either purely linear (Pouget and Sejnowski, 1997) or purely local gain fields (Denève and Pouget, 2003) and different types of gaze-direction input are employed (either space code or rate code). In the DNF model below, I will focus on the variant with localized gain fields. This form has several advantages from a computational point of view: It allows an easier read-out of the transformed spatial information, using only excitatory projections with a simple connection pattern; it ensures that the representation in the gain-modulated population can be a stabilized state with localized activation peaks; and, as I will show below, it can be used in a straightforward fashion to perform a reference frame transformation on multiple perceptual items simultaneously.

4.2.3 Retinocentric remapping as an alternative to gaze-invariant representations

A separate line of electrophysiological experiments has found a surprising property of visual spatial representations in the cortex of macaque monkeys, namely in the LIP region. Duhamel et al. (1992) measured the visual responses of neurons around the time of saccadic eye movements. They determined the spatial receptive field of each recorded neuron, and observed a vigorous response when a visual stimulus was presented inside this receptive field. This response persisted for several hundred milliseconds after the stimulus was extinguished. A first interesting observation was that this visual response decreased sooner and more sharply when a saccadic eye movement was performed that moved the neuron's receptive field away from the stimulus location. This is unexpected because the visual sensory input to the neuron should be affected in the same way whether the stimulus is turned off or is moved out of the receptive field.

Even more interesting is the neural response pattern in a different stimulus condition. Here, the authors presented a stimulus and extinguished it immediately before a saccadic eye movement to a new fixation point was made. The stimulus location was chosen in such a way that it would be within the recorded neuron's receptive field *after* the completion of the eye movement. This means that the stimulus never actually appeared in the neuron's receptive field, but it would have appeared there if the stimulus had not been turned off directly before the saccade. In the this situation, the

neuron produced a brief burst of activity around the time of the saccadic gaze shift, even though it should never experience any visual stimulation. The same response behavior has also been reproduced for neurons in spatial representations in the FEF by Sommer and Wurtz (2006). In many cases, these neurons start to show activity even before the saccade is completed.

Colby and colleagues described this phenomenon as peri-saccadic shift of the spatial receptive fields of LIP neurons. This is potentially misleading as it may imply that the origin of the observed effect lies in a change or modulation of the feed-forward connectivity from the retina to the investigated neurons. In fact, the underlying mechanism is still unclear, and Cavanagh et al. (2010) have argued that it should more properly be described as a remapping of attentional pointers in the retinocentric representation. I will therefore refer to this phenomenon as retinocentric remapping in the following.

This phenomenon of retinocentric remapping offers a mechanism for trans-saccadic spatial working memory that does not depend on an explicitly gaze-invariant representation. A purely retinocentric spatial representation can be used that is updated during every gaze change in such a way that all memorized locations are shifted by the inverse of the saccade metrics. Using such a representation, tasks like the double step saccade may be performed without using any reference frame transformations. This does, however, require information about the metrics of the upcoming saccade in order to determine the appropriate shift in the retinocentric representation. This information may be obtained from an efference copy of the saccade motor command, and experimental evidence indicates that such an efference copy is indeed needed to produce the remapping. Specifically, when a synaptic pathway from the superior colliculus (where the saccade signal is generated) to the FEF is interrupted, the remapping activity during a saccadic eye movements is no longer observed in that cortical region (Sommer and Wurtz, 2008).

The proposed mechanism of achieving trans-saccadic spatial working memory through retinocentric remapping has been implemented in a neural network model Quaia et al. (1998). Their model assumes a population representation of visual space in a retinocentric reference frame. Working memory in this neural map is modeled as neural activity sustained through self-excitation (analogous to a two-dimensional neural field with self-sustained activation peaks). The neurons within this map are connected to each other by means of lateral projections, implementing an all-to-all connection pattern. These projections are gated by an input that reflects the efference copy for a saccadic eye movement. If, for instance, a 5° horizontal saccade to the right is executed, all lateral connections in this map are activated that project from some point in visual space to another point 5° to the *left* of it. This implements the basic remapping of activity. Additional control mechanisms are used in this neural network to ensure that the activity is

extinguished at the old location, sustained at the new location, and that only exactly one remapping operation is performed during every saccade. The model was successful at demonstrating a functional neural mechanism for retinocentric remapping and at reproducing observed patterns of neural activity.

Based on the experimental evidence and these modeling results, some researchers argued that retinocentric remapping provides a sufficient explanation for trans-saccadic spatial memory, making accounts based on gaze-invariant representations unnecessary (Colby and Goldberg, 1999; Wurtz, 2008). The remapping approach is appealing in that it relies only on retinocentric representations, which are known to predominate in the visual system for sensory processing, spatial attention and eye movement control. Nonetheless, reference frame transformations would still be needed for sensor fusion and sensory-motor mapping. Moreover, some modeling efforts have indicated that activity patterns consistent with retinocentric remapping can also appear as the result of bi-directional reference frame transformation. Xing and Andersen (2000) presented a model for sequential saccades, using reference frame transformations based on gain-modulated neurons. The model uses separate representations for the first and second target in a double step saccade task, and an effect consistent with remapping occurs during the first saccade for the target of the second saccade. Comparable observations were also reported in the model of White III and Snyder (2004).

The DNF model that I will present in this chapter expands on these works, and can provide a full account of the experimental observations regarding retinocentric remapping in a framework based on neural reference frame transformations. Through the continuous coupling to a gaze-invariant representations, stimulus locations are autonomously remapped within the retinocentric representation. This can occur for multiple perceptual or memory items simultaneously, independent of whether or not they have been selected as saccade targets. The model also addresses the origin of the gaze direction signal used in the transformation. By using an internal representation of gaze direction that is predictively updated based on an efference copy of saccade motor commands, the model can account for the time course of the remapping activity and its dependence on signals from the saccade motor system.

4.3 Basic mechanisms

4.3.1 DNF model of reference frame transformation

In this section, I will describe the implementation of a basic reference frame transformation mechanism in a DNF architecture. The general mechanism is analogous to previous radial basis function models of reference frame trans-

formations (Pouget and Sejnowski, 1997). This system performs a transformation from a retinocentric visual representation to a body-centered representation, using a signal that specifies the current gaze direction. This signal combines the orientation of the head and the orientation of the eyes within the head into a single variable, giving the gaze direction relative to the forward direction of the body. The reference frame transformation then corresponds to a variable shift of object locations in the two-dimensional visual representation.

Algorithmically, the reference frame transformation for an individual stimulus location can be performed as a simple vector addition. To this end, the retinocentric stimulus position in the two-dimensional visual image and the gaze direction (as the position of the fixation point within the two-dimensional visual scene) are both described as vectors in units of visual angle. The body-centered position $\vec{p}_b$ of a stimulus can then be determined as the sum $\vec{p}_b = \vec{p}_r + \vec{p}_g$ of the retinal position $\vec{p}_r$ and the current gaze direction $\vec{p}_g$ relative to the forward direction of the body. In the DNF model, this operation has to be implemented through fixed synaptic projections between different population representations.

I will first introduce the reference frame transformation mechanism for one-dimensional spatial information, dealing only with horizontal visual stimulus positions and gaze direction. The DNF model architecture for this case is shown in Figure 4.1. The model contains two DNFs that represent sensory inputs. The *retinal field* is defined over retinocentric visual space, and forms activation peaks for the angular positions of salient visual stimuli. The evolution of activation u_r in this field is governed by the differential equation

$$\tau \dot{u}_r(x) = -u_r(x) + h_r + i_r(x) + [k_{rr} * f(u_r)](x), \tag{4.1}$$

with external input $i(x)$ reflecting the locations of salient visual stimuli and a lateral interaction kernel k_{rr} of the difference-of-Gaussians type.

The second DNF is the *gaze field*, which spans the space of horizontal gaze directions relative to the straight forward direction of the body. It is governed by a similar differential equation

$$\tau \dot{u}_g(y) = -u_g(y) + h_g + i_g(y) + [k_{gg} * f(u_g)](y). \tag{4.2}$$

The main difference between the fields is in the shape of the interaction kernels, with the kernel k_{gg} featuring strong global inhibition to allow only a single activation peak at any time. This reflects the fact that the gaze direction should always take a single value, whereas in the retinal field multiple stimuli may be represented simultaneously. The retinal field and the gaze field are initially independent of each other, since they receive input form separate sensory systems.

To bring these two representations together and to capture the function of the gain-modulated neurons in the parietal cortex, a two-dimensional *transformation field* is defined spanning both the dimension of retinal position

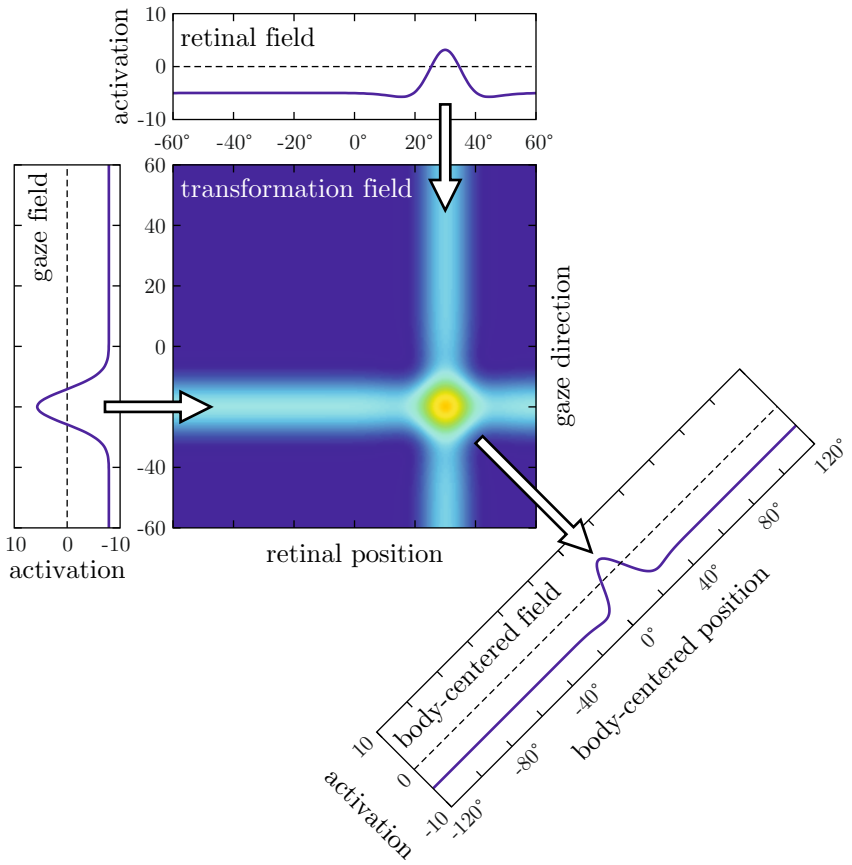


Figure 4.1: DNF architecture for a reference frame transformation from a retinocentric to a body-centered spatial representation, using the current gaze direction. One-dimensional fields are shown as activation plots over their respective feature spaces, the two-dimensional transformation field is shown as color-coded activation distribution. The feature space axes of the one-dimensional fields are aligned with the corresponding axes of the two-dimensional fields, the projections between fields (white arrows) run perpendicular to them.

and gaze direction. The connectivity between the fields is set up as shown in Figure 4.1. In the figure, the one-dimensional fields are aligned to the corresponding dimension of the transformation field, with the retinal stimulus position on the horizontal axis, the gaze direction on the vertical axis. The one-dimensional fields now project ridge inputs into the transformation field, in the same way as described in the previous chapter. The activation peak in the retinal field induces a vertical ridge input, the activation peak in the gaze field induces a horizontal ridge input in the transformation field.

The connectivity is captured in the field equation for the transformation field u_t as

$$\begin{aligned}\tau \dot{u}_t(x,y) = &- u_t(x,y) + h_t + [k_{tr} * f(u_r)](x) \\ &+ [k_{tg} * f(u_g)](y) + [k_{tt} * f(u_t)](x,y)\end{aligned} \tag{4.3}$$

As in all DNF models described here, the projections between fields are mediated by interaction kernels that smooth the output of one field before it is fed into another field. Here, the interaction kernels k_{tr} and k_{tg} are one-dimensional Gaussians, while the lateral interaction kernel k_{tt} is a difference of Gaussians defined over two dimensions.

The pattern of projections from two one-dimensional fields to a single two-dimensional field in the transformation mechanism is analogous to the basic space-feature association mechanism described in the previous chapter. There is a key difference between these two architectures though: The two-dimensional field in the space-feature association architecture receives direct external input that is localized in both dimensions, reflecting the existence of visual feature detectors with localized spatial receptive fields. The ridge inputs from the one-dimensional fields only modulates this activation pattern and strengthens specific existing peaks. In the reference frame transformation system, the ridge inputs are the only inputs to the two-dimensional transformation field. Since there is no sensory system that directly detects a combination of gaze direction and visual stimulus position, there are no localized inputs to the transformation field.

Activation peaks form in the transformation field at the intersection points of one horizontal and one vertical ridge input. In the field equations, all inputs are combined additively, so that the field activation is driven to the highest levels at those points where two ridge inputs come together. To obtain peaks only at these intersection points, the connection parameters are chosen in such a way that the activation induced by each individual input ridge remains below the field's output threshold, but the activation induced by the combination of two such inputs is sufficient to pierce the threshold. Lateral interactions within the two-dimensional transformation field then drive the formation of stabilized activation peaks at these intersection points.

The position of such a peak in the two-dimensional transformation field combines the retinocentric position of a visual stimulus and the current gaze direction in a single representation. It therefore provides all information needed to determine the stimulus position in a body-centered reference frame (as $\vec{p}_b = \vec{p}_r + \vec{p}_g$). The body-centered representation is implemented in the DNF architecture as another one-dimensional DNF, plotted diagonally in Figure 4.1. The reference frame transformation can now be implemented as a fixed synaptic connection pattern from the transformation field to this *body-centered field*. From each point in the transformation field specifying a

pair $[\vec{p}_r, \vec{p}_g]$, a projection is defined to the corresponding position $\vec{p}_b$ in the body-centered field. The field equation for the body-centered field u_b with this input then takes the form

$$\tau \dot{u}_b(z) = -u_b(z) + h_b + [k_{bt} * F_b(u_t)](z) \tag{4.4}$$

with

$$F_b(u_t)(z) = \int f(u_t(x, z - x))dx. \tag{4.5}$$

The projections from the transformation field to the body-centered field form a specific geometric pattern, which can be seen as follows: Assume that there is a single visual stimulus in the scene, and that you are initially fixating it. Now you shift your gaze to the right in little steps by making a series of small saccades. With each small saccade, the gaze direction shifts to the right by a certain amount, while the retinocentric stimulus position shifts to the left by that same amount. In the DNF architecture as shown in Figure 4.1, this means that the peak in the gaze field shifts upward and the peak in the retinal field shifts to the left by the same amount. Consequently, the intersection point of the activation ridges in the transformation field shifts along a diagonal line through that field (this may also be seen directly from the relationship $\vec{p}_b = \vec{p}_r + \vec{p}_g$). The points on this diagonal reflect all possible combinations of gaze direction and retinocentric stimulus position that correspond to a certain body-centered position, and therefore project to the same position in the body-centered field. The body-centered field is drawn perpendicular to this projection, analogous to the other one-dimensional fields in the architecture.

Note that this geometry of a diagonal projection only reflects the relationship between the different feature dimensions in the model, and does not reflect any physiological properties in the biological neural system. Moreover, analogous mechanisms may also be implemented for operations other than the simple addition of two inputs, as long as the result varies smoothly with changes in the input. The read-out projection then takes a different and potentially more complex shape than the diagonal projection employed here, but the operation principle remains the same.

The DNF architecture now performs the reference frame transformation in a temporally continuous and autonomous fashion. A single gaze direction peak is assumed to be present at all times, changing its position when gaze shifts occur (typically only during saccades). When a visual stimulus appears, it induces a peak in the retinal field, which automatically drives peak formation first in the transformation field, then in the body-centered field. If the stimulus moves, the activation peaks shift accordingly (albeit with slight delay) to reflect its latest position. Conversely, if the stimulus remains fixed but the gaze direction changes, then the peaks in the retinal field and gaze field shift to new locations, the peak in the transformation field shifts along

the diagonal, but the peak in the body-centered field remains at its original location (although it may fluctuate in strength during this process).

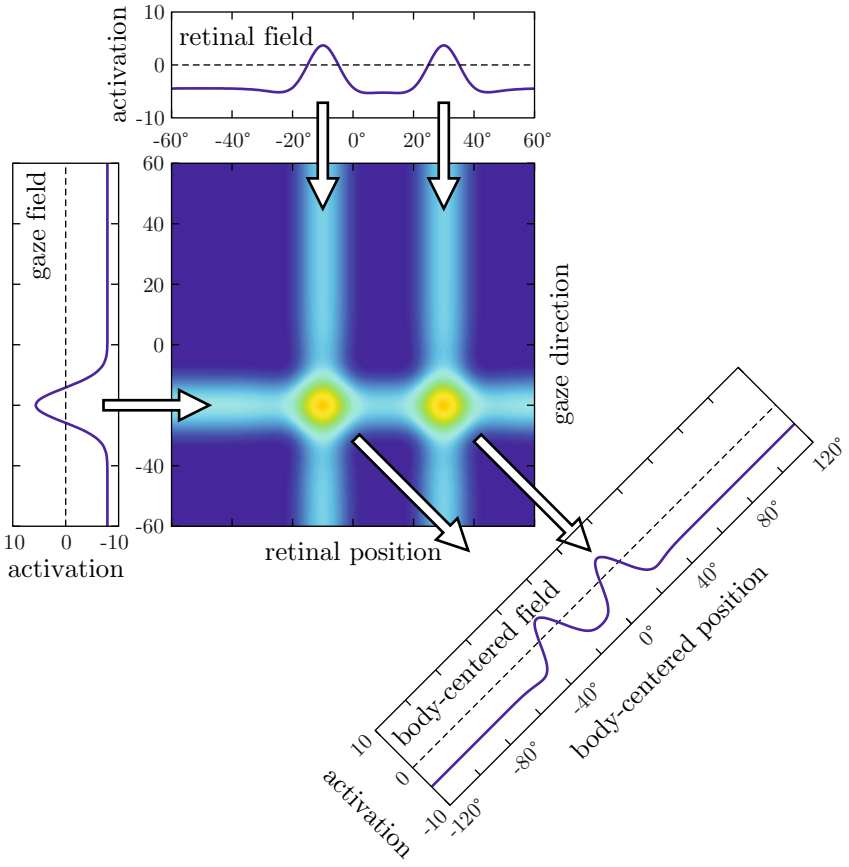


Figure 4.2: Simultaneous mapping of two retinal stimulus positions into a body-centered spatial representation in the DNF model.

Moreover, when set up with appropriate lateral connection patterns, the DNF architecture may perform the transformation for multiple stimulus positions in parallel, as illustrated in Figure 4.2. This requires that lateral inhibition in the retinal field, the transformation field, and the body-centered field is of the local surround type (rather than global), such that multiple activation peaks may coexist in these fields without competition between them. When two or more peaks are now present in the retinal field, they project parallel vertical ridge inputs into the transformation field. These ridges intersect at different locations with the single horizontal ridge from the gaze field, and two separate peaks form in the transformation field. These in turn project in parallel to the body-centered field, and again induce two activation peaks that reflect the positions of the visual stimuli relative to the

body. To some degree, the transformation mechanism can also retain intensity information of different peaks, in that a larger stimulus-induced peak in the retinal field will also create a larger peak in the body-centered field. Differences in peak size are reduced, however, by the normalizing effects of lateral interactions at all stages of the transformation.

4.3.2 Transformation operations in different directions

The mechanism described so far implements the transformation from a retinocentric to a body-centered frame of reference. It thereby provides a gaze-invariant representation of visually perceived spatial locations. But as discussed in the beginning, to fully utilize this representation we also need a way to perform the inverse transformation, from the body-centered to the retinocentric reference frame. This is necessary, for instance, to initiate a saccadic eye movement to a location memorized in a body-centered representation, since the motor command is specified in a retinocentric format (relative to the current gaze direction) in the saccade motor system.

This inverse transformation can be implemented in a DNF architecture that is analogous to the one described in the previous section. The only difference is that the directions of certain synaptic projections are reversed. The resulting architecture is shown in Figure 4.3. If an activation peak is present in the body-centered field of this system, it projects a diagonal ridge input into the transformation field. The single gaze direction peak that is assumed to be present at all times still projects a horizontal ridge input into the transformation field (as in the previous setting), and an activation peak forms in the transformation field at the intersection of these two ridges. By the same geometric considerations as before, it is clear that the horizontal position of this peak corresponds to the retinocentric position $\vec{p}_r$ for which the relationship $\vec{p}_b = \vec{p}_r + \vec{p}_g$ holds. To explicitly represent this retinocentric location information, the transformation field is read out by integrating over its vertical dimension, and the result is projected into the retinal field. This yields for the retinal field the modified differential equation:

$$\tau \dot{u}_r(x) = -u_r(x) + h_r + i_r(x) + [k_{rt} * F_r(u_t)](x) + [k_{rr} * f(u_r)](x) \quad (4.6)$$

with

$$F_r(u_t)(x) = \int f(u_t(x,y))dy. \quad (4.7)$$

The reverse transformation now provides a mechanism to project either memorized location information or spatial information from body-centered sensory modalities (such as touch) into the retinocentric reference frame of vision. It shares all the features of the forward transformation described above. It provides a continuous coupling between spatial representations, supporting online updating and tracking of moving stimulus locations. It

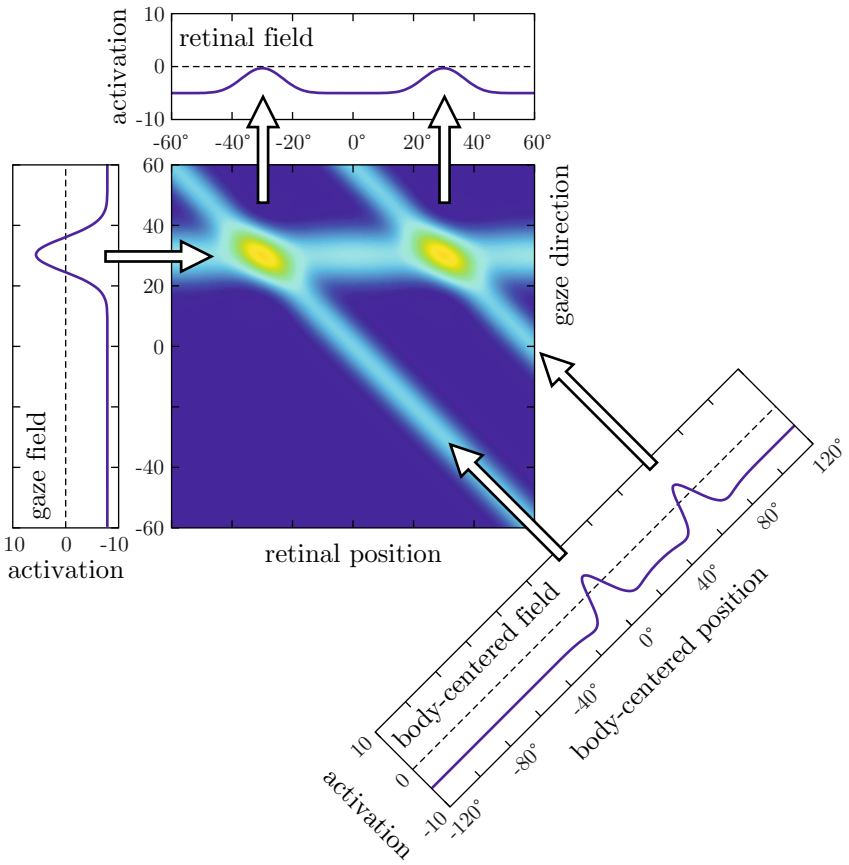


Figure 4.3: DNF architecture for a backward transformation from a body-centered representation into a retinocentric representation of space, applied to two stimulus positions simultaneously.

can be applied to multiple locations in parallel (as shown in Figure 4.2), with each peak in the body-centered field sending a separate diagonal ridge into the transformation field and ultimately inducing one peak in the retinal field. And it allows for propagation of graded information (intensity of activation peaks) to a certain degree.

There is a third possible direction of transformation that can be implemented by reversing projections in this architecture. Assume that you have memorized a visual scene containing a few objects in a certain spatial arrangement, and now view the same scene again, but under some different gaze direction. You can now perform an *alignment* operation on the memorized (body-centered) pattern and the currently perceived (retinocentric) pattern of spatial locations to determine how they are shifted against each other and thereby estimate your own gaze direction. At first glance, it may

appear unnecessary to use such matching processes for estimating your own gaze direction, because it can be determined just from proprioceptive signals indicating the current states of muscles and joints. But it is critical for many spatial cognition tasks to keep different spatial representations exactly aligned, and experiments have shown that humans do indeed use landmarks perceived as stable to estimate the metrics of their gaze changes (Deubel et al., 1998; Deubel, 2004).

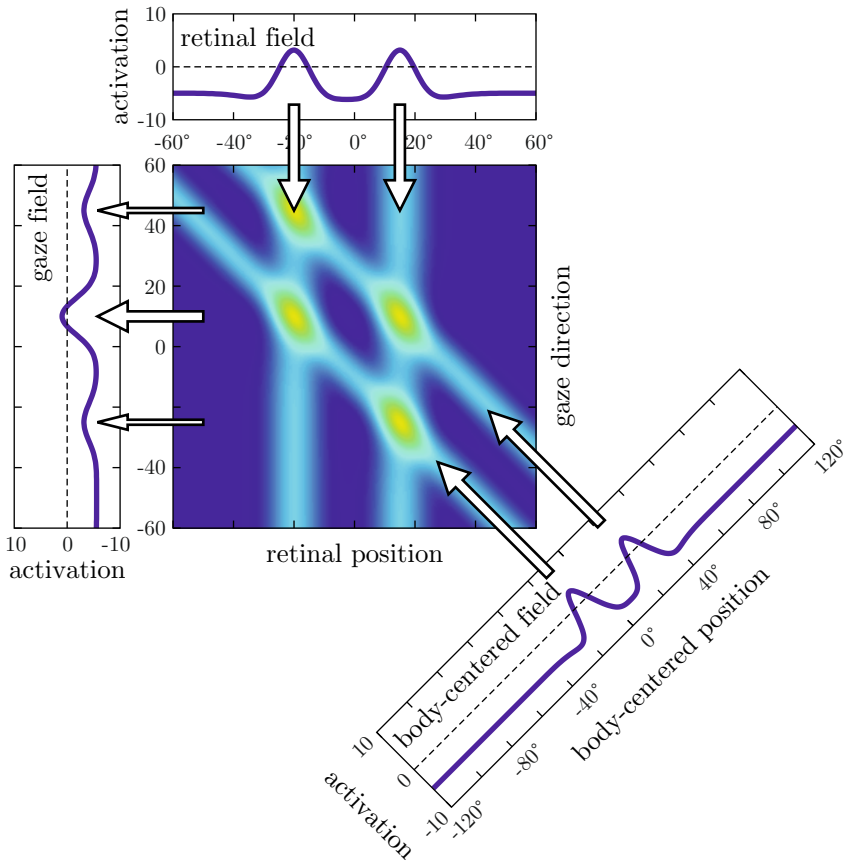


Figure 4.4: DNF architecture for finding the alignment between a retinocentric and a body-centered spatial representation and thereby estimating the current gaze direction.

An alignment mechanism can be set up in the architecture by having both the retinal field and the body-centered field project into the transformation field, and having a horizontal read-out of the transformation field into the gaze field (Figure 4.4). The field equation for the gaze field is then adjusted

to

$$\tau \dot{u}_g(y) = -u_g(y) + h_g + i_g(y) + [k_{rt} * F_g(u_t)](y) + [k_{gg} * f(u_g)](y) \quad (4.8)$$

with

$$F_g(u_t)(y) = \int f(u_t(x, y))dx. \quad (4.9)$$

Assuming that there is only a single peak in both the retinal field and the body-centered field, the process is entirely analogous to the two other connectivities discussed before. The retinal field projects a vertical ridge of activation into the transformation field, the body-centered field induces a diagonal activation ridge, and a peak forms at the intersection point. The vertical position of this peak then corresponds to that gaze direction that brings the two locations into alignment (that is, under which the given retinal location would correspond to the given location in the body-centered reference frame). Through the read-out projection, a peak for this gaze direction is induced in the gaze field.

The situation is slightly more complex if there are multiple peaks in the retinal field and the body-centered field (the situation shown in Figure 4.4). In this case, both fields induce multiple activation ridges in the transformation field, and peaks form at all possible intersection points between any pair of ridges from the two fields. For each of these intersection points, its vertical position reflects the gaze direction that would bring the two corresponding locations in the retinal field and the body-centered field into alignment. But many of these intersections are spurious: They reflect, for instance, an alignment between the right-most peak in the retinal field and the left-most peak in the body-centered field. However, if the pattern of peaks in the body-centered field is indeed an accurate memory of the stimulus pattern currently represented in the retinal field, then the patterns in the two fields will be shifted versions of each other. In this case, there must be one gaze direction that brings each retinal peak into alignment with the corresponding body-centered peak. At this gaze direction, the maximum number of intersection points will be lined up, as shown in Figure 4.4.

The input to the gaze field will be strongest for this gaze direction, since here the outputs of all the aligned peaks add up. If the gaze field is set up with local self-excitation and global inhibition (as was already assumed above), then a selection decision will take place in the field. A peak forms at the location of the strongest input, and suppresses activation everywhere else in the field. This selection decision will reliably yield the correct alignment between the two peak patterns if they are indeed shifted versions of each other. It is also robust against moderate deviations from this assumption (e.g, if the memory in the body-centered field is inaccurate or individual items in the scene have changed their position), and will still produce a reasonable estimate of the correct gaze direction under such conditions.

4.3.3 Transformations with multi-directional coupling

It is possible to merge the three different transformation mechanisms, which already share the same representations, into a single architecture. To this end, the projections between the fields are made bidirectional. Each one-dimensional field projects a ridge input into the transformation field—horizontal, vertical, or diagonal—and in turn receives input from the transformation field that is computed by integrating along the same direction. In such a system, there is no longer a predetermined direction of the transformation process. Each one-dimensional field acts as a potential input for the transformation, and each can provide a result. Which role each field actually takes is no longer determined by the connectivity in the architecture, but by the activation states and external inputs of the fields.

How such a system can be used has been demonstrated by Denève et al. (2001). Their neurodynamic architecture realizes the same basic mechanism I have sketched here for a DNF model, with neural fields for retinal positions, body-centered positions, and gaze direction coupled to each other in a bidirectional fashion. There are however several differences in the implementation. In particular, they do not use an actual two-dimensional neural representation for the coupling, and employ divisive normalization instead of lateral inhibition to control the spread of activation (which allows only a single stable peak to exist in each representation). The model is operated in the following manner: An initial activation pattern is specified for each one-dimensional field. This is typically a noisy or ambiguous representation of a stimulus position or gaze direction. It is also possible to initialize one of the three fields with a completely flat activation pattern. Then the activation patterns are allowed to evolve under the influence of mutual interactions until they have settled into a stable state, which can be read out as the result of the operation. If all fields are initialized with noisy patterns, the system performs a cue integration that yields a smooth, mutually consistent representation in all fields. If one field is initialized with a flat activation distribution, it will be filled according to the transformation operation from the two inputs that are provided.

While this work of Denève et al. (2001) demonstrates the capabilities and the flexibility of the multi-directional transformation system, its mode of operation is not actually compatible with a fully autonomous dynamical system. It requires that inputs are fed into the three fields synchronously at a fixed time, then the system is allowed to run on its own, and the termination (after settling into an attractor state) has to be determined by an external process. So even though it is formulated as a temporally continuous dynamical system, it effectively performs a single, discrete operation. This contrasts with a system such as the biased competition model presented in the previous chapter, which is coupled directly to simulated sensory and motor systems and can continuously receive input and generate behavior.

Operating a DNF model of multi-directional transformation in such a fully autonomous mode is problematic because of conflicting constraints. First, to allow transformations in all directions, excitatory coupling has to be strong enough such that if input is provided to two of the one-dimensional fields, an activation peak will be elicited in the third one. Second, activation levels must be held in check once peaks are present in all three activation fields, which all project input to the transformation field and reinforce each other. This requires sufficiently high levels of lateral inhibition in the fields. But now in such a strongly coupled system with strong lateral interactions, the importance of external input beyond the initial state is greatly diminished. Once the system has reached a stable state, it will tend to persevere in this state, and will no longer be sensitive to changes in input patterns.

To avoid these problems, the autonomous neurodynamic model of gaze-invariant working memory I now present does not implement a fully multi-directional transformation mechanism. Instead, the gaze field is coupled only unidirectionally, and for the retinocentric representation, inputs to and outputs from the transformation field are held separate to reduce back-coupling. The full multi-directional transformation will however be used in Chapter 6 to support the flexible use of relational spatial language.

4.4 DNF model of retinocentric remapping

4.4.1 Model overview

Based on the general DNF mechanisms for reference frame transformation, I will now describe a specific neurodynamic model for gaze-invariant spatial working memory. The goal of this model is two-fold: On the one hand, it should provide a functional and autonomous system for spatial working memory, in which new visual stimulus positions can be entered at any time. These memorized stimulus positions should then be available at all times in both the current retinal and the body-centered reference frames, independent of intervening gaze changes. On the other hand, the model should explain the neurophysiological findings about peri-saccadic remapping, which indicate a shift in retinocentric spatial representations preceding every saccade. An overview of the model architecture is shown in Figure 4.5.

The model architecture introduces a number of extensions and modifications of the basic transformation mechanism. In order to more fully capture the properties of visual space, two spatial dimensions are covered both for stimulus positions and gaze directions. Consequently, the retinal field, the gaze direction field, and the body-centered field are all two-dimensional. This requires that the transformation field, which combines spatial position and gaze direction, now has to span a four-dimensional space. The gaze-direction field and the retinal field in this system project to the transformation field,

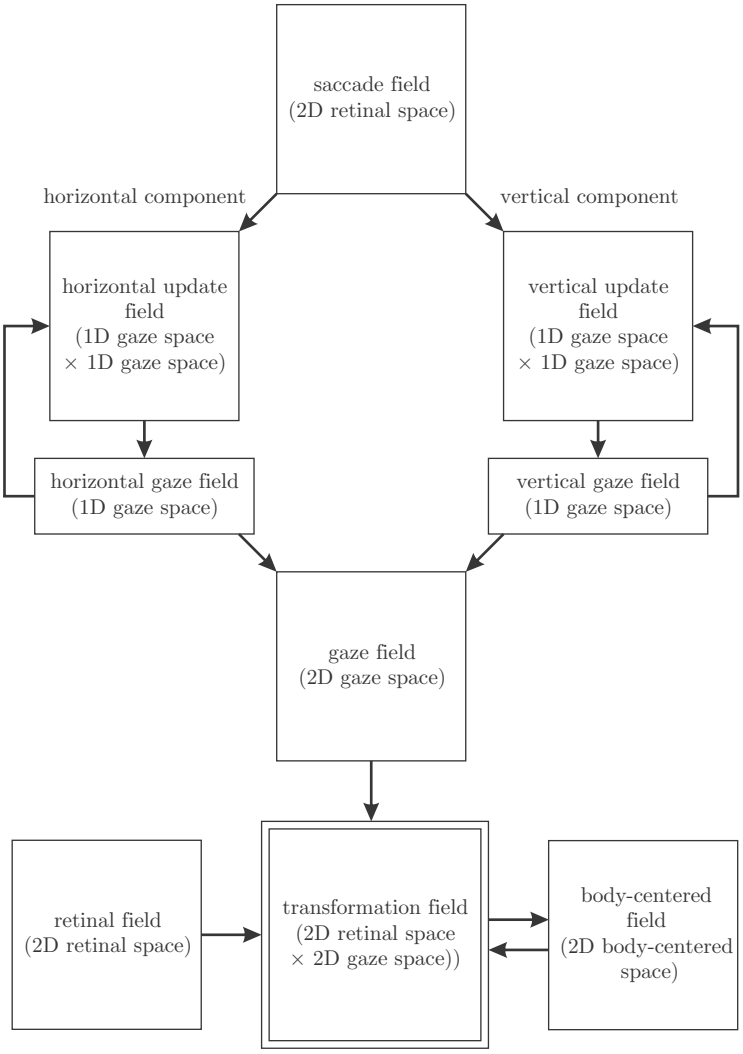


Figure 4.5: Overview of the DNF architecture for gaze-invariant spatial working memory and retinocentric remapping. The elements of the architecture are shown as boxes with the name of the DNF and the feature space over which each is defined. Arrows indicate projections between these fields along shared feature spaces.

but do not receive input back from it. The body-centered field is coupled to the transformation field in a bi-directional manner. In addition, there is a retinal readout of the transformation field that shows the locations of all memorized positions in the current retinocentric reference frame.

The model also incorporates an additional subsystem that is serves to

update the internal representation of gaze direction during a saccadic eye movement. This is motivated by the aim to account for the electrophysiological findings on retinocentric remapping. As described above, the shift in the retinocentric representation occurs before a saccadic eye movement is completed, and it depends on an efference copy of the saccade motor command. In the present model, such an efference copy is used to update an internal representation of gaze direction. The update corresponds to adding the saccade metrics to the current gaze direction, and it is implemented in a mechanism analogous to the reference frame transformation.

There are two important differences here, however. First, the gaze direction field appears in two different roles in this operation, providing one of the inputs but also receiving the result of the update. A specific sequence of instabilities in the field dynamics is employed to ensure that the update process is executed stably under these conditions. Second, in the gaze update, the two spatial dimensions are split up, and horizontal and vertical gaze shift are treated separately. This is possible because there is always just a single gaze direction at all times, so there can be no confusion which horizontal gaze direction value belongs to which vertical value. This is different in the reference frame transformation, where multiple stimulus locations can be present simultaneously, and a separate processing of the two spatial dimensions would result in a binding problem.

The following subsections first describe the subsystem for the reference frame transformation and retinocentric remapping, then the subsystem for the predictive gaze update that supports the remapping operation.

4.4.2 Remapping subsystem

The elements of the remapping subsystem are identified with the same indices as used in the description of the general mechanism above: r for retinal field, g for gaze field, t for transformation field, and b for body-centered field. The parameter values of fields and lateral interactions are given in Table 4.1, the parameter values for projections between fields are given in Table 4.2.

The retinal field is defined over two-dimensional visual space, covering a range of -30° to 30° from the fovea both horizontally and vertically. It is governed by the differential equation

$$\tau \dot{u}_r(x,y) = -u_r(x,y) + h_r + i_r(x,y) + [k_{rr} * f(u_r)](x,y). \tag{4.10}$$

The field receives visual input i_r providing the locations of salient visual stimuli. During saccadic eye movements, this input is suppressed. Lateral interactions in the field are described by a difference-of-Gaussians kernel k_{rr}, which allows the simultaneous presence of multiple activation peaks to reflect the locations of different visual stimuli.

The gaze field is analogously defined over a two-dimensional space of gaze-directions, ranging horizontally and vertically from -30° to 30° relative

to the straight forward direction of the body. Its field equation is

$$\tau \dot{u}_g(p,q) = -u_g(p,q) + h_g + i_g(p,q) + [k_{gg} * f(u_g)](p,q). \tag{4.11}$$

The input i_g indicates the current gaze direction. It is provided by two one-dimensional gaze fields (for horizontal and vertical gaze direction) that form a part of the gaze update subsystem, which is described below.

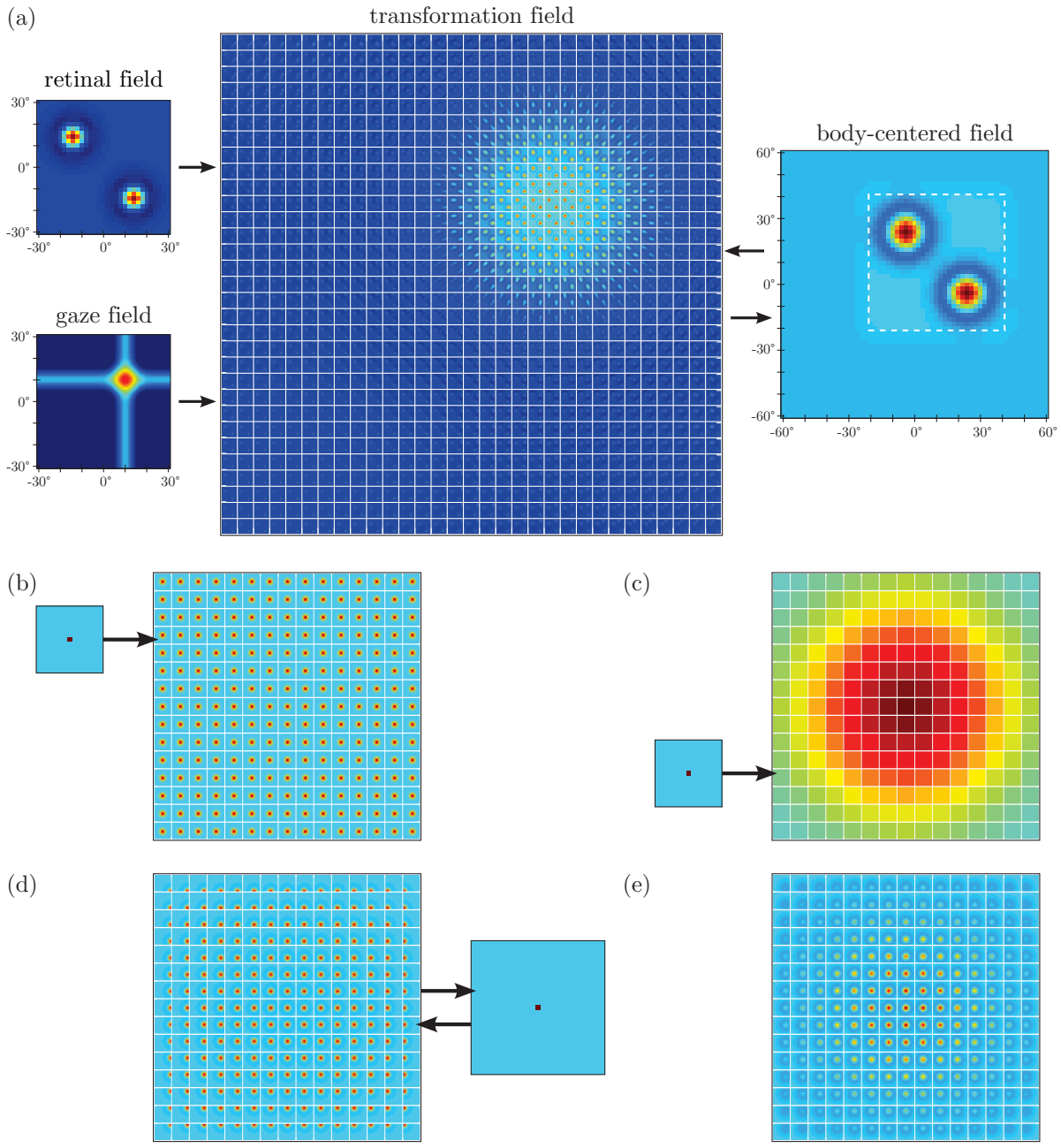
The transformation field is defined over a four-dimensional space, spanned by the two retinal and two gaze dimensions as defined above. The field receives input from the retinal field, the gaze field, and the body-centered field, u_b, specified below. This yields the field equation

$$\begin{aligned} \tau \dot{u}_t(p,q,x,y) = &- u_t(p,q,x,y) + h_t + [k_{tg} * f(u_g)](p,q) \\ &+ [k_{tr} * f(u_r)](x,y) + [k_{tb} * f(u_b)](p+x, q+y) \\ &+ [k_{tt} * f(u_t)](p,q,x,y). \end{aligned} \tag{4.12}$$

The projection patterns from the two-dimensional fields to the four-dimensional field are illustrated in Figure 4.6. They are natural extensions of the patterns described for the general mechanism in the lower-dimensional case above: The projections of the retinal field are localized along the two retinal dimensions and homogeneous along the gaze dimensions. Conversely, the projections from the gaze field are localized along the gaze dimensions and homogeneous along the retinal dimensions. Finally, the projections from the body-centered field are localized along the retinal dimensions for each single gaze direction, with the location systematically shifting between different gaze directions.

The projection weights for the different inputs are scaled to fulfill the following condition: Peaks should only form at the intersection point of the gaze input with either a retinal input or a body-centered input (or both of them). To achieve this, the gaze input is weighted approximately twice as strong as the two other inputs, but still not so strong that it can drive the transformation field activation beyond the output threshold by itself. The retinal and body-centered inputs are weaker, so that an intersection between them is not sufficient to induce activation peaks (these would correspond to the spurious intersections in the alignment mechanism described earlier).

The lateral interactions in the transformation field are of the difference-of-Gaussians type along both the retinal and the gaze dimensions, so that multiple activation peaks may persist in the field simultaneously. To approximate the properties of gain-modulated neurons in the parietal cortex, the interactions are broader along the gaze dimensions (and likewise, the inputs from the gaze field are broader than the retinal inputs), so that each single point in the transformation field can be active for relatively wide range of different gaze directions. To ensure that these wide lateral interactions can still effectively support activation peaks driven by input from the body-centered field, a special modification is applied: The kernel $k_{tt}(p,q,x,y)$ is


(a)
transformation field
retinal field
gaze field
body-centered field
30°
0°
-30°
60°
-60°
(b)
(c)
(d)
(e)

rotated in the px- and in the qy-plane by an angle of $\phi = \frac{3}{16}\pi$ toward the diagonal axes along which the body-centered field projects.

The body-centered field is defined over two-dimensional space in a body-centered reference frame, covering twice the range of the retinal field (-60° to 60° from the forward direction in both dimensions). This extended range means that the field can capture all body-centered locations that correspond to any combination of positions in the gaze field and the retinal field—such as when a stimulus appears in the top left in the retinal field while gaze is also directed to the top left. The body-centered field is governed by the field equation

$$\tau \dot{u}_b(x,y) = -u_b(x,y) + h_b + [k_{bt} * F_b(u_t)](x,y) + [k_{bb} * f(u_b)](x,y). \quad (4.13)$$

Here, $F_b(u_t)$ is the output of the transformation field, integrated along the diagonals:

$$F_b(u_t)(x,y) = \iint f(u_t(p,q,x-p,y-q))dpdq \quad (4.14)$$

The lateral interactions, described by the kernel k_{bb}, are again of the difference-of-Gaussians type to allow multiple peaks to co-exist.

Finally, to obtain the retinocentric positions of memorized stimuli, the transformation field is integrated over the two dimensions of gaze direction. This yields the retinal read-out

$$F_r(u_t)(x,y) = \iint f(u_t(p,q,x,y))dpdq. \quad (4.15)$$

Figure 4.6 *(preceding page)*: Activation patterns and connectivity in the reference frame transformation model for two-dimensional visual space. (a) Activation patterns in the different DNFs in response to two localized stimuli presented to the retinal field. The activation pattern in the four-dimensional transformation field is depicted through a set of tiles, each reflecting a two-dimensional cut through the four-dimensional space. Each individual tile shows the activation distribution over the two dimensions of retinal position for one specific gaze direction, and the tiles are arranged according to this gaze direction along both x- and y-axis. (b) Connection pattern from one point in the retinal field (marked as red dot) into the transformation field. Connection weights are color coded (light blue for connection weight of zero, red for highest weight). Field sizes along all dimensions are reduced for clarity. (c) Connection strengths from one point in the gaze field to the transformation field. (d) Connection strengths between the transformation field and one point in the body-centered field (bi-directional). (e) Connection strengths for lateral interactions in the transformation field, showing connections to and from the central point in the central tile of this field (darker blue indicates negative connection weights).

For numerical simulations, the feature spaces of all fields in the remapping subsystem are sampled with one unit per two degrees of visual angle, except for the gaze field that is sampled with one unit per degree. All fields use the same time constant $\tau = 10\,\text{ms}$ and steepness parameter $\beta = 4$ for the sigmoid output function.

field index	h	c^{exc}	σ^{exc}	c^{inh}	σ^{inh}	c^{gi}
r	-2	5	3°	7.5	6°	0
g	0	0	-	0	-	0.075
t	-2	7.5	3°/9°	25	6°/18°	0
b	-2	9	3°	15	6°	0
s	-2	-	0	-	0	0
a	-2	10	3°	0	-	0.075
d	0	8	3°	0	-	0.55

Table 4.1: Field parameters and parameters of lateral interactions. For the transformation field (with index t,), the first interaction width is the value for the dimensions of retinal space, the second value is for the dimensions of gaze direction.

projection index	c^{exc}	σ^{exc}	c^{inh}	σ^{inh}	c^{gi}
as	0.45	6°	0	-	0
ad	0.7	6°	0	-	0
da	1.125	3°	0	-	0
gd	7.5	3°	0	-	0.1
tg	9.5	9°	0	-	0
tr	1.2	3°	0	-	0
tb	1	3°	0.5	6°	0
bt	0.125	3°	0	-	0

Table 4.2: Parameters of interactions between fields.

4.4.3 Subsystem for gaze update

The gaze update subsystem provides a self-sustained representation of the current gaze direction during fixation phases, and updates this representation when a gaze change is initiated based on the saccade motor plan. All parameter values are again given in Tables 4.1 and 4.2. For numerical simulations, the fields of the gaze update subsystem are sampled with one unit per degree of visual angle.

The motor plan for a gaze update is provided in a format that matches neural population code representations found in the superior colliculus and

the frontal eye field. It is an activation distribution over two-dimensional retinal space (covering a range from -60° to 60° in both dimensions), with an activation peak indicating the intended saccade endpoint (relative to the current fixation point). This is equivalent to the saccade motor field in the biased competition model, although here, the actual saccade generation from the field dynamics is not modeled. Instead, the saccade signal is simply generated as an artificial input that is passively reflected in the field. This yields the field equation

$$\tau \dot{u}_s(x, y) = -u_s(x, y) + h_s + i_s(x, y). \tag{4.16}$$

The input i_s is a Gaussian of fixed strength centered on the planned saccade endpoint, starting 50 ms before the onset of the actual saccadic eye movement and lasting for 100 ms. This emulates in simplified form the corollary discharge signal as described by Sommer and Wurtz (2004), which has been found to be critical for the peri-saccadic remapping of retinocentric representations.

The central representation of the current gaze direction is provided by two one-dimensional gaze fields, u_d^{hor} for the horizontal and u_d^{ver} for the vertical component. Each of these is governed by a field equation of the form

$$\tau \dot{u}_d(x) = -u_d(x) + h_d + [k_{da} * F_d(u_a)](x) + [k_{dd} * f(u_d)](x). \tag{4.17}$$

The update fields, u_a, provide input to these gaze fields while a saccade is in progress, as detailed below. During fixations, there is no external input to the field. The lateral interaction kernel k_{gg} consists of local (Gaussian) self-excitation and global inhibition, producing competition between active regions in the field. Moreover, these interactions are strong enough to support a self-sustained activation peak that continues to reflect the current gaze direction in the absence of external input.

The actual gaze update mechanism determines the upcoming gaze direction, v_d^{new}, from the current gaze direction, v_d^{cur}, and a saccade motor command, v_s, as soon as the latter becomes available. With the saccade motor command given in a retinocentric reference frame, the mathematical operation to be implemented is the addition $v_d^{\text{new}} = v_d^{\text{cur}} + v_s$. This can be computed separately for the vertical and the horizontal component of the gaze direction, in a manner largely analogous to the one-dimensional reference frame transformation. Update fields u_a^{hor} and u_a^{ver}, respectively, are defined over a two-dimensional space spanned by current gaze direction and upcoming gaze direction, as depicted in Figure 4.7. These fields receive two inputs: The first one comes from the corresponding one-dimensional gaze-direction field, which projects a ridge input into the field that is localized along the dimension of current gaze direction (the horizontal axis in the figure) and homogeneous along the the dimension of upcoming gaze direction

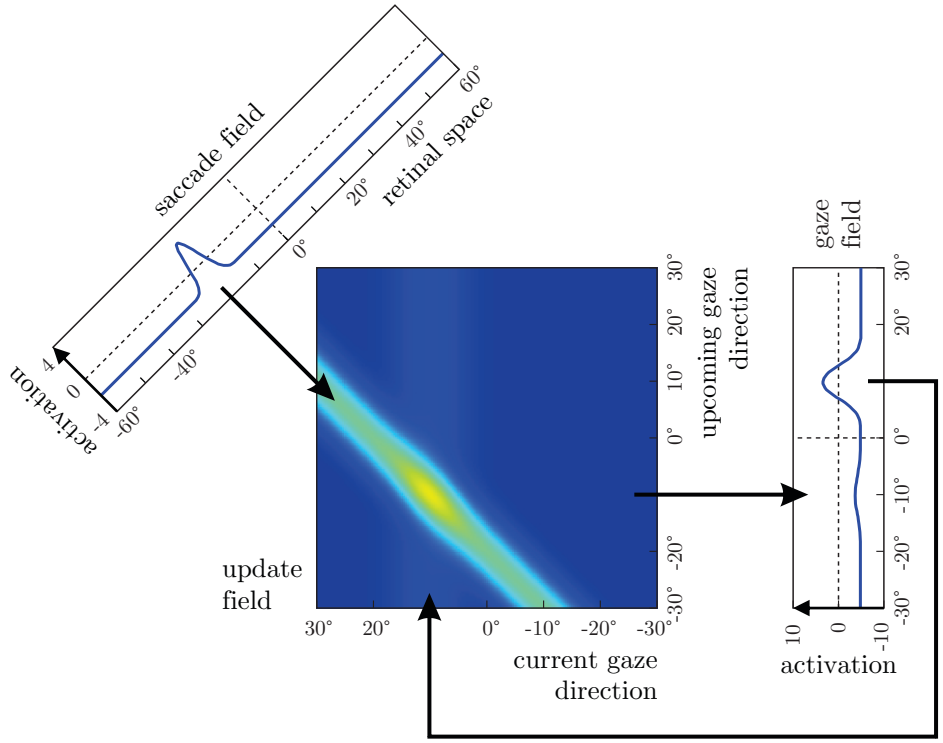


Figure 4.7: DNF architecture for a gaze update in one spatial dimension. Arrows indicate projections between fields. Note that the saccade field is actually two-dimensional, but only a cut through the activation along the relevant spatial dimension is shown here.

(the vertical axis). The second input comes from the saccade field. The horizontal and vertical components of the saccade signal are first separated by integrating the field output over the corresponding dimensions:

$$F^{\text{hor}}(u_s)(x) = \int f(u_s)(x,y)dy \tag{4.18}$$

$$F^{\text{ver}}(u_s)(y) = \int f(u_s)(x,y)dx \tag{4.19}$$

These one-dimensional saccade signals are then fed into the corresponding update field along the diagonal, after smoothing it with the Gaussian kernel.

The resulting field equation for the update fields then takes the form

$$\begin{aligned}\tau \dot{u}_a^{\text{dim}}(x,y) = &- u_a(x,y) + h_a + [k_{ad} * f(u_d^{\text{dim}})](x) \\ &+ [k_{as} * F^{\text{dim}}(u_s)](x+y) + [k_{aa} * f(u_a)](x,y)\end{aligned} \tag{4.20}$$

with $\text{dim} \in \{\text{hor}, \text{ver}\}$. The lateral interactions in the field consist of local excitation and global inhibition. Peaks in the field are induced at the inter-

section between the gaze input ridge and the diagonal input ridge from the saccade field when a saccade motor signal is present. These peaks remain stable due to the lateral interactions as long as the stronger saccade input is present, even when the relatively weak gaze input changes. They also prevent any other peaks from forming in the field during this time.

The vertical position of a peak in the update fields provides the new gaze direction as the result of the addition $v_d^{\text{new}} = v_d^{\text{cur}} + v_s$. To read out this new gaze direction, the output from each update field is integrated along the horizontal dimension and fed into the corresponding gaze field, yielding the term $[k_{da} * F_d(u_a)](x)$ in equation 4.17, with

$$F_d(u_a)(y) = \int f(u_a(x, y))dx. \tag{4.21}$$

Finally, the two one-dimensional gaze fields provide input to the single two-dimensional gaze field, linking the gaze update subsystem to the remapping subsystem. The input takes the form of one horizontal and one vertical activation ridge, with a peak induced at the intersection point. This is expressed in the term i_g in equation 4.11 as

$$i_g(x, y) = [k_{gd} * f(u_d^{\text{hor}})](x) + [k_{gd} * f(u_d^{\text{ver}})](y). \tag{4.22}$$

4.5 Results

4.5.1 Evaluation of the gaze update mechanism

The gaze update process is illustrated for the horizontal component of the gaze direction in Figure 4.8 (analogous processes occur simultaneously for the vertical component). Note that only a one-dimensional cut through the two-dimensional saccade field is shown here, reflecting the horizontal saccade metrics relevant for update of the horizontal gaze direction. During fixations, there is a self-sustained activation peak in the one-dimensional gaze field, reflecting the current gaze direction (Figure 4.8a). The gaze field projects a constant input into its associated update field, inducing a vertical ridge of activation, but this input alone is not sufficient to induce a peak. In the saccade field, there is no activation during fixation phases.

When a saccade motor command is generated, an activation peak appears in the saccade field that reflects the intended metrics of the upcoming saccade. The saccade field projects a diagonal ridge input into the update field. This input intersects with the present vertical ridge induced by the gaze field, and an activation peak forms at the intersection point (Figure 4.8b). The update field now projects back to the gaze field. Critically, this readout runs along the horizontal axis, not along the vertical axis as the reverse projection. So while the gaze field provides the old gaze direction to the update field, it receives the new gaze direction as input, provided by the

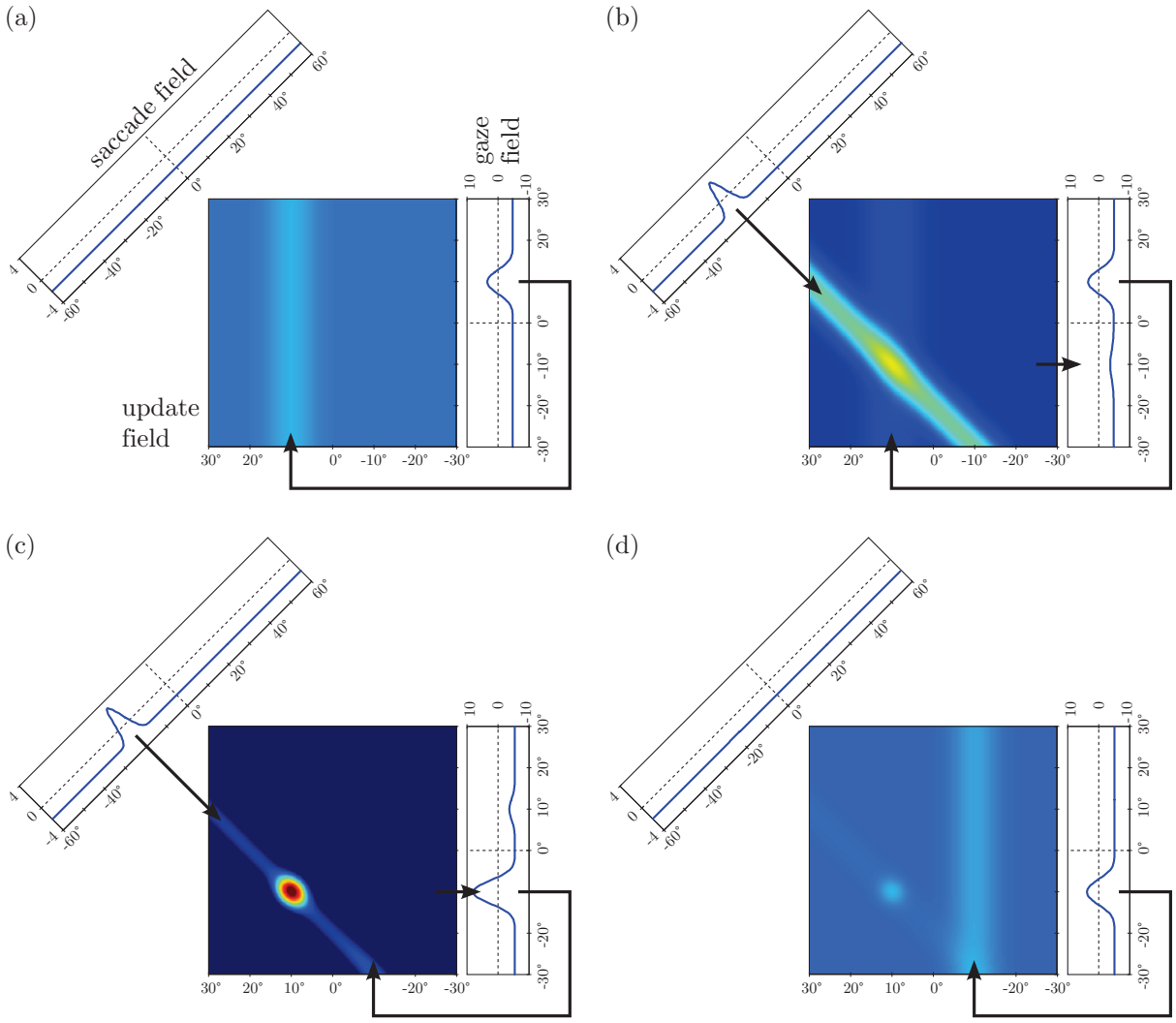


Figure 4.8: Development of activation patterns during a saccadic eye movement in the horizontal gaze update system. (a) Situation before a saccade motor signal arrives. (b) Activation peak forming in the update field after a saccade motor signal has appeared in the saccade field. (c) The representation in the gaze field has been updated, the peak in the update field remains stable as long as the saccade signal is present. (d) Situation briefly after the completion of the gaze update.

vertical position of the peak in the update field according to the relationship $v_d^{\text{new}} = v_d^{\text{cur}} + v_s$.

This input induces an activation peak in the gaze field at a new position. Due to the global inhibitory interactions in the gaze field, this peak competes with and ultimately suppresses the previous activation peak, which is not

supported by any external input (Figure 4.8c). Once the new peak has been established, it is again self-sustained by the lateral interactions. When the saccade motor signal ceases, the peak in the update field disappears, but the gaze peak remains active to reflect the new gaze direction until the next saccade (Figure 4.8d). As the peak locations change in both of the one-dimensional gaze fields (for horizontal and vertical gaze direction), they also provide a new input to the combined two-dimensional gaze field, and the activation pattern in that field is updated as well.

An important detail of this mechanism is how it ensures that exactly one complete update of the gaze direction field occurs for each saccade motor signal that arrives. Based on the coupling of the fields, one might expect that a whole cascade of gaze shifts may occur for every saccade signal: Once a new peak has formed in the gaze field, it projects a new input ridge to the update field (as in Figure 4.8c). A new ridge intersection occurs, and might once again provide a shifted input back to the gaze field. In fact, however, the new intersection point between the two ridges does not induce an activation peak. This is ensured by the lateral interactions and the input characteristics in the update field. The input from the gaze field is relatively weak, the input from the saccade field significantly stronger. While the input ridge from the gaze field determines where a peak first forms in the update field, it is then no longer needed to sustain the peak. Moreover, the competitive interactions in the update field mean that the activation peak, once it has formed, suppresses activation in the remaining field and prevents any new peaks from forming (this is visible in Figure 4.8c). So even though the gaze input changes, the peak in the update field remains at its original position as long as the saccade signal is present. Only once the saccade signal ceases does this peak disappear, which allows the formation of another peak in a new location when the next saccade occurs. This is an example of how the field interactions and the resulting stability properties of the field can be used to structure the continuous neural dynamics into discrete, controlled processing steps.

The saccade update mechanism was tested by simulating saccades of all possible amplitudes and directions within a range from $0°$ to $40°$ both horizontally and vertically, in steps of $1°$. The initial gaze direction for each trial was set to $(-20°, -20°)$, such that the final gaze direction was always well within the range of possible gaze directions covered by the gaze fields. Note that this covers all relevant saccade configurations within the core range covered by the architecture, since saccade direction (positive or negative) and initial gaze direction do not change the size of the resulting shift in peak position due to symmetry properties of the fields. The saccade command was generated by inducing an activation peak at the appropriate location and with a time course as specified above in the saccade field. The resulting final gaze direction was read out from the two-dimensional gaze field, determined

as center of mass of the field output. By the time the saccade signal ended, the gaze update was always completed. The mean error in gaze direction (deviation of the represented direction from the expected one) at this time was 0.08°, with a maximal error of 0.53°.

The main source of errors in the mechanism lies in interactions between old and new activation peaks in the one-dimensional gaze fields. For small to moderate saccade sizes, the activation peaks partly overlap. When the new peak is induced by input from the update field, its location can therefore be biased toward the location of the old peak. A second source of error at small saccade amplitudes is drift of the activation peak in the update field. This occurs when the new peak in the gaze field projects an input ridge into the update field that partly overlaps with the existing peak there. The peak then drifts along the diagonal activation ridge induced by the saccade field toward this new input, and consequently also projects a biased signal back to the gaze field. These two sources of errors act in different directions (the first leading to undershoot, the second to overshoot in the shift of the gaze direction peak), and partly cancel each other out.

These results demonstrate that the gaze update mechanism fulfills the key requirements to be used in an account of retinocentric remapping. It uses only signals that are available before the beginning of the actual gaze change—namely, the saccade motor plan that is modeled after the experimentally determined properties of a an efference copy from the superior colliculus—, and it performs a fast direct computation of the new gaze direction (rather than, e.g., a slow integration over instantaneous motor signals as used in the oculomotor system, see Goossens and Van Opstal, 2006). This makes it possible to predict the new gaze direction before the saccade is actually completed, and to use this prediction in the generation of pre-saccadic remapping as described below.

4.5.2 Emergence of retinocentric remapping in the model

A retinocentric remapping of activation peaks during gaze changes is an inherent property of the architecture presented here. I will describe the detailed sequence of instabilities that leads to this remapping. The process is illustrated in Figure 4.9 for the case of one-dimensional spatial representations, allowing an easier visualization and verbal description of activation patterns in the transformation field. The process for the full two-dimensional spatial representations is in all respects a direct extension of what is described here. The figure shows the architecture in the same format as Figure 4.1, but with the retinal readout added below the retinal field. This retinal readout is where the remapping can be observed.

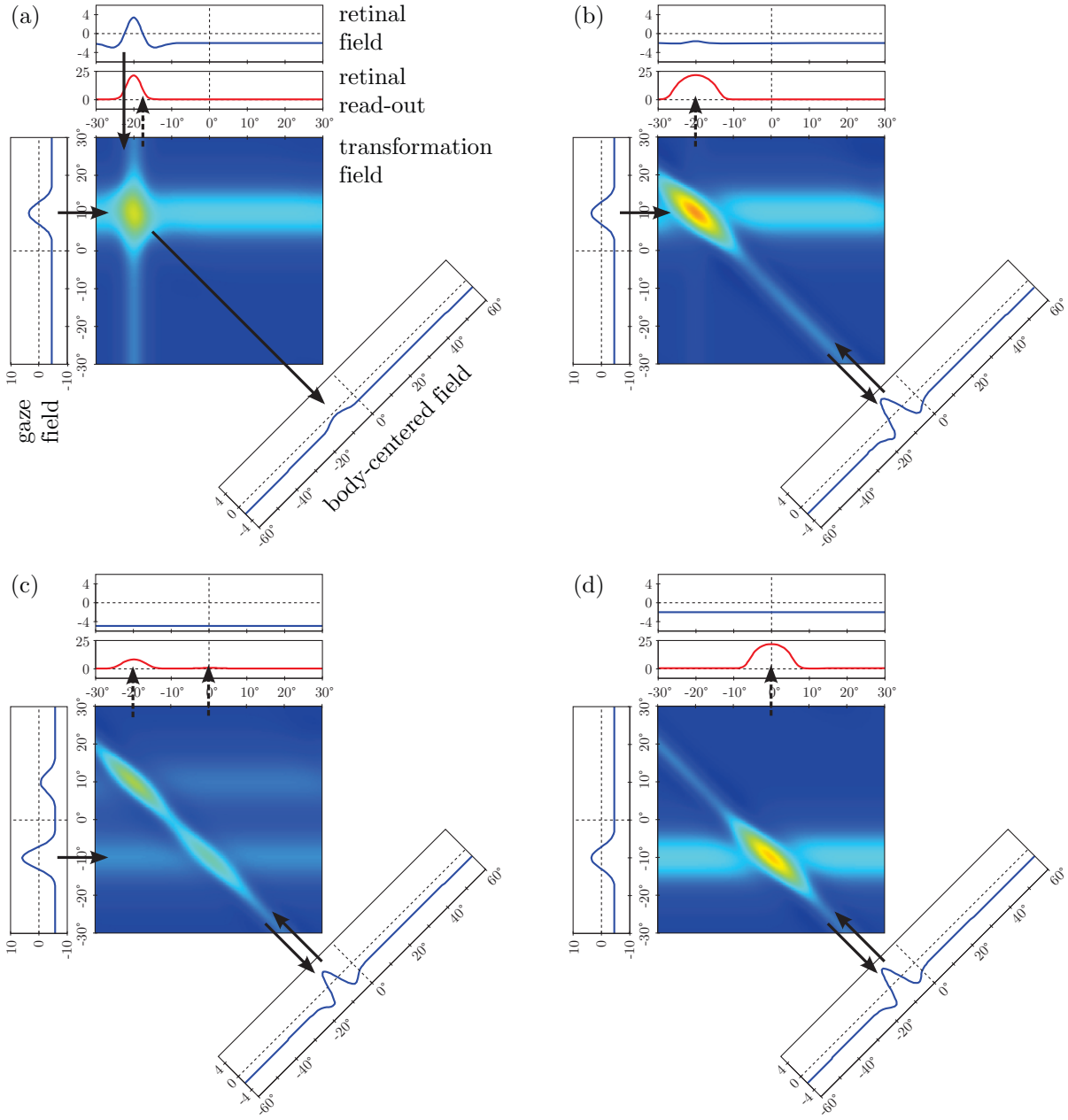
In the absence of any current or memorized visual stimuli, the only activation peak in the system is in the gaze field, which continuously maintains a representation of the current gaze direction, as described above. The gaze

field projects a broad input into the transformation field, visible as horizontal activation ridge in Figure 4.9a, but does not by itself induce an activation peak. When a salient visual stimulus appears, it first induces an activation peak in the retinal field. This field projects another input to the transformation field (weaker, but sharper than the gaze input; see vertical ridge in Figure 4.9a). An activation peak forms in the transformation field at the intersection of the two inputs, and projects further to the body-centered field, following the general mechanism for the forward transformation. Here, another activation peak forms to explicitly reflect the stimulus position in the body-centered reference frame.

This peak in the body-centered field now projects a diagonal ridge input back into the transformation field, which runs through the already present activation peak. When the stimulus is turned off, the coupled peaks in the transformation field and the body-centered field remain active (Figure 4.9b). Each of these peaks is stabilized by lateral interactions within the respective field, but these interactions in themselves would not be strong enough to sustain the peaks without input. With the mutual coupling between them along the diagonal axis, however, they remain stable. These coupled peaks now provide a working memory representation of the stimulus position in two reference frames: The body-centered field directly reflects the position in the body-centered frame. The horizontal position in the transformation field shows the retinal position of the stimulus, which can be made explicit in the retinal readout (Figure 4.9b); here, the active region matches the original retinal stimulus position.

A remapping of this retinal position now occurs when a gaze change is performed while a stimulus position is held in working memory, as shown in Figure 4.9c. The gaze change is reflected in the gaze field by the formation of a new activation peak that replaces the previous peak, driven by the gaze update mechanism. When the old peak in the gaze field disappears, so does the input it projects into the transformation field (the upper horizontal ridge in Figure 4.9). The peak in the transformation field that was located on this input ridge becomes unstable and decays. But at the same time, the new peak in the gaze field projects a new input to the transformation field (the lower horizontal ridge in Figure 4.9). This new input ridge again intersects with the diagonal activation ridge induced by the body-centered field. A new peak forms in the transformation field at this new intersection point, nearly simultaneously with the disappearance of the previous peak.

In effect, the activation peak in the transformation field shifts its position along the diagonal ridge (or more precisely, it jumps from one point on this ridge to another point). The peak in the body-centered field remains stable during this time—reflecting the fact that the remembered position does not change relative to the body-centered reference frame. Although the input that the body-centered field receives from the transformation field fluctuates


(a)
retinal field
retinal read-out
transformation field
gaze field
body-centered field
(b)
(c)
(d)

during the shift of the peak, the lateral interactions keep the body-centered peak sufficiently stable.

The retinocentric remapping of memorized locations is visible in the retinal read-out of the transformation field (red plot in Figure 4.9). As the activation peak in the transformation field shifts along the diagonal, the peak in the retinal read-out likewise changes its position. Due to the geometry of the projections between the fields, it is shifted exactly by the inverse of the gaze shift. Therefore, the new peak position in the retinal read-out now reflects the retinocentric location where the original stimulus would be visible if it had been sustained (or, where it reappears after the gaze shift if it actually was sustained). The system thus implements the desired function of updating a retinocentric representation of memorized locations to compensate for shifts in gaze direction.

The temporal pattern of the peak shift in the model is consistent with experimental data. In particular, the peak does not slide smoothly between the two locations, but disappears at the old location and almost simultaneously reappears at the new location. This matches the electrophysiological measurements of neural activity during saccadic remapping in the frontal eye field of the frontal cortex (Sommer and Wurtz, 2006). It is also consistent with behavioral data from Golomb et al. (2011) investigating how spatial attention to a fixed visual location is updated during a saccade. This study found that there is no attentional facilitation for intermediate retinal locations when the retinocentric locus of attention is shifted. Moreover, due to the predictive update of gaze direction in the present model, the remapping takes place before the saccadic eye movement is completed. This will be shown in greater detail below.

In order for this mechanism to fully account for experimental data and to be useful for an autonomous system of spatial memory and scene representation, it is important that the remapping can be applied to multiple items in parallel. This is indeed the case, as shown in Figure 4.10. During the forward transformation, multiple peaks in the retinal field may project parallel inputs into the transformation field (vertical ridges in Figure 4.10), which all form at their intersection with the single input from the gaze field.

Figure 4.9 *(preceding page)*: Evolution of activation patterns in the DNF model of peri-saccadic remapping during stimulus presentation and gaze change. (a) Forward transformation of retinal stimulus location briefly after stimulus onset. (b) Maintenance of a distributed working memory representation in the transformation field and body-centered field after the stimulus is turned off. (c) Shift of the gaze peak during a saccadic eye movement and resulting shift of the peak in the transformation field. (d) Prediction of new retinocentric stimulus position after the gaze change is completed.

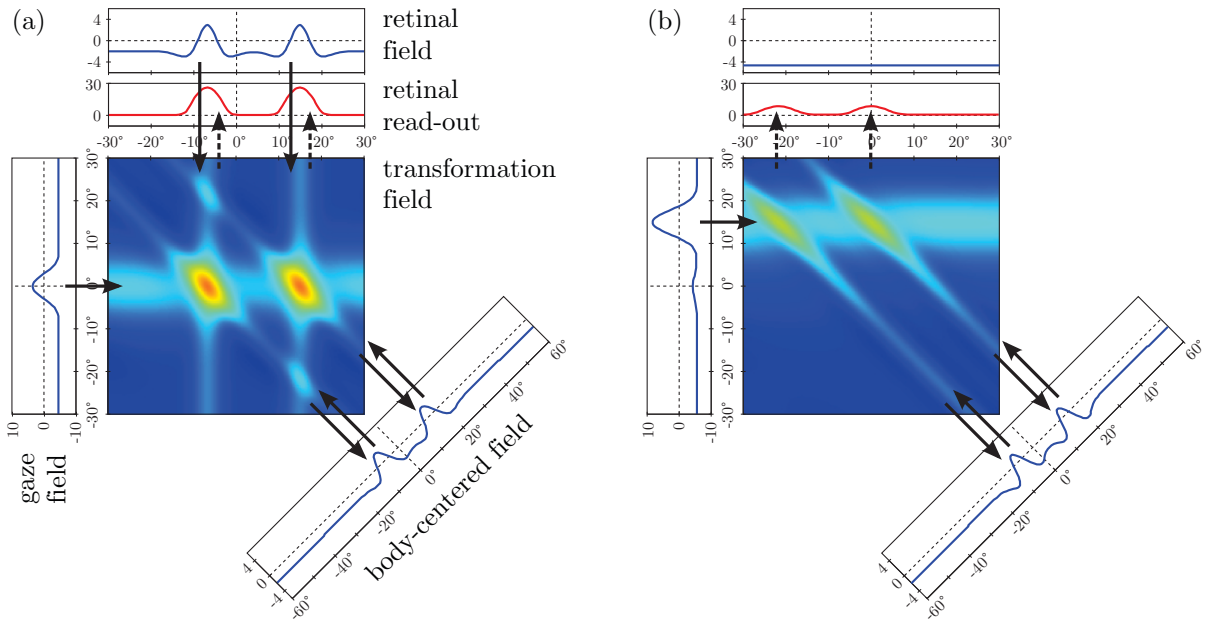


Figure 4.10: Simultaneous remapping of two stimulus locations in the DNF model. (a) Distributed representation of two stimulus locations during fixation. (b) Shift of activation peaks during gaze change and resulting predictions of new retinocentric stimulus positions.

These peaks then project in parallel to the body-centered field, where they produce one separate activation peak for each retinal stimulus. These in turn project back to the transformation field. Note that while the retinal stimuli are active, additional intersections occur in the transformation field between the inputs from one retinal peak and a non-matching body-centered peak (in Figure 4.10a, the intersections between vertical and diagonal input ridges above and below the horizontal gaze input ridge). These spurious intersections do not induce supra-threshold activation peaks, since the intersecting retinal and body-centered inputs are relatively weak compared to the gaze input. When a gaze change occurs, the peaks in the transformation field all shift in parallel along their corresponding diagonal ridges, driven by the changing input from the gaze field (Figure 4.10b). Consequently, the peaks in the retinal read-out shift to reflect the new retinocentric positions of all memorized stimuli.

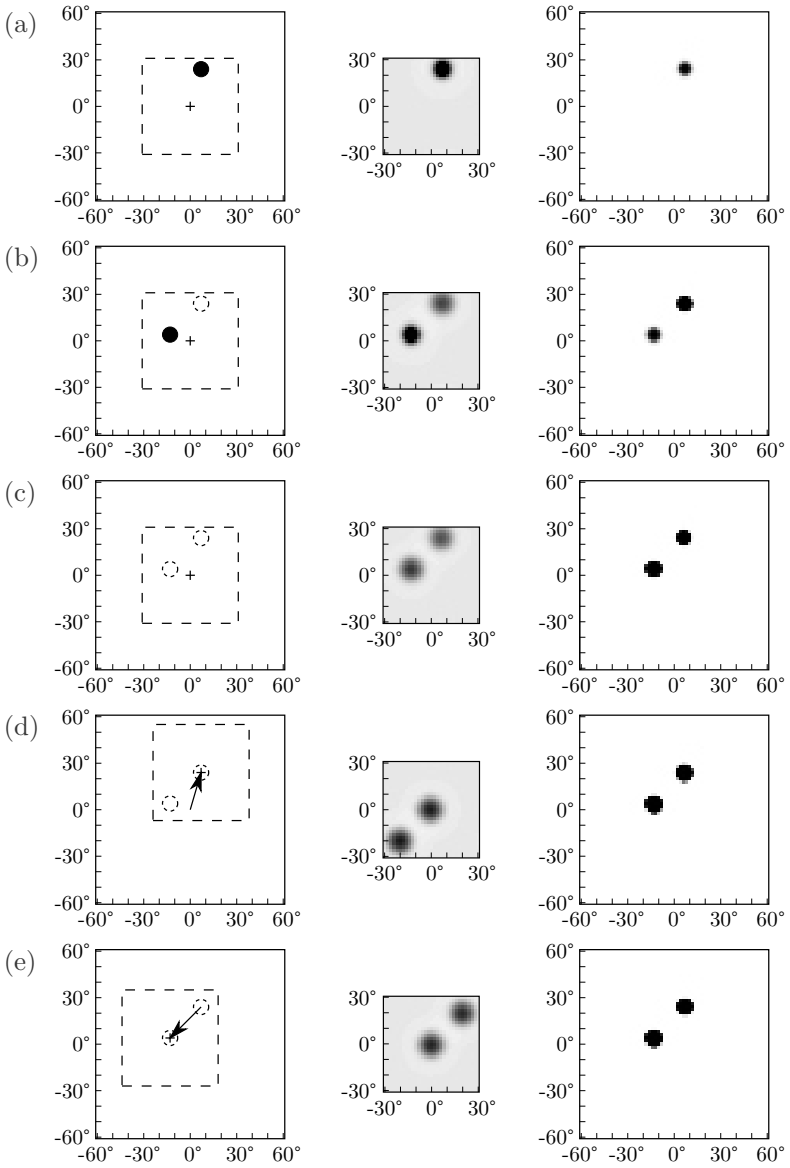
4.5.3 Evaluation of the remapping mechanism in double step saccades

One of the key reasons for having a continuously updated retinocentric representation of space in the brain is its use in saccade planning. Even if stimulus

locations are memorized in a gaze-invariant frame of reference, they must be transformed back into the retinocentric reference frame to be used as saccade targets, since the saccade motor command is always encoded retinocentrically. To test whether the model can provide the required functionality to plan saccades to memorized locations after an intervening gaze change and to evaluate its performance in this task, I employed a simulated double step saccade task (Hallett and Lightstone, 1976).

The task is illustrated in Figure 4.11 (left column), together with the activation profiles of the two-dimensional retinal read-out (middle column) and the two-dimensional body-centered field (right column). Two visual stimuli are presented to the model sequentially for 50 ms each, with stimulus onsets at t = 0 ms and t = 200 ms (Figure 4.11a and b). Activation peaks form to represent their locations both in the retinal and the body-centered reference frame, and remain sustained after the stimuli are turned off (Figure 4.11c). At t = 400 ms, a saccade signal is generated to fixate the first stimulus. The simulated saccade begins 50 ms later and takes another 50 ms to completion (Figure 4.11d). As an effect of this gaze change, the peaks in the retinal read-out are shifted to reflect the updated retinocentric stimulus locations, while the body-centered peaks remain unchanged. This is the same effect described in detail above for the simplified scenario with a single spatial dimension, now for two-dimensional spatial locations and gaze shifts. At t = 600 ms, the saccade signal to fixate the second stimulus is generated (Figure 4.11e), leading to another shift in the retinal read-out. Note that the remapping mechanism is needed to plan this second saccade. The saccade metrics cannot be determined based only on the retinal position of the visual stimulus, but must take into account the intervening gaze shift.

The required saccade metrics for both gaze shifts can be obtained from the retinal read-out of the transformation field, by determining the position of the corresponding peak at the time when the saccade signal is to be generated. For instance, the retinal location of the bottom left peak in Figure 4.11d (middle row) provides the required metrics for the saccade executed in Figure 4.11e. Note that for the task to be solved fully autonomously, the model would also have to accomplish the sequential selection of the two peaks in the correct order. This is not covered in the present architecture, although the continuous coupling of the two reference frames can potentially be utilized to accomplish this. If the temporal order of the two stimuli is memorized in the body-centered reference frame, it may send an additional biasing input into the body-centered field at the time when a saccade to a specific stimulus should be initiated. This would strengthen the corresponding memory peak, and the stronger activation would immediately be propagated to the retinal representation, where it could be used to select the correct peak and determine the saccade metrics from its position. In the experiment, the correct peak is selected manually instead, as the one that


(a)
(b)
(c)
(d)
(e)
60°
30°
0°
-30°
-60°

is closer to the expected location. In all trials, the mechanism provided a close approximation of the exact remapping of all memorized locations, so that this selection was never ambiguous.

The accuracy of the remapping was tested by performing the double step saccade task with different positions of the first stimulus, and consequently different metrics of the first saccade. Stimulus positions were varied between 0° and 25° both horizontally and vertically in steps of 1°. The second visual stimulus was always located 20° below and to the left of the first stimulus. This stimulus distance avoided effects of repulsion or attraction between activation peaks (Simmering et al., 2008), which are not the subject of this study. The location of the peak corresponding to the second stimulus was then determined in the retinal read-out at time t = 600 ms (when the motor signal for the second saccade was generated) to evaluate the accuracy with which the system could provide the metrics for the second saccade.

The mean amplitude of the error (deviation from the exact remapped location) was 0.29°. The maximal error over all trials was 0.85°, with a standard deviation from the exact remapped location of 0.35°. Errors in the remapping occur when the peak in the transformation field does not shift exactly along the diagonal input ridge from the body-centered field. This can occur due to partial overlap and lateral interactions between the decaying old activation peak and the newly forming activation peak in the transformation field during the remapping. The error typically manifests in a slight under-compensation of the gaze shift and occurs mainly for saccades of intermediate amplitude. Note that the accuracy measures reported here for the remapping also incorporate the errors in the gaze update described earlier, since the remapping is based on the internally generated estimate of the new gaze direction after a saccade.

Figure 4.11 *(preceding page)*: Stimuli and activation patterns in a double step saccade task. The left column depicts the simulated stimulus positions on a screen (black dots for active stimuli, dashed circles for positions of previously presented stimuli). The black cross indicates the current fixation point, the dashed square shows the current field of view, arrows indicate gaze changes. The middle row shows the activation patterns in the retinal read-out of the transformation field, the right column shows the output of the body-centered field. (a) First stimulus presentation. (b) Second stimulus presentation. (c) Delay period. (d) Situation after the saccade to the first stimulus. (e) Situation after the saccade to the second stimulus.

4.5.4 Time course of remapping and comparison to electrophysiological data

As a final result, I will show comparisons between the activation time course in the DNF model and electrophysiological data of retinocentric remapping from the work of Duhamel et al. (1992). These experimental results have been viewed as key support for the hypothesis that trans-saccadic spatial memory relies on remapping of retinocentric representations and therefore does not require any gaze-invariant neural representations of space (Wurtz, 2008). By modeling the experiment with the DNF model, I will show that the findings are entirely consistent with the use of a gaze-invariant representation (here, the body-centered field) that is continuously coupled to representations in a retinocentric reference frame.

In the original experiment, activity was measured for single neurons in the LIP region of macaque monkeys. First, the visual receptive field of each neuron in a retinocentric reference frame was determined. Then neural activity was measured while visual stimuli were either presented statically within the receptive field, or were moved into or out of the receptive field by saccadic eye movements. Here, I compare the activity of these neurons to the activation time course in the retinal read-out of the transformation field, sampled at individual locations that correspond to the receptive field centers of the neurons in the experimental study. This comparison is based on the hypothesis that the experimentally observed neurons are either gain-modulated themselves (which was not tested in the experiment), and thus correspond to nodes at certain retinocentric locations in the transformation field; or that they are driven by retinocentric input from the transformation field.

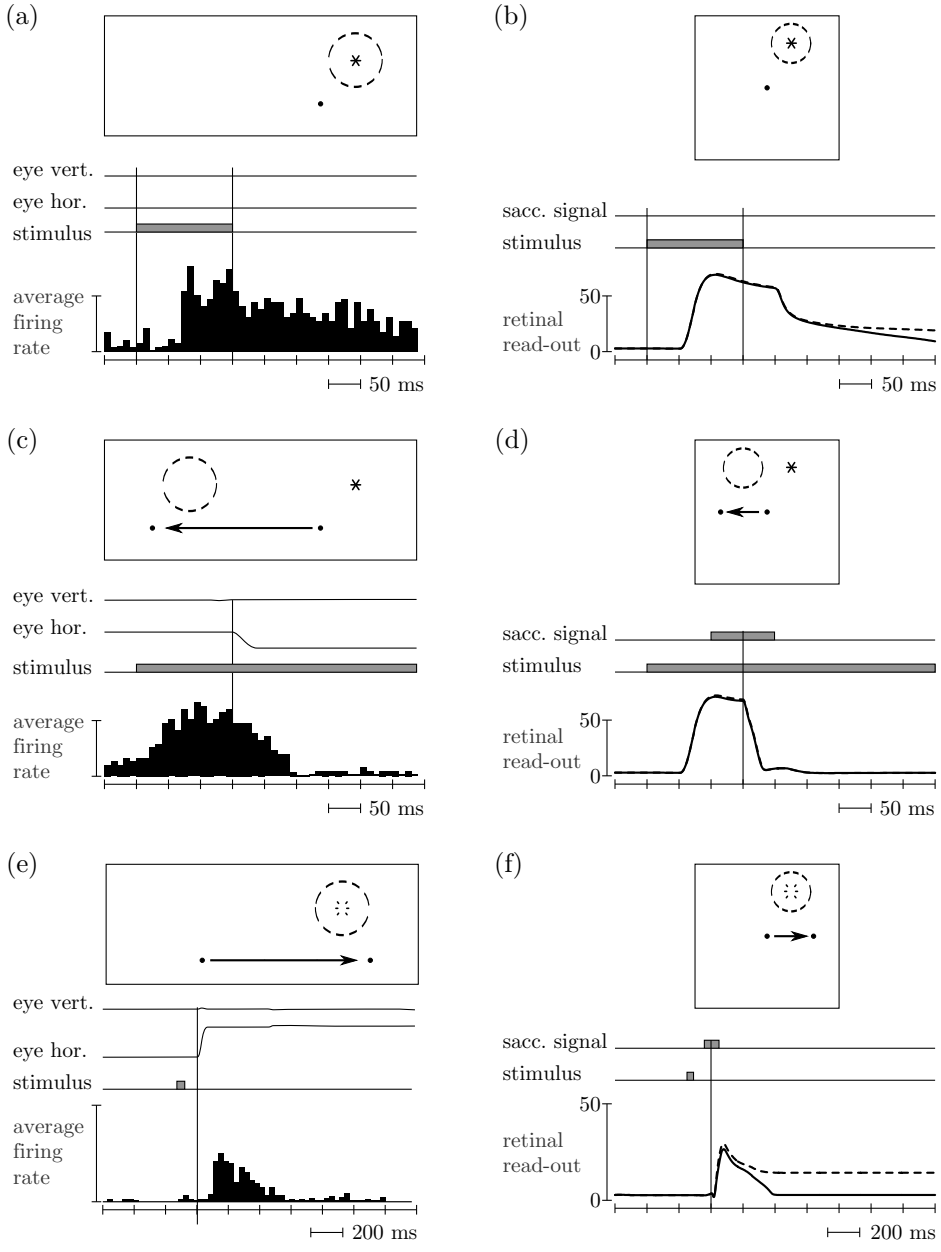
There is one adjustment made in the model to accommodate for this task. The monkeys in the experiment were not required or trained to memorize any spatial locations, and no sustained activity was observed in the neural recordings after stimuli were turned off. To emulate this in the model, the resting level in both the transformation field and the body-centered field is slightly reduced so that the coupled activation peaks that form in these fields are no longer self-sustained. This can be achieved in a neural system by a global control input without requiring any structural changes, and is analogous to the control of the feature working memory field in the DNF model of biased competition described in the previous chapter. I therefore consider this as the same model operated in a different mode to reflect task instructions, and will refer to it as the *perceptual mode* in contrast to the *memory mode* that was used so far. For completeness, I describe activation time courses for both modes of operation below.

The neural recordings from the experiment are shown side-by-side with the activation time courses from the model in Figure 4.12. In the first experimental condition (Figure 4.12a), a visual stimulus is briefly presented

inside the neuron's receptive field. The neuron increases its activity with a brief delay after the stimulus onset (due to synaptic transmission time from the retina), then activity slowly decays after the stimulus is turned off. The model responds similarly when operated in the perceptual mode (Figure 4.12b), with fast increase of activation followed by a slow decline when the stimulus is turned off. The latter effect is due to the slow decay of stabilized activation peaks even in the perceptual mode. In the memory mode, the activation decays to a lower level when the stimulus is turned off, but then remains stable at that level. Note that a fixed delay between modeled stimulus onset and activation of the field input is employed in the model to emulate the synaptic transmission time.

In the second condition (Figure 4.12c), a sustained stimulus is presented inside the neuron's receptive field, but then a saccade is made to a peripheral visual target such that the neuron's receptive field is moved away from the stimulus. As a result, the neural activity decays much more rapidly, even though the effect on the visual input to the neuron should be indistinguishable. The model reproduces this effect (Figure 4.12d). Both in the perceptual and in memory mode, the activation quickly goes back to zero at the time of the simulated saccade. The reason is that during the saccade there is not only a suppression of the visual input to the retinal field, but there is also the shift of the gaze input. Losing both of its supporting inputs, the original peak in the transformation field decays almost immediately (while a new peak forms at the remapped location, see Figure 4.9). Consequently, the activation at the original retinocentric stimulus location in the retinal read-out quickly declines.

In the third and final condition, the saccade moves the stimulus location into the receptive field of the neuron that is being recorded from. However, the stimulus is only flashed briefly (for 50 ms), and it is extinguished before the actual gaze shift begins (Figure 4.12e). Here, the key signature of retinocentric remapping is observed: Briefly after the saccade, there is a transient stimulus-related response of the neuron, even though there was never any direct visual stimulation at the retinocentric position of the neuron's receptive field. The same effect can be observed in the retinal read-out of the DNF model, due to the remapping mechanism detailed above (Figure 4.12f). When the visual stimulus is presented before the saccade, it induces coupled activation peaks in the transformation field and the body-centered field. Then, during the gaze shift, the peak in the transformation field is shifted to reflect the new retinocentric location of the stimulus. In the perceptual mode of the model, the activation peaks in the transformation field and body-centered field decay over time, but the activation is sustained long enough to produce the transient response in the retinal read-out. When the model is operated in the memory mode, the activation at the new location is sustained, producing the updated spatial working memory in a


(a)
eye vert.
eye hor.
stimulus
average
firing
rate
50 ms
(b)
sacc. signal
stimulus
50
retinal
read-out
0
50 ms
(c)
eye vert.
eye hor.
stimulus
average
firing
rate
50 ms
(d)
sacc. signal
stimulus
50
retinal
read-out
0
50 ms
(e)
eye vert.
eye hor.
stimulus
average
firing
rate
200 ms
(f)
sacc. signal
stimulus
50
retinal
read-out
0
200 ms

retinocentric reference frame that was used in the double-step saccade task.

Note that in the response of the single neuron shown in Figure 4.12e, the remapping activity occurs only with a brief delay after the saccadic eye movement, and thus later than in the simulation. Sommer and Wurtz (2006) have investigated the exact timing of neural responses associated with retinocentric remapping in a separate study. They found significant variability between individual neurons, but report that the mean response onset is very close to the time of saccade initiation. Thus, the remapping often takes place before the eye movement is completed. These findings are consistent with the activation time course in the DNF model, which can explain the early remapping activity through the integrated mechanism for the predictive update of gaze direction.

As a final remark on these simulations, I would like to point out that all visual stimuli relevant in the different conditions were modeled, not just the probe stimulus in the receptive field of the recorded neuron. In the third condition, for instance, the shifting fixation stimulus that provides the saccade target for the monkey is included in the simulation. These stimuli are processed in the same way as the probe stimulus, and are likewise subject to remapping during gaze shifts. This is consistent with experimental results showing that the saccade target itself is also remapped during the gaze shift, and its new location is predicted with good accuracy even when it is not foveated by the saccade (Collins et al., 2009). This feature highlights the fact that the remapping in the present model is a general property of the coupled spatial representations, whereas in other theoretical accounts it is often interpreted as special operation applied only to intended future saccade targets (Xing and Andersen, 2000).

Figure 4.12 *(preceding page)*: Time course of neural activity in LIP and activation in the DNF model during stimulus presentation and saccadic gaze shifts. (a) Response of a neuron to a stimulus (star in the stimulus display at the top) in the neuron's receptive field (dashed circle). (b) Time course of the retinal read-out in the DNF model in the same condition, measured at the retinal position of the a visual stimulus. The solid line shows the retinal read-out in the perceptual mode of the model, the dashed line in the memory mode. (c) Neural response when the stimulus is moved out of the neuron's receptive field by a saccade (indicated by the arrow in the stimulus display). (d) Model activation in the same condition. (e) Remapping activity in an LIP neuron when the position of a previously extinguished stimulus (dashed star) is brought into the neuron's receptive field by a saccade. (f) Model activation in the same condition.

4.6 Discussion

In order to understand the neural mechanisms underlying spatial cognition, it is essential to consider how the brain deals with the different reference frames in its spatial representations. Mappings between different reference frames are needed to enable trans-saccadic working memory in active looking, to link sensory representations to motor acts, and to integrate spatial representations from different sensory modalities. While transformations between different reference frames are a general issue also for algorithmic approaches to sensory processing and motor planning, for instance in robotics, specific difficulties arise in neural approaches to this problem. Here, spatial information is generally represented in a distributed fashion through population codes, and operations on these representations must be implemented through fixed synaptic connections. Reference frame transformations must therefore be realized in a form that is fundamentally different from the algorithmic operations that can be applied to numerical vectors.

The mechanism proposed here for neural reference frame transformations is based on the observation of gain-modulated neurons in the parietal cortex and other brain areas (Andersen et al., 1985). The general idea is that a combined representation is formed by bringing together spatial information in one reference frame and a representation of the required shift between reference frames (in the case of a transformation from the retinocentric to the body-centered frame, this is given by the current gaze direction). From the activation pattern in this combined representation, the spatial information in the new reference frame can then be read out through fixed synaptic projections. This general mechanism has previously been implemented in different neural network models (Zipser and Andersen, 1988; Pouget and Sejnowski, 1997). Here, I have presented an implementation in the framework of Dynamic Field Theory, first as a general mechanism, then in the form of a concrete model for trans-saccadic working memory and retinocentric remapping.

The DNF model focuses on the autonomous operation of the transformation mechanism, such that new visual stimuli can be processed at any time to form a representation of their locations in two different reference frames simultaneously. This is aided by the fact that the DNF model can perform the reference frame transformation for multiple locations in parallel, a feature not present in previous models. Early neural network models did not generalize to multiple stimulus locations (Zipser and Andersen, 1988), and the radial basis function network by Denève et al. (2001) employed divisive normalization that forces the representation to converge on a single location. The model of Xing and Andersen (2000) did deal with two visual stimuli (the current and the subsequent saccade target), but held these in separate representations, with the full transformation process only applied to

one of them. With the parallel processing and autonomous operation in the present model, the reference frame transformation is no longer interpreted as a discrete operation applied to an individual spatial location. Instead, it is understood as a continuous process that provides a dynamical coupling between two spatial representations in different reference frames and keeps them synchronized.

It is important to note, however, that this system is not intended to apply the reference frame shift to a complete visual image. Instead, the transformation is only applied to a limited number of discrete locations that are represented by stabilized activation peaks in the fields. This is consistent with experimental findings showing that humans do not integrate even moderately complex shape information across saccadic eye movements (Irwin et al., 1983). Humans can, however, integrate information about individual locations across saccades, and for instance make accurate and precise judgments about their spatial arrangements without having seen them in the same fixation (Hayhoe et al., 1991). In limiting the spatial shift to a few spatial locations, the present model differs from approaches such as dynamic routing circuits (Olshausen et al., 1993), which propose that a variable spatial shift and scaling operation is integrated directly into the visual processing stream. In this approach, the shift is also applied to representations of surface feature information (such as edge orientations or texture), so that it can serve for position-invariant object recognition. Unlike the approach of Olshausen et al., the present model does not rely on special scalable synapses to perform the shift operation.

The goal of the DNF model also differs somewhat from the radial-basis function models of Pouget and colleagues (Pouget and Sejnowski, 1997; Denève et al., 2001), which show strong similarities in the underlying mechanisms (despite several differences in implementation pointed out earlier). These models aim more strongly to achieve optimality in sensor fusion over different reference frames. They tend toward an interpretation of the activation pattern as a probability distribution (which is also expressed in the explicit normalization in each representation), and are thereby linked to graphical models such as Bayesian networks. This link is made explicit in Denève and Pouget (2004). The DNF model does not aim to explicitly represent probability distributions. While DNFs can under certain conditions act in a way that approximates Bayesian integration of different information sources, the activation peaks in the present architectures are interpreted as discrete stimulus locations. This discretization occurs through the detection instability in the DNFs, which transforms a graded input signal into a binary decision to either form a peak or not (although a certain amount of graded information can still be represented by the size of the activation peak).

With the DNF model, I have shown how the phenomenon of retinocentric remapping that has been observed in neural data can be explained

as an emergent effect of bidirectional coupling between retinocentric and gaze-invariant spatial representations. Retinocentric remapping has been reported in the same cortical areas in which gain-modulated neurons are found, but has previously been explained by mechanisms distinct from reference frame transformations. Models of retinocentric remapping proposed a direct shift of retinocentric representations based on the metrics of a saccade (Quaia et al., 1998; Keith et al., 2010), and such explanations have been taken as arguments against a role of gaze-invariant representations in trans-saccadic spatial memory (Colby and Goldberg, 1999; Wurtz, 2008). The DNF model can directly account for the original experimental data (Duhamel et al., 1992), due to the parallel remapping of multiple items and the integrated mechanism for a predictive gaze update.

This gaze update module is another important novelty of the present model. It allows the remapping to occur simultaneously with the saccadic eye movement, and explains the experimental observation that the remapping activity is dependent on an efference copy of the saccade motor signal. The gaze update module uses an architecture analogous the reference frame transformation, but is operated in a different dynamic regime. The strong global interactions employed in this system create a sequence of instabilities and stabilized states that deviates from the one in the transformation module, and that produces the single, stabilized update of the gaze direction once a saccade motor signal is provided. The coupling between the gaze update system and the remapping system demonstrates how the field dynamics simplify the modularization of a complex architecture. Due to the effects of lateral interactions, the activation states of the gaze fields always take the form of a single activation peak of a relatively stereotyped shape. The input to the remapping system is therefore qualitatively always the same, and the details of the processing within the gaze update system do not affect the remapping system.

The broader functional achievement of the DNF model is that it can autonomously keep different spatial representations aligned and synchronized. The general mechanism allows contributions from different sensory and cognitive systems to be integrated and to interact in a single, distributed representation of space, without imposing any constraints on the number or timing of such contributions. In particular, the system allows a selection decision to be propagated between reference frames, due to the continuous coupling of graded representations. If, for instance, the location of an individual stimulus is selected by attentional mechanisms at the retinal level (e.g., because the stimulus is particularly salient), then the increased activation for this location is automatically projected to the body-centered reference frame. The selection decision is thereby propagated to the working memory representation. This is critical if spatial location is to be used as a form of pointer that allows referencing a specific object (compare Cavanagh et al., 2010). To be

functional, such a pointer must be able to traverse the different levels of a neural representation, from the sensory level to working memory and more abstract cognitive representations. In this role, the transformation operation will be used in the following chapter, where working memory for spatial locations is combined with surface feature representations to form a scene representation in memory.

Chapter 5

Neurodynamic Model of Human Scene Representation

5.1 Introduction

In this chapter, I will address the question how humans form an internal representation of a visual scene that can be used to guide behavior and plan actions. The two previous chapters have highlighted the importance of active looking with frequent gaze changes to perceive different parts of the environment, have introduced the problems that these gaze changes create for spatial cognition and memory, and proposed a solution to keep track of locations in the world despite shifts of the retinal reference frame. I will now build on the two models described so far and present a DNF architecture that can not only keep track of object locations in a visual scene, but that integrates the spatial information with information of object features at these locations. The result is a gaze-independent scene representation in working memory.

When a scene representation is formed by fixating objects individually, this necessarily requires a sequential processing of the visual scene. However, humans appear to employ such sequential processing even when the gaze remains fixed during scene viewing, for instance when memorizing arrays of simple visual stimuli presented around the fixation point. Observations from psychophysical experiments indicate that participants focus attention on each item in the array sequentially, even though their gaze does not change. It has been suggested that this focused attention is necessary to bind the individual features of an object together into a coherent representation. The DNF model that I present here adopts this strategy of sequential processing of visual items. It provides a mechanistic explanation for the problem of feature binding, and explains how focused attention can solve this problem.

In the following sections, I will first review experimental findings on

human working memory for visual scenes, and discuss theoretical accounts for these findings. I will then motivate and describe a DNF architecture that implements a neurodynamic process model for the formation and use of scene representations in working memory. The model captures two elementary surface feature dimensions and one spatial dimension for object location, and autonomously iterates over objects to memorize visual scenes. I will present tests of this model in different change detection tasks, which constitute a key experimental tool to assess visual working memory in humans, and I will show that the model mechanisms are consistent with human behavior. The model and the results have previously been described in Schneegans et al. (in press b).

The psychophysical model I describe here has been developed in parallel with a robotic model (Zibner et al., 2010a,b, 2011a). In the robotic scenario, the robot builds a representation of objects in a table-top scene based on a camera image, and can then use this representation to guide arm movements or answer questions about the scene. The robotic version of the model uses two spatial dimensions to cover actual visual space, and has also been combined with a full object recognition system to go beyond elementary surface features for identifying objects (Zibner et al., 2011b). On the other hand, it is somewhat less neurally realistic in that it uses algorithmic shortcuts for certain operations, and it lacks certain structural elements to deal with the change detection tasks discussed here.

5.1.1 Scene representation and working memory in humans

Working memory in humans and animals is a type of memory representation that forms rapidly, but has limited capacity (Vogel et al., 2006). The underlying neural mechanism is generally believed to be the sustained activity of neurons that excite themselves through synaptic loops (Wang, 2001), as captured in the sustained activation peaks in DNF models. The specific type of memory addressed here has been termed the *visuo-spatial sketchpad* in the influential classification of Baddeley and Hitch (1974). It captures information about objects and their locations in the world. It is distinguished from the articulatory loop, which is used to store verbal information, and a central executive working memory component. All these forms of working memory can again be contrasted with two other forms of memory. Long-term memory is a high-capacity memory mechanism that is based on the formation or modification of synaptic connections between neurons. In contrast, iconic memory is often likened to a neural afterimage, that provides a detailed memory representation for brief periods of time (less than one second), but is very susceptible to interference from new sensory input. These two forms of memory are not addressed in the present chapter.

How does the human nervous system form an internal representation of a visual scene in working memory? What properties do these internal rep-

resentations have, and how are they used in the generation of behavior? A large number of behavioral studies have explored different aspects of these questions. Change detection experiments constitute one central experimental tool in exploring human scene representation in working memory. If a participant can reliably detect a difference between a previously viewed scene and a changed version of it, this demonstrates that the changed aspects were held in working memory. The opposite is not necessarily true: Not detecting a change does not prove that something was not memorized, since failures may also occur in the comparison between working memory and new stimuli.

One quite radical theory posits that no internal scene representations are built up in working memory or used to guide behavior when viewing natural scenes. This view is based on findings of so-called *change blindness* (Rensink et al., 1997; Simons and Levin, 1997). In one type of experiment, participants are asked to view a photograph of a natural scene on a computer screen and to report any changes that occur in the image while they view it. Participants frequently fail to notice even drastic alterations in the image, as long as the change is masked in such a way that it cannot be detected as motion. This can be achieved by altering the image either while the participant is blinking or performing saccadic eye movements (O'Regan et al., 2000), during which visual perception is greatly diminished, or while a masking stimulus is flashed over the image (Rensink et al., 1997). Analysis of participants' performance indicated that they only detect the change reliably if they are currently fixating the changing region or object, which is taken as an indication that only the currently attended item is actually kept in an internal representation.

This interpretation of change blindness results has been questioned by the work of Hollingworth and colleagues. Hollingworth and Henderson (2002) tracked participants' eye movements during a similar task, and found that salient changes were detected at least in about half of the trials as long as participants fixated the affected object both before and after the change. In another variant, the previously simultaneous tasks of memorization and change detection were separated. A forced choice test for a specific object was presented at some point during viewing the scene, and participants had to decide whether that object was changed or unchanged. Participants showed moderately high performance (>80%) when they had previously fixated the tested object in the scene, although performance decreased with increasing number of intervening fixations.

Other studies looked specifically at the capacity of working memory for scene representations. A classical experimental paradigm for this is change detection with arrays of artificial stimuli (Vogel et al., 2001; Treisman and Zhang, 2006). The stimuli are characterized by one or more elementary surface features, such as color, shape, or edge orientation. In each trial, first a sample array is shown, then after a delay a second stimulus array is

presented as test array. The two arrays may either be identical, or different types of change may be introduced, and subjects have to report whether they detected a difference (and possibly provide additional information, such as indicating the changed item).

One key parameter in these tasks is the size of the stimulus arrays, that is, the number of presented items. Performance decreases significantly for larger arrays, and the performance statistics over different array sizes are consistent with the notion of a limited working memory capacity. Humans appear to be able to memorize between three and five different items, dependent on task details (with a few studies indicating a capacity limit of up to seven items). When more items are presented in the stimulus array, only a subset is memorized (Irwin, 1992; Irwin and Andrews, 1996). Interestingly, this capacity limit does not seem to operate on the level of individual features to be memorized, but on the level of items or objects. A study by Luck and Vogel (1997) found that participants could remember the same number of items independent of whether they were characterized by a single feature (e.g., color) or a combination of features (e.g., color, edge orientation, and size). This is consistent with the idea that different visual features may be represented and memorized in separate neural populations.

A second type of manipulation in the change detection task is the type of the change that can be introduced in the test array. In particular, results indicate that it makes a significant difference whether a novel feature value is introduced (such as a color that was not present in the sample array), or whether only the binding between features is changed (such as switching the colors between two items from the sample array). The latter type of tasks is generally found to be harder for humans. Moreover, several studies indicate that memorizing the binding between different features of an object requires focused attention on that object, either through fixation or covert spatial attention (Treisman and Zhang, 2006; Hyun et al., 2009a). The special role of feature binding in perception and memory is formulated in the Feature Integration Theory (Treisman and Gelade, 1980; Kahneman et al., 1992), described below. These findings are generally consistent with the results for natural scenes, where sequential fixation of visual objects in a scene was found to be a critical factor for successful change detection.

Taken together, these results clearly show that humans are capable of forming a representation of a visual scene in working memory, but they also show the significant limitations in this ability. Clearly, the scene representation in working memory is not a photographic image that is acquired instantly when we first see a visual scene. Instead, only certain aspects of a scene are retained in memory, such as the features of individual objects, and also for those the capacity of working memory is very limited. Forming the scene representation is an active process that takes time and uses neural resources. In particular, it requires visual attention to be sequentially

directed at different objects in a scene. The same is true for using such a scene representation once it is formed, for instance for detecting changes in the visual image. Explaining the experimental findings on human scene representation abilities requires addressing not only the properties of working memory representations themselves, but also the processes involved in the formation and utilization of these working memory representations.

5.1.2 Theoretical accounts of human scene representation

An eminent theory in psychology of human scene perception is the Feature Integration Theory, proposed by Treisman and Gelade (1980; see also Treisman, 1988, 1996). Based on convergent evidence collected from visual search tasks, texture segmentation, illusory conjunctions, and change detection, this theory states that individual features (like color, edge orientation, and spatial frequency) can be detected and processed in parallel over a whole visual scene. Forming conjunctions between these features dimensions, however, requires focused attention.

For instance, visual search for an object of a certain color in an array of distractor objects is a parallel process, in which the search time does not increase with increasing number of distractors (Treisman, 1988). In contrast, search for a conjunction of features (like search for a red vertical line in an array of red horizontal lines and green vertical lines) is often a serial process. The response time increases approximately linearly with the number of distractors, indicating that the items in the search array (or at least a subset of them) are inspected one by one to determine whether they match all target features. Attention to individual items in such tasks may be either overt (making a saccade to fixate an object) or covert. In the latter case, attention is focused on a single object (or location) without producing any eye movements; this is the more common strategy in humans to rapidly inspect simple and closely spaced stimuli in typical visual search tasks. Both forms of attention appear to work equally to bind features together.

Based on this theory of scene perception, Kahneman et al. (1992) developed a theory for how the objects from a scene are stored in memory. They proposed so-called *object files*, which hold the different features of an object in an integrated fashion. According to this theory, an object file is created for a visual object when it is first focused by spatial attention. Features may then be added to that object file (although not necessarily all features are added immediately, but only task-relevant ones). The object file can be addressed by the location of the object and accessed at a later point in time (e.g., when the same object is attended again), and features can be added when they become available, or updated when they change.

Through these properties, object files mediate continuity of object representations. For instance, if you see an object moving in the sky, you may first only perceive its approximate shape and assume that it is a bird; but

after briefly observing it, you realize that it is actually a plane. Still, you are certain throughout this identification that there is only a single object, and only your assumption about its identity changed. In such a case, a single object file is created and sustained, and its content is updated. Conversely, multiple distinct object files with equal content can accomplish individuation of objects, when several indistinguishable objects are present in a visual scene.

It is assumed in the theory that movement of objects is tracked in the object file (so that addressing it via the object's location remains possible), and that a correspondence process exists to link object files to the visual scene when the view changes, for instance, due to an eye movement. There is an important constraint, however: Only a limited number of object files can be active at the same time, matching the estimated capacity limit of about four objects that is found in working memory experiments. When this limit is reached, one of the existing object files is lost as soon as a new one is created. Note that Object File Theory does not claim that object files are the only form of visual working memory. There can also be memory of unbound object features, as shown in different psychophysical experiments. But if the features belonging to an object are stored in a bound form, the theory claims that this happens in the form of object files.

Concepts similar to the object files are also used in other theoretical work. Pylyshyn (2001) focused on the spatial aspect of scene perception, aiming in particular to explain experimental results on multi-object tracking. He proposed that a limited number of visual indexes are available that can be assigned to objects in the world and used to track them over movements. Object features may be associated with these visual indexes in working memory, yielding a representation analogous to object files. Such ideas have also been taken up in the work of Ballard (Ballard et al., 1997, discussed in the introduction of this thesis), who proposes a kind of pointer to solve the problem of variable binding in cognitive programs. In this approach, overt fixation or spatial attention to a visual object is one form of this pointer, but additional ones can be held in working memory. They provide access to more complex object representations (either sensory or in memory), and can be handed to cognitive routines so that these can be applied flexibly to different objects.

Interestingly, all of these conceptual models of human scene memory directly employ terminologies from computer science, namely pointers and files. In this analogies, biological working memory is treated as a form of addressable memory (albeit with some specific properties), in which different features of an object can be laid down in a cluster that constitutes a file. This file (as a location in memory) can then be addressed by a pointer. (It appears, however, that the direction of this pointer is not quite consistent between different theoretical approaches: In the Object File Theory, it is

assumed that the location of an object in the world can be used to address the object file. The visual indexes of Pylyshyn, in contrast, are described as pointers to objects and their locations in the world. It appears that both directions would be needed for different tasks.)

The fundamental problem with these analogies from computer science is that none of the theories specifies how the required structures and functions can actually be realized in neural systems. In particular, there is no obvious way to realize a variable pointer in a network of neurons. Working memory in biological systems is realized as sustained activity of specific populations of neurons. All interactions between neurons are mediated by synaptic connections, which can be considered as fixed in the context of scene viewing and memorization (although they may be adapted by learning on longer time scales). This means that a group of neurons may activate a certain working memory representation that it is synaptically connected to, but it can not switch to activate another working memory representation a few moments later. Groups of neurons thus cannot act as variable pointers, and working memory representations are not organized as addressable memory.

An additional, more specific issue for the neural implementation of object files is the exact structure of the working memory representations. The behavioral evidence described above points to separate WM capacities for different feature dimensions. This is consistent with the assumption of separate neural populations, sensitive for different surface features such as colors, shapes or orientations, as the basis for working memory. These capacity limitations still hold when the features of objects are memorized in a bound form, as tested by Wheeler and Treisman (2002). While the Feature Integration Theory does posit separate neural populations for representing different feature dimensions at the perceptual level (Treisman, 1988), its extension to the Object File Theory does not specify how the memory for different features within one object file is organized. Based on behavioral evidence and from what is known about the neural basis of working memory, it is quite clear that the image of a "file" as a segment of memory into which arbitrary information can be written is not fitting here.

So while the theoretical accounts presented here provide a conceptual model that can explain the behavioral data, they do not address the neural implementation of the model. Consequently, they also do not address the neural processes underlying the elementary operations that are necessary for working memory tasks like change detection. The conceptual theories do not specify how the sequential processing of individual objects is realized, how attention is shifted from one to the other, how the working memory representation—or object file—for each item is actually formed, and how it is compared to new perceptual items to detect changes. Moreover, the conceptual theories cannot explain the origin of the capacity limit, beyond stating that only a fixed number of object files or visual indexes are avail-

able; and they cannot give a reason why the sequential attentional focusing of objects is necessary to form bound representations, given the inherently parallel processing that the neural systems are capable of. The present chapter aims to address these questions by proposing a functional and autonomous model that implements the conceptual model of the Object File Theory in a neurally plausible fashion.

5.2 DNF model of scene representation

5.2.1 Outline of the model

The model to be described in this chapter should capture the neural processes underlying the formation of a scene representation in working memory. Moreover, the model should be able to use such a scene representation to perform change detection tasks, which constitute a key behavioral paradigm in testing human scene memory. The model is restricted to processing simple visual features as used in many of these tasks, and is not designed to deal with complex natural scenes. This makes it possible to focus on the general mechanism underlying the formation and use of scene working memory, without requiring a complex object recognition system (but see Zibner et al., 2011b for integration of a DNF model of scene representation and object recognition).

A simplified sketch of the full model architecture is shown in Figure 5.1, illustrating the basic organization of perceptual and working memory organizations. The perceptual part of the model (right part in the figure) is based directly on the biased competition model (Chapter 3). It is now extended to two elementary surface feature dimensions (color hue value and edge orientation) to allow modeling of feature bindings. The initial visual representations for these two feature dimensions are feature maps over retinal space. They are implemented as DNFs over a two-dimensional space, called visual sensory fields, spanned by one spatial dimension (the same for both feature maps) and one surface feature dimension. Note that the restriction to a single spatial dimension in the biased competition model is retained here, to reduce computation complexity and allow easier visualizations. The general mechanisms in the model would not change if a second spatial dimension was added, as done in the robotic implementation (Zibner et al., 2010a).

The separate feature maps do not directly code feature conjunctions, in the way that a single, three-dimensional field over the dimensions of color, orientation, and space would do. This avoids a combinatorial explosion when further dimensions, either for space or more surface features, are added. However, the feature maps are indirectly coupled to each other via the shared spatial dimension, in that they project to a single spatial attention field and

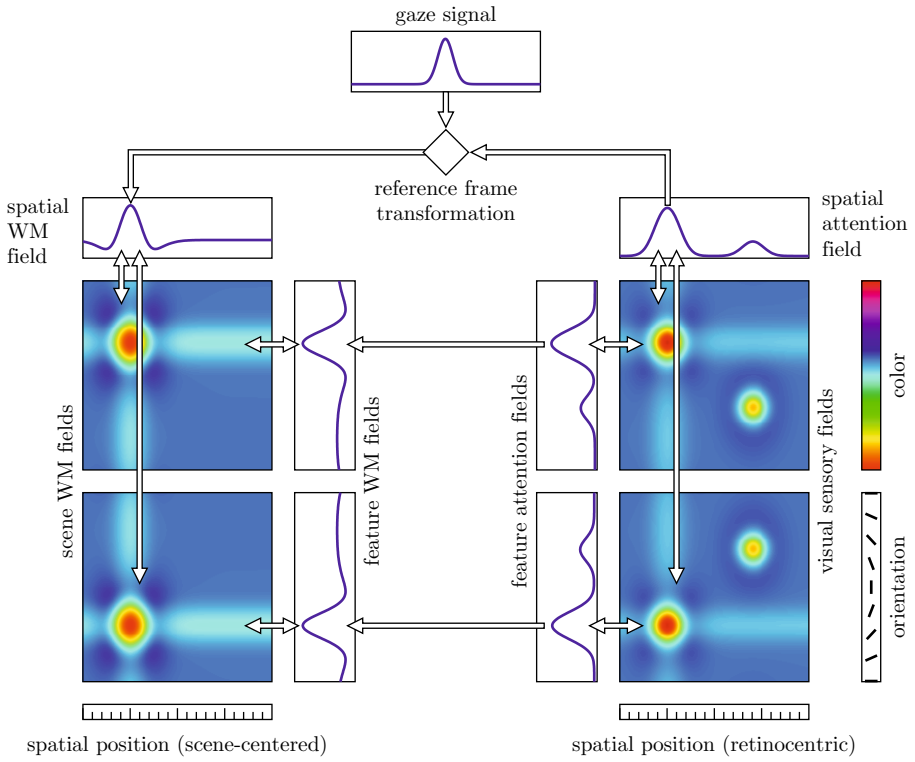


Figure 5.1: Simplified DNF architecture for a model of scene representation in working memory. Working memory for feature conjunctions is implemented as a stack of feature maps over space, coupled via a shared spatial dimension (left part of the figure), analogous to the representations at the perceptual level (right part of the figure).

both receive feedback from it. This layout is consistent with what is known about the neural architecture in early visual cortex (Krüger et al., 2013). Here, separate populations of neurons exist that are sensitive to different surface features, while at the same time having localized receptive fields in retinal space. The layout also matches the assumptions underlying Feature Integration Theory (Treisman, 1988).

The model also retains the separate pathways for spatial location and surface features from the biased competition model. Separate one-dimensional fields are used to mediate attention and working memory for each surface feature dimension and for space. These separate working memory representations reflect the psychophysical findings that unbound features can be memorized, and appear to be memorized more easily and robustly than feature conjunctions (Treisman and Zhang, 2006).

The core design question for the scene representation architecture is how

the working memory for feature conjunctions is realized (left part in Figure 5.1). This working memory representation should provide the functionality ascribed to object files, but in a neural implementation. A straightforward hypothesis is that the working memory representation for feature conjunctions is organized in a fashion directly analogous to the perceptual representations. This is implemented here by employing feature maps for the different surface features over space as the substrate for scene working memory.

Unlike on the perceptual level, the spatial dimension for the working memory representation should not employ retinal space. One of the functions of scene working memory is to integrate information over different fixations, and to this end, the working memory representation should be invariant over gaze shifts. I assume for the model that the spatial dimension for the working memory representation uses a reference frame that is aligned with the visual scene (not necessarily strictly allocentric, but invariant under different views of the scene). This is consistent with behavioral evidence that rearrangements of items in a scene has a disruptive effect on change detection, whereas a shift of a whole array of objects only weakly affects change detection performance (Hollingworth, 2007).

The use of different spatial reference frames between perceptual and scene working memory representations makes it necessary to introduce a reference frame transformation between these two levels. The DNF model of reference frame transformation presented in the previous chapter is an ideal candidate for this task, since it allows the continuous coupling between different multi-object representations. It is adopted in the present model in a simplified form, employing a direct convolution of the spatial representations with a shift signal. This is neurally less realistic, but reduces the complexity of this large architecture and speeds up simulations. The model then includes a spatial attention field in the retinal frame, and both a spatial attention field and a spatial working memory field in the scene reference frame, coupled to each other via the transformation system.

The full structure of the scene working memory is then as follows: A stack of feature maps over space, with one separate map for each surface feature dimension, provides the substrate for memory of bound features. The binding of different feature dimensions is achieved by coupling of these feature maps to the purely spatial working memory representation via their shared spatial dimension. This is supported by psychophysical findings that show a special role of location for feature binding in working memory (Pertzov and Husain, 2013). The feature maps are also coupled bidirectionally to the pure feature working memory via the shared feature dimension. This visual working memory architecture can therefore represent either only the unbound feature values, or form actually bound object representations by additionally employing the coupled feature maps. This is neurally plausible,

as recent fMRI results show that additional brain regions are engaged when memorizing feature bindings compared to memorizing individual feature values (Parra et al., 2014).

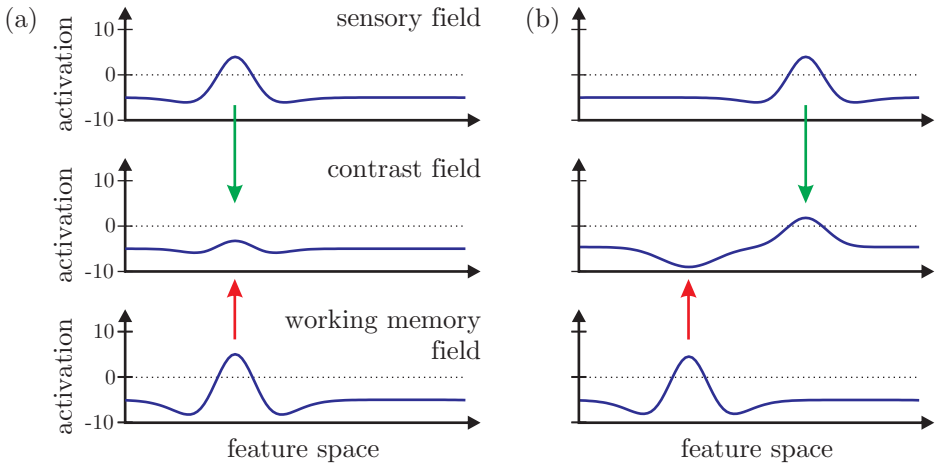


Figure 5.2: DNF mechanism for change detection. A sensory field provides excitatory input to a contrast field (green arrows), a working memory field provides inhibitory input (red arrows). (a) Feature match. (b) Mismatch of features.

The final building block for the architecture is the change detection mechanism. This is adapted from prior DNF models of change detection in spatial memory and feature memory (Schutte and Spencer, 2009; Johnson et al., 2009a,b). Change detection in these models is achieved in a contrast field, which receives excitatory sensory input and inhibitory input from a working memory field, all defined over the same feature space (Figure 5.2). In the earlier publications, this inhibitory input was achieved indirectly through a shared inhibitory layer, but in the present work it is implemented as a direct inhibitory projection. When the working memory matches the current sensory stimulus, the two inputs to the field cancel each other out (Figure 5.2a). However, when no memory was previously formed or when there is a mismatch between memory and stimulus features, the excitatory sensory input can induce an activation peak in the contrast field (Figure 5.2b). This peak indicates that a novel feature value is being detected, and can thus serve as change signal in a change detection tasks.

The present architecture contains a one-dimensional contrast field for each of the surface feature dimensions. Change detection thus occurs separately for different feature dimensions and does not operate directly on the bound features. As will be described below, the contrast fields in this configuration are sufficient to allow both the parallel detection of single feature

changes in a stimulus array, and the sequential detection of changes in feature binding. The full implementation contains an analogous contrast field in the spatial pathway, but this will be omitted in the following descriptions since it is not relevant for the tasks addressed here.

5.2.2 Binding problem and sequential processing

The perceptual processes in the neural field architecture are inherently parallel. If multiple visual stimuli are presented simultaneously, then multiple peaks will form in each visual sensory field to reflect all the stimuli. So why is there a need to sequentially attend to individual items to establish and memorize the feature bindings? In the present architecture, the reason lies in the separate pathways for spatial and surface feature information. Such separation (at least a partial one) is widely accepted in the neurobiological literature (Krüger et al., 2013). Going up in the cortical hierarchy for visual processing, neural representations of surface features in the ventral pathway show sensitivity for increasingly complex features and feature combinations, but at the same time show decreasing sensitivity for the spatial location of a feature. In the dorsal pathway, on the other hand, neurons are sensitive to spatial locations and certain spatio-temporal features such as movement, but not sensitive for surface features like color.

If now multiple stimuli are present in the visual scene, their locations and individual surface feature values are transmitted via these separate pathways. A binding problem occurs when these are going to be recombined in working memory, since the separate representations contain no information about which feature values belong together to the same object. In the DNF model, the recombination is implemented as an intersection of horizontal and vertical ridge inputs in a two-dimensional field. As discussed in Chapter 3, this recombination induces a single peak unambiguously recombining the individual features if only one vertical and one horizontal ridge inputs intersect. But if there are two or more ridge inputs in each dimensions (reflecting, e.g., the locations and colors of two stimuli), then they form four intersection points, and spurious combinations of locations and features are formed (see Figure 3.4 in the earlier chapter).

For this reason, the main transmission paths along the separate pathways run through the feature and spatial attention fields. These fields are operated in a selection regime, so they can only stably support a single activation peak at any time. The model autonomously performs a coupled selection decision to select a single stimulus item in both the spatial and feature attention fields, in the same way as described for the biased competition model. There are additional parallel projections via the one-dimensional spatial and feature working memory fields, that allow the system for instance to smoothly track changes in object features or object locations. However, the selection of an object by focused attention is necessary to form bound representations in

working memory, consistent with the Object File Theory.

This account of the binding problem still leaves open the question why a separation into multiple streams is used at all, in particular if these are to be recombined in the end. A closer look at the spatial pathway can give an answer to this question. The transmission via this pathway includes the reference frame transformation, which, as shown in the previous chapter, is a rather complex and resource-intensive operation when realized purely by synaptic connectivity. Retaining the surface feature information in this operation would require a further substantial increase in neural resources. Transforming a full three-dimensional feature map (two spatial and one feature dimensions) with the same mechanism as used in the purely spatial transformation would require a five-dimensional transformation field—and this would have to be repeated for each surface feature dimension. Moreover, in a biological system, these transformations have to be learned from experience, and the synaptic connections for the transformation would have to be learned for each feature dimension and each feature value, even though they are identical for different features. The separation into spatial and surface feature pathways therefore provides a significant benefit in terms of efficiency and generality of the system.

Similar considerations may independently hold for the surface feature pathway. In the biological neural system, neurons along this pathway show specific responses to increasingly complex feature combinations, while selectivity for spatial location decreases. This likely allows neurons at higher levels of the processing hierarchy to aid in identifying objects independent of their location in the retinal image. Keeping the full spatial information in this pathway would likely decrease the efficiency and generalization capabilities of this processing pathways, similar as in the spatial pathway. In the present model, this increase in feature complexity is not captured, and the feature pathway simply propagates elementary surface features. This is sufficient to solve change detection tasks with simple artificial stimuli.

Given the need for sequential processing of individual items, the challenge for the model is now to generate this sequentiality based on its temporally continuous and inherently parallel mode of operation. The emergence of discrete steps from the continuous evolution of activation patterns can be effected by the field interactions, as has already been described in the previous chapters. The selection of a single item, for instance, is achieved by a selection decision (constituting a bifurcation in the field dynamics) in a set of coupled fields with competitive lateral interactions. The present architecture must go beyond that by creating a series of attentional selections (and releases of the previous selection) to sequentially process the items of a visual scene, without any change in external input to drive this sequence.

To meet this goal, the architecture contains a set of specific structures to generate the sequential selection of items. First, a set of peak detector

nodes is used to detect the successful memorization or completed comparison of a single item. The nodes are driven by the scene attention fields that are coupled to the scene working memory fields. These nodes then drive the formation of activation peaks for an *inhibition-of-return* field, defined over a spatial dimension in the scene reference frame. This field forms peaks for already visited locations, and suppresses spatial attention to them. This both creates the release of attention from the presently selected location, and ensures that the same stimuli are not selected repeatedly as focus of attention. An inhibition-of-return effect is also known in the behavioral literature. For instance, humans are slower to initiate saccades to locations they have recently attended to in certain tasks (Posner and Cohen, 1984).

5.2.3 Model architecture

The model combines parallel processing of multiple items, mediated by multi-peak DNFs, with sequential processing of individual items, mediated by selective DNFs that support only a single activation peak at any time (Figure 5.3). Both types of processing occur both along the spatial and the feature pathways. The parallel processing supports the simultaneous detection of changes in feature values for all memorized items, and keeps the working memory representations aligned with the visual input. It also allows the parallel tracking of smooth changes in object locations or feature values to a limited degree (compare Spencer et al., 2012), but this will not be addressed in detail for the present architecture. The sequential processing is needed for memorizing feature conjunctions. It also underlies the detection of changes in feature conjunction, using the structures as the parallel change detection. I will describe the DNFs and connections for the different forms of processing separately below.

Multi-item representations for parallel processing

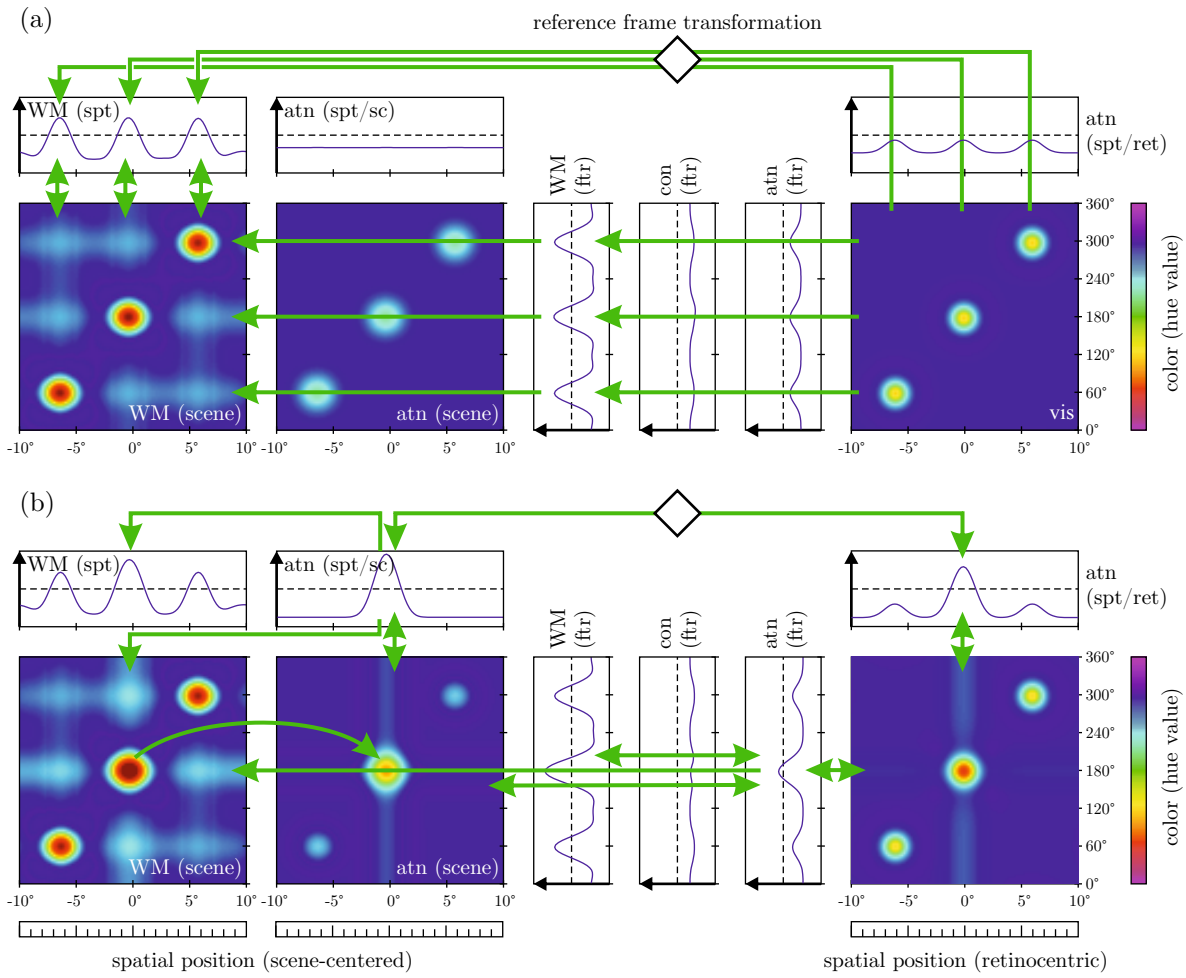
The external visual input to the system first induces localized activation peaks in a set of *visual sensory fields* (Figure 5.3a, right side). There is one of these fields for each surface feature dimension, namely color and orientation. The fields are defined over the two-dimensional space spanned by feature dimension and retinal space. The logarithmic scaling of retinal space that was used in the biased competition model is omitted here for simplicity. Otherwise, these DNFs behave in the same way as in that previous model, with moderate lateral interactions (local self-excitation and local surround inhibition) to support stabilized representations of visual stimuli, but without creating a working memory or selection regime. It is assumed in the model that each visual stimulus creates one activation peak in each of the visual sensory fields. These peaks are aligned with each other in the spatial dimension, while the feature values in different surface feature dimensions

are independent of each other.

Each visual sensory field is read out along both the spatial dimension and the surface feature dimension by integrating the field output over the disregarded dimension. The surface feature read-out provides input to the *feature WM fields* (Figure 5.3a, middle), defined over the one-dimensional space of surface feature values (color or orientation; note that only one surface feature is shown in the figure). These fields are set up in a multi-item working memory regime, with strong local self-excitation and local surround inhibition, such that they sustain multiple activation peaks without external input. The spatial read-out analogously projects to the *spatial working memory field*, which is set up in the same fashion. This read-out, however, is first convolved with a gaze signal indicating the current gaze direction within the visual scene. This implements a reference frame transformation from the retinal to the scene reference frame (Figure 5.3a, top).

For each surface feature dimension, a *scene WM field* is defined as activation distribution over the two-dimensional space spanned by the surface feature and one spatial dimension in the scene reference frame (leftmost fields in Figure 5.3). Like the one-dimensional WM fields, the scene WM fields are set up with strong lateral interactions to support multiple self-sustained peaks simultaneously. These scene WM fields are bidirectionally coupled to the one-dimensional WM fields: Each feature WM field is coupled to the corresponding scene WM field along the surface feature dimension, projecting horizontal ridge inputs into the scene WM field and receiving feedback from it (integrated over the spatial dimension). The spatial WM field projects vertical ridge inputs into both of the scene WM field and likewise receives input from both (integrated over the surface feature dimensions). Through this coupling, WM peaks in all these fields mutually support and stabilize each other, and retain their alignment with each other when activation peaks drift over time.

Note that the scene WM fields cannot form peaks based on the parallel inputs of the one-dimensional WM fields alone, due to the binding problem that occurs when intersecting multiple ridge inputs. They require additional inputs for a single, attentionally focused item to form activation peaks, as described below. Likewise, the formation of peaks in the one-dimensional WM fields is tied to visual attention under normal conditions, as it was also implemented in the biased competition model. However, they can form peaks from the parallel input provided by the visual sensory fields when they are globally boosted, reflecting a possibility for parallel memorization of features that is indicated by some experimental results.


(a)
reference frame transformation
WM (spt)
atn (spt/sc)
atn (spt/ret)
WM (ftr)
con (ftr)
atn (ftr)
WM (scene)
atn (scene)
vis
color (hue value)
(b)
spatial position (scene-centered)
spatial position (retinocentric)

Single-item representations for sequential processing

The representation of the present visual stimuli in the visual sensory fields forms the basis not only for the parallel processing, but also for the sequential processing. Each visual sensory field is bidirectionally coupled to a *feature attention field* and to the single *spatial attention field (retinal)*, with the same set-up as used in the biased competition model. The one-dimensional attention fields each receive the read-out along the corresponding dimension from the visual sensory fields, and project ridge inputs back to them (right part of Figure 5.3b). Lateral interactions in these fields create a selection regime, allowing only a single activation peak to form even in the presence of multiple inputs. As in the biased competition model, the coupling between the fields ensures that the same item is selected in all of the one-dimensional attention fields. Compared to that previously described model, the selection is now extended to span two surface feature dimension, which are coupled to each other via the single spatial attention field. To make sure that the simultaneous selection in the attention fields is always consistent with each other, the coupling between the visual sensory fields to the spatial attention field is slightly stronger than the coupling to the feature attention fields, so that the effect of the shared spatial selection is dominant over feature-based selection.

Each feature attention field provides input to the corresponding feature WM memory field (in addition to the parallel input from the visual sensory field), ensuring that a self-sustained peak is formed in that field when a visual item is attended. The feature attention field in turn receives feedback from the feature WM field in the same way as introduced in the biased competition model. In the spatial pathway, the reference frame transformation has to be applied again. The retinal spatial attention field is first convolved with the gaze signal, and the result drives activation in the *scene spatial attention*

Figure 5.3 *(preceding page)*: Parallel and selective processing along separate spatial and feature pathways in the DNF model of scene representation. The connectivity is shown only for a single surface feature dimension (color), the connection patterns for the second surface feature dimension (orientation) are identical. (a) Projections (green arrows) between fields that can contain multiple activation peaks simultaneously for the parallel processing of feature values and locations. (b) Projections between fields that are operated in a selection regime for the processing of a single visual item. Note that the two sets of fields on the left are defined over the same spatial dimension, although they are not aligned along that dimension in the figure. Abbreviations: vis - visual sensory fields, atn - attention fields, con - contrast fields, WM - working memory fields, ftr - surface feature representations, spt - spatial representations, ret - retinocentric, sc - scene-centered.

field. These fields are typically tightly coupled and represent the same spatial information, only in different reference frames. However, they can be decoupled for certain tasks as described in below. The scene spatial attention field provides an additional input to the spatial WM field, analogous to the connectivity in the feature pathway.

The spatial and feature attention fields also provide additional inputs to the two-dimensional scene working memory fields. Each scene WM field receives one ridge input from the feature attention field of the corresponding surface feature (horizontal ridge in Figure 5.3b), and one ridge input from the spatial attention field (vertical ridge in the figure). These add to the ridge inputs that the scene WM fields receive from the one-dimensional spatial and feature WM fields (which may provide multiple input ridges along each dimension). Only the combination of all these inputs is sufficient to bring the field activation to the output threshold and induce self-sustained activation peaks. Consequently, a new peak can only form for the single, currently attended item, even if there are multiple stimuli in the visual scene.

An additional set of attention fields is employed in the model, the *scene attention fields*, that enable the system to select individual item from the scene WM representation. These fields are defined over the same two-dimensional spaces as the scene WM fields, spanned by one surface feature dimension and a spatial dimension in the scene reference frame. Each scene attention field receives input from its corresponding scene WM field (curved arrow in Figure 5.3b). This input is localized, meaning that an activation peak in the scene WM field induces a local hill of activation at the corresponding location in the scene attention field. The scene attention fields are bidirectionally coupled to the one-dimensional attention fields. They receive ridge inputs in the same configuration as for the scene WM fields, which can be used to select an item from scene working memory by inducing a supra-threshold activation peak from the sub-threshold hills of activation. Like the one-dimensional attention fields, the scene attention fields feature lateral interactions with local excitation and strong global inhibition, producing a selection regime that support only a single activation peak at any time.

Change detection

Changes between a memorized scene and the current visual stimuli can be detected using the feature contrast fields. There is one feature contrast field for each of the surface feature dimensions, and each of them receives one pair of inputs from multi-item fields, and one pair of inputs from single-item fields (Figure 5.4).

The excitatory multi-item input for the contrast fields comes directly from the visual sensory fields (three parallel, long green arrows in Figure 5.4). An inhibitory input from the one-dimensional feature WM fields

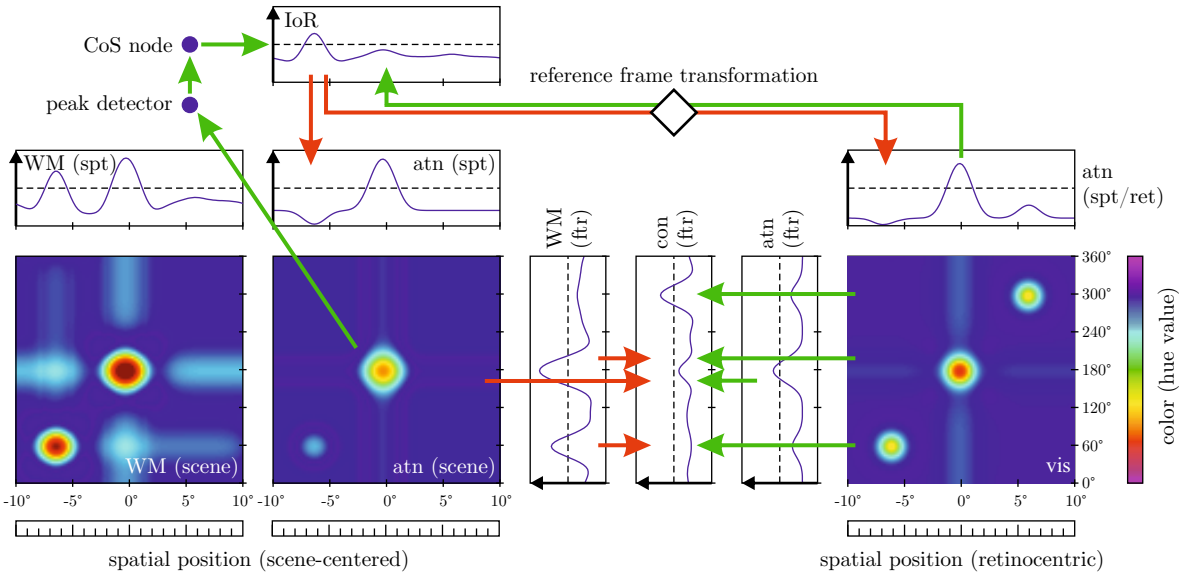


Figure 5.4: Change detection and autonomous sequence generation in the DNF model of scene representation. Change detection is achieved by the combination of excitatory (green arrows) and inhibitory inputs (red arrows) to the feature contrast fields (bottom center). The sequence generation involves the peak detector nodes driven by input from the scene attention fields, the CoS node, and the IoR field.

acts as antagonist for these excitatory inputs (two parallel short red arrows in Figure 5.4; the third item has not yet been memorized in the depicted situation). When WM representations and present visual stimuli match each other, these excitatory and inhibitory inputs cancel each other out. Simultaneously, the contrast fields receive excitatory inputs from the one-dimensional attention fields, and inhibitory input from the scene attention fields (read out along the feature dimension). These connections allow the field to detect a difference between a specific item in the current visual scene and a selected item from scene WM.

The contrast fields also project weakly back to the one-dimensional attention fields in an excitatory fashion, and thereby biases attention toward changed items in a scene, or items that have not yet been focused during memorization. This implements and autonomous deployment of attention to the locations of feature changes, as observed in change detection experiments (Hyun et al., 2009b).

Autonomous sequence generation

Several components in the architecture contribute to the transition of focused attention from one item to the next, either when it has been memorized successfully or when the sequential change detection operation for an item is completed. Based on previous models of sequential order and behavior organization in DNF architectures (Sandamirskaya and Schöner, 2010; Sandamirskaya et al., 2011; Richter et al., 2012), the system does not rely on a fixed timing for the sequential inspection of items. Instead, a *condition of satisfaction* is defined that indicates the processing of one item is completed. This condition of satisfaction is the formation of salient activation peaks in all of the scene attention fields. As detailed below in the demonstrations, the scene attention fields form peaks after the memorization of an item is complete (because then they receive converging input from the scene WM fields and the one-dimensional attention fields). They also form peaks during the change detection process when a specific item in scene WM is being compared to an item in the current visual input.

The condition-of-satisfaction system is implemented via a *peak detector node* defined for each of the scene attention fields (top left in Figure 5.4). The node is driven by the field output, integrated over the whole field, and it becomes activated when this total output exceeds a certain threshold. The peak detector nodes for all surface feature dimensions project to a single *condition-of-satisfaction (CoS) node*, with connection weights chosen in such a fashion that all peak detector nodes need to be active in order to activate the CoS node (note that only a single surface feature dimension is shown in the figure, but a second one is present in the model).

The CoS node then acts by globally increasing the activation level in the *inhibition-of-return (IoR) field*. This field is defined over the spatial dimension in the scene reference frame. It receives localized input from the retinal spatial attention field as well as from the visual sensory fields (both first transformed to the scene reference frame), and it features strong lateral interactions that support a multi-peak working memory regime. When this field is globally activated by the CoS node, it forms a peak for the currently attended spatial location. These peaks then remain sustained by lateral interactions in combination with weak input from the visual sensory field, and provide a memory of which locations in the scene have been inspected. The IoR field projects local inhibition to both spatial attention fields (Figure 5.4), which suppresses the peaks for the current attentional selection. This release from attention is supported by the CoS node globally inhibiting the scene attention fields and feature attention fields.

The inhibition from the IoR field for previously attended locations promotes the direction of spatial attention to items in the scene that have not yet been inspected. The continuous spatial input from the visual sensory field allows spatial tracking of item's locations by the sustained peaks in the

IoR field to a certain degree. Thus, items will not be treated as novel if they have moved to a new location, consistent with experimental findings that inhibition of return is object-based, rather than strictly location-based (Tipper et al., 1994). When a visual stimulus (or the whole stimulus array) is removed from the scene, the input from the visual sensory field ceases and the activation peaks in the IoR field collapse after a brief time. Newly appearing stimuli, even at previously inspected locations, will then be treated as new and can be attentionally selected again.

5.2.4 Field equations and parameter values

As in the DNF model described in the previous chapters, the complete DNF architecture for scene representation and change detection is implemented in a set of differential equations that describe the evolution of activation patterns over time. A unique two-letter index is used to identify each field in the equations, given in Table 5.1. The fields representing surface features are furthermore identified by a superscript specifying the surface feature dimension (col for color and orn for orientation). Connection patterns and parameter values are identical between these two surface feature dimensions, so they are generally treated jointly in the equations.

field name	field index	c^{exc}	c^{inh}	c^{gi}
visual sensory fields	vs	7.5	7.5	0.002
retinal spatial attention field	ra	8	0	0.6
scene-level spatial attention field	sa	6	0	0.5
spatial contrast field	sc	20	20	0
spatial WM field	sm	27	25	0
IoR field	ir	20	15	0
feature attention fields	fa	4	0	0.5
feature contrast fields	fc	20	20	0
feature WM fields	fm	27	25	0
scene attention fields	ca	2.5	0	0.0175
scene WM fields	cm	25	27.5	0.05
peak detector nodes	pd	4	0	0
CoS node	cs	4	0	0

Table 5.1: Field indices and lateral interaction parameters.

The two visual sensory fields are governed by the field equations

$$\begin{aligned}\tau \dot{u}^d_{\text{vs}}(x,y) = &- u^d_{\text{vs}}(x,y) + h + i^d_{\text{vs}}(x,y) + [k_{\text{vs,vs}} * f(u^d_{\text{vs}})](x,y) \\ &+ [k_{\text{vs,ra}} * f(u_{\text{ra}})](x) + [k_{\text{vs,fa}} * f(u^d_{\text{fa}})](y) + q_{\text{vs}}\xi(x,y)\end{aligned} \tag{5.1}$$

for $d \in \{\text{col}, \text{orn}\}$. The input pattern i^d_{vs} is a superposition of two-dimensional non-normalized Gaussians with width $\sigma = 4$ and amplitude $a = 6$, centered

on the combinations of stimulus positions and feature values for the present stimulus array. For projections along the separate spatial and feature pathways, integrated outputs are computed as

$$F_{\text{ftr}}(u_{\text{vs}})(y) = \int f(u_{\text{vs}}(x,y))dx \tag{5.2}$$

$$F_{\text{spt}}(u_{\text{vs}})(x) = \int f(u_{\text{vs}}(x,y))dy. \tag{5.3}$$

Equivalent integrals are defined for the other two-dimensional DNFs in the architecture, u_{ca} and u_{cm}.

The field equation for the retinal spatial attention field is

$$\begin{aligned}\tau\dot{u}_{\text{ra}}(x) = &- u_{\text{ra}}(x) + h + [k_{\text{ra,ra}} * f(u_{\text{ra}})](x) \\ &+ \sum_{d\in\{\text{col,orn}\}} [k_{\text{ra,vs}} * F_{\text{spt}}(u^d_{\text{vs}})](x) \\ &+ c_{\text{ra,sa}}[i_{\text{gd}} \star f(u_{\text{sa}})](x) - c_{\text{ra,ir}}[i_{\text{gd}} \star f(u_{\text{ir}})](x) + q_{\text{ra}}\xi(x).\end{aligned} \tag{5.4}$$

The notation $[i_{\text{gd}} \star f(u_{\text{sa}})]$ is used here for the correlation operation that implements the reference frame shift from the scene reference frame back to the retinocentric reference frame, based on the gaze signal i_{gd}. It is equivalent to the convolution of the field output with a mirrored version of the gaze signal (for real-valued inputs). The gaze signal itself is a Gaussian pattern with width $\sigma = 4$ centered on the current gaze direction.

The remaining fields in the spatial pathway are defined over space in the gaze-invariant scene reference frame. Retinocentric input to these fields is shifted by a convolution with the gaze signal, yielding terms of the form $[i_{\text{gd}} * f(u_{\text{ra}})]$. The resulting field equations for the scene-level spatial attention field (u_{sa}), spatial contrast field (u_{sc}), spatial WM field (u_{sm}), and IoR field (u_{ir}) are as follows:

$$\begin{aligned}\tau\dot{u}_{\text{sa}}(x) = &- u_{\text{sa}}(x) + h + [k_{\text{sa,sa}} * f(u_{\text{sa}})](x) \\ &+ c_{\text{sa,ra}}[i_{\text{gd}} * f(u_{\text{ra}})](x) + [k_{\text{sa,ir}} * f(u_{\text{ir}})](x) \\ &+ [k_{\text{sa,sc}} * f(u_{\text{sc}})](x) + [k_{\text{sa,sm}} * f(u_{\text{sm}})](x) \\ &+ \sum_{d\in\{\text{col,orn}\}} [k_{\text{sa,sm}} * F_{\text{spt}}(u^d_{\text{ca}})](x) + q_{\text{sa}}\xi(x)\end{aligned} \tag{5.5}$$

$$\begin{aligned}\tau\dot{u}_{\text{sc}}(x) = &- u_{\text{sc}}(x) + h + [k_{\text{sc,sc}} * f(u_{\text{sc}})](x) \\ &+ \sum_{d\in\{\text{col,orn}\}} c_{\text{sc,vs}}[i_{\text{gd}} * F_{\text{spt}}(u^d_{\text{vs}})](x) + [k_{\text{sc,sa}} * f(u_{\text{sa}})](x) \\ &- [k_{\text{sc,sm}} * f(u_{\text{sm}})](x) - \sum_{d\in\{\text{col,orn}\}} [k_{\text{sc,sm}} * F_{\text{spt}}(u^d_{\text{ca}})](x) \\ &+ q_{\text{sc}}\xi(x)\end{aligned} \tag{5.6}$$

$$\begin{aligned}\tau\dot{u}_{\mathrm{sm}}(x) =& -u_{\mathrm{sm}}(x) + h + [k_{\mathrm{sm,sm}} * f(u_{\mathrm{sm}})](x) \\ &+ \sum_{d\in\{\mathrm{col,orn}\}} c_{\mathrm{sm,vs}}[i_{\mathrm{gd}} * F_{\mathrm{spt}}(u^d_{\mathrm{vs}})](x) + [k_{\mathrm{sm,sa}} * f(u_{\mathrm{sa}})](x) \\ &+ \sum_{d\in\{\mathrm{col,orn}\}} [k_{\mathrm{sm,cm}} * F_{\mathrm{spt}}(u^d_{\mathrm{cm}})](x) + q_{\mathrm{sm}}\xi(x)\end{aligned} \tag{5.7}$$

$$\begin{aligned}\tau\dot{u}_{\mathrm{ir}}(x) =& -u_{\mathrm{ir}}(x) + h + [k_{\mathrm{ir,ir}} * f(u_{\mathrm{ir}})](x) \\ &+ \sum_{d\in\{\mathrm{col,orn}\}} c_{\mathrm{ir,vs}}[i_{\mathrm{gd}} * F_{\mathrm{spt}}(u^d_{\mathrm{vs}})](x) \\ &+ c_{\mathrm{ir,ra}}[i_{\mathrm{gd}} * f(u_{\mathrm{ra}})](x) + c_{\mathrm{ir,cs}} f(u_{\mathrm{cs}}) + q_{\mathrm{ir}}\xi(x)\end{aligned} \tag{5.8}$$

In the feature pathways, the differential equations for the feature attention fields (u_{fa}), feature contrast fields (u_{fc}), and feature WM field (u_{fm}) are given as

$$\begin{aligned}\tau\dot{u}^d_{\mathrm{fa}}(y) =& -u^d_{\mathrm{fa}}(y) + h + [k_{\mathrm{fa,fa}} * f(u^d_{\mathrm{fa}})](y) \\ &+ [k_{\mathrm{fa,vs}} * F_{\mathrm{ftr}}(u^d_{\mathrm{vs}})](y) + [k_{\mathrm{fa,fc}} * f(u^d_{\mathrm{fc}})](y) \\ &+ [k_{\mathrm{fa,fm}} * f(u^d_{\mathrm{fm}})](y) - c_{\mathrm{fa,cs}} f(u_{\mathrm{cs}}) + q_{\mathrm{fa}}\xi(y)\end{aligned} \tag{5.9}$$

$$\begin{aligned}\tau\dot{u}^d_{\mathrm{fc}}(y) =& -u^d_{\mathrm{fc}}(y) + h + [k_{\mathrm{fc,fc}} * f(u^d_{\mathrm{fc}})](y) \\ &+ [k_{\mathrm{fc,vs}} * F_{\mathrm{ftr}}(u^d_{\mathrm{vs}})](y) + [k_{\mathrm{fc,fa}} * f(u^d_{\mathrm{fa}})](y) \\ &- [k_{\mathrm{fc,fm}} * f(u^d_{\mathrm{fm}})](y) - [k_{\mathrm{fc,ca}} * F_{\mathrm{ftr}}(u^d_{\mathrm{ca}})](y) + q_{\mathrm{fc}}\xi(y)\end{aligned} \tag{5.10}$$

$$\begin{aligned}\tau\dot{u}^d_{\mathrm{fm}}(y) =& -u^d_{\mathrm{fm}}(y) + h + [k_{\mathrm{fm,fm}} * f(u^d_{\mathrm{fm}})](y) \\ &+ [k_{\mathrm{fm,vs}} * F_{\mathrm{ftr}}(u^d_{\mathrm{vs}})](y) + [k_{\mathrm{fm,fa}} * f(u^d_{\mathrm{fa}})](y) \\ &+ [k_{\mathrm{fm,cm}} * F_{\mathrm{ftr}}(u^d_{\mathrm{cm}})](y) + q_{\mathrm{fm}}\xi(y)\end{aligned} \tag{5.11}$$

for $d \in \{\mathrm{col}, \mathrm{orn}\}$.

The two pathways converge in the scene attention field (u_{ca}) and scene WM field (u_{cm}), both defined over a combination of spatial and feature dimensions, with field equations

$$\begin{aligned}\tau\dot{u}^d_{\mathrm{ca}}(x,y) =& -u^d_{\mathrm{ca}}(x,y) + h + [k_{\mathrm{ca,ca}} * f(u^d_{\mathrm{ca}})](x,y) \\ &+ [k_{\mathrm{ca,sa}} * f(u_{\mathrm{sa}})](x) + [k_{\mathrm{ca,fa}} * f(u^d_{\mathrm{fa}})](y) \\ &+ [k_{\mathrm{ca,cm}} * f(u^d_{\mathrm{cm}})](x,y) - c_{\mathrm{ca,cs}} f(u_{\mathrm{cs}}) + q_{\mathrm{ca}}\xi(x,y)\end{aligned} \tag{5.12}$$

$$\begin{aligned}\tau\dot{u}^d_{\mathrm{cm}}(x,y) =& -u^d_{\mathrm{cm}}(x,y) + h + [k_{\mathrm{cm,cm}} * f(u^d_{\mathrm{cm}})](x,y) \\ &+ [k_{\mathrm{cm,sa}} * f(u_{\mathrm{sa}})](x) + [k_{\mathrm{cm,sm}} * f(u_{\mathrm{sm}})](x) \\ &+ [k_{\mathrm{cm,fa}} * f(u^d_{\mathrm{fa}})](y) + [k_{\mathrm{cm,fm}} * f(u^d_{\mathrm{fm}})](y) \\ &+ q_{\mathrm{cm}}\xi(x,y).\end{aligned} \tag{5.13}$$

Note that in the feature WM fields, lateral inhibition is local only in the spatial dimension, and global along the feature dimension.

Finally, the peak detector nodes for the two surface feature dimensions are governed by the differential equation

$$\tau \dot{u}^d_{\text{pd}} = -u^d_{\text{pd}} + h + c_{\text{pd,pd}} f(u^d_{\text{pd}}) + c_{\text{pd,ca}} \iint f(u^d_{\text{ca}}(x,y))dxdy, \qquad (5.14)$$

and the CoS node that drives the sequential processing of stimuli is governed by the equation

$$\tau \dot{u}_{\text{cs}} = -u_{\text{cs}} + h + c_{\text{cs,cs}} f(u_{\text{cs}}) + c_{\text{cs,pd}} \left(f(u^{\text{col}}_{\text{pd}}) + f(u^{\text{orn}}_{\text{pd}}) \right). \qquad (5.15)$$

projection	c	projection	c	projection	c
ra, sa	3	fa, cs	4	vs, fa	1.25
ra, ir	10	fa, vs	1	vs, ra	1.5
ra, vs	0.8	fa, fc	1.75	ca, fa	2
sa, ra	10	fa, fm	1	ca, sa	2.75
sa, ir	10	fc, vs	1.25	ca, cm	6
sa, ca	0.75	fc, fa	7.5	ca, cs	2
sc, sa	2.5	fc, ca	0.75	cm, sa	1
sc, sm	8	fm, vs	0.4	cm, sm	1.75
sc, vs	0.625	fm, fa	3.5	cm, fa	1
sc, ca	0.375	fm, fc	8	cm, fm	1.75
sm, sa	3.5	fm, cm	0.1	ir, ra	0.5
sm, vs	0.2	pd, ca	0.05	ir, cs	2.25
sm, cm	0.05	cs, pd	3	ir, vs	0.15

Table 5.2: Connection strengths for projections between fields.

All feature spaces (horizontal stimulus position, gaze direction, color hue values, and edge orientation) are sampled with 100 units for numerical simulations. The widths of interaction kernels are given in these field units. The temporal step size for the Euler method is $\Delta t = 1$ (in arbitrary units, a mapping to real time is not yet specified for this model).

All fields and nodes use the same resting level $h = -5$, steepness parameter $\beta = 4$ for the sigmoid output function, and time constant $\tau = 10$. The lateral interactions in all fields can be described by a difference-of-Gaussians kernel with $\sigma^{\text{exc}} = 4$ and $\sigma^{\text{inh}} = 8$, and a global inhibitory component for some fields. The lateral interaction weights are given in Table 5.1. All interaction kernels for projections between fields are Gaussians with width $\sigma = 4$, the weights for these projections are listed Table 5.2.

5.3 Demonstrations

In this section, I will describe in detail how the DNF architecture for scene representation can solve three different classes of change detection tasks used

in the psychophysical literature. Each task consists of the presentation of a sample array, followed after a brief delay by the presentation of the test array. Each array here consists of three colored, oriented bar stimuli. The DNF system has to memorize the sample array and maintain this memory for the following change detection in the test scene.

In the first type of change detection task, the feature change detection, the system has to detect whether any novel feature values (like a new color) were introduced in the test array that were not present in the sample array. In the feature location change detection task, the system must detect whether any feature values have changed their locations (e.g., if colors were swapped between two items in the array). Finally, in the feature conjunction change detection task, the system must determine whether the same items (defined by a combination of a color and an orientation) are present in sample and test array, independent of their location.

5.3.1 Sequential formation of visual working memory

The first step in every change detection task is the memorization of the sample array. In accordance with the Feature Integration Theory, the model performs this operations as a sequential process in which each visual item is attended individually. The sequential approach is directly supported by evidence from human psychophysical experiments (Vogel et al., 2006). I will first address the sequential selection of visual items through covert attention, without any actual gaze changes, and then address the scanning of a scene through a series of fixations and saccades below.

The beginning of the memorization process in the DNF model is illustrated in Figure 5.5. The sample array to be memorized consists of three colored, oriented bars, shown in the top left. These visual stimuli first induce activation peaks in the visual sensory fields, with one peak for every item in each of these fields. These peaks provide input to the one-dimensional attention fields along the separate spatial and surface feature pathways. When activation levels reach the output threshold in the attention fields, the lateral interactions within these fields and their bidirectional coupling with the visual sensory fields autonomously initiate a selection process. Through this process, a single visual item is focused by attention, with its features and location represented by peaks in the separate attention fields, and the corresponding peaks in the visual sensory fields enhanced by feedback.

Following the attentional selection, peaks for the individual features of the selected item are induced in the one-dimensional WM fields. The feature attention fields project directly to the feature WM fields, and their input, together with direct input from the visual sensory fields, induces an activation peak in each WM field. In the spatial pathway, the retinal spatial attention field first projects to the scene spatial attention field via the reference frame transformation. The scene spatial attention field then induces a peak in the

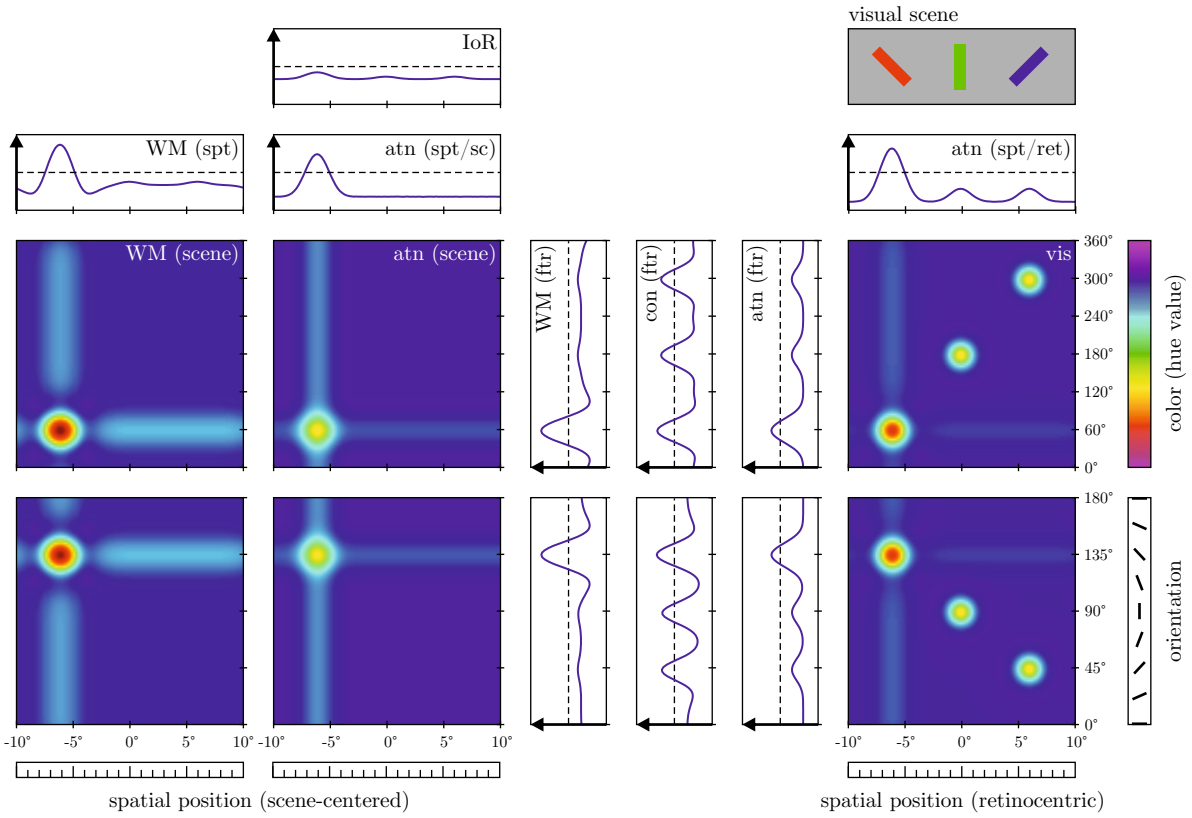


Figure 5.5: Activation patterns in the DNF model of scene representation at the beginning of the memorization period. The leftmost item in the visual scene has been selected by attention, and a distributed working memory representation for the item's location and surface features has been formed. The gaze direction is assumed to be straight ahead, so that retinocentric and scene-centered reference frames are directly aligned in the figure.

spatial WM field.

Both the one-dimensional attention fields and the one-dimensional WM fields now project ridge inputs into the scene WM fields (vertical ridges from the spatial representations, horizontal ridges from the surface feature representations). For the first item to be memorized, the inputs from the one-dimensional WM fields are completely aligned with the inputs from the attention fields, so effectively only a single vertical and a single horizontal activation ridge is induced in each scene WM field. Activation peaks form at the intersections of these input ridges, reflecting the combination of feature values and location in the scene for the selected item.

When an activation peak has formed in the scene WM field, it projects localized input to the corresponding scene attention field. This field also

receives ridge inputs from the one-dimensional attention fields, and the position of the localized input matches the intersection point of these ridges. This combination of inputs induces an activation peak in the scene attention field. This peak does not directly contribute to the memorization process itself, but it indicates that the formation of a memory peak for the currently attended item in the scene WM field is completed. This activates the condition of satisfaction system described above. An activation peak for the currently attended location forms in the IoR field, and suppresses activation for this location in the spatial attention fields. The CoS node simultaneously inhibits the scene attention fields and the feature attention field, so that the attentional selection of the current item is released simultaneously in all attention fields. After the peaks in the scene attention fields have disappeared, the activation values of the nodes in the condition-of-satisfaction system fall back to their resting level.

When the focus of attention is released from the item that was selected first, the remaining items start anew to compete for the formation of activation peaks in the one-dimensional attention fields. The location of the previously selected item is inhibited in this competition by the sustained activation peak in the IoR field, so that it cannot be selected again. A new item from the visual scene is selected, and is memorized in the same fashion as the first item.

It is important that during this memorization of additional items, the WM fields already contain one or more peaks, but these do not interfere with the process. This is shown in Figure 5.6 for the memorization of the last of three items in a stimulus array. Self-sustained peaks exist in the spatial WM field, feature WM fields, and scene WM fields, and these are coupled by mutually projecting inputs to each other, visible in particular in the multiple vertical and horizontal input ridges in the scene WM fields. These different ridge inputs also intersect and could give rise to incorrect bindings, but by themselves these inputs are too weak to induce any peaks. Only the addition of the input ridges from the feature and spatial attention fields produce new peaks in the scene WM field, as shown in the figure, and these newly created peaks correctly reflect the features and location of the single, currently attended item.

Note that the lateral inhibitory interactions between peaks in the WM fields creates a natural capacity limit for the number of items that can be memorized. The more activation peaks are present in the field, the more long-range inhibition is generated that depresses the activation of all peaks and can ultimately make them collapse. When a new peak is added while the system of coupled WM fields is already at its capacity limits, this new peak is likely to prevail (due to the strong convergent inputs that support it during creation), but other, neighboring peaks are likely to decay. Since corresponding peaks in the different WM fields for space, features, and feature

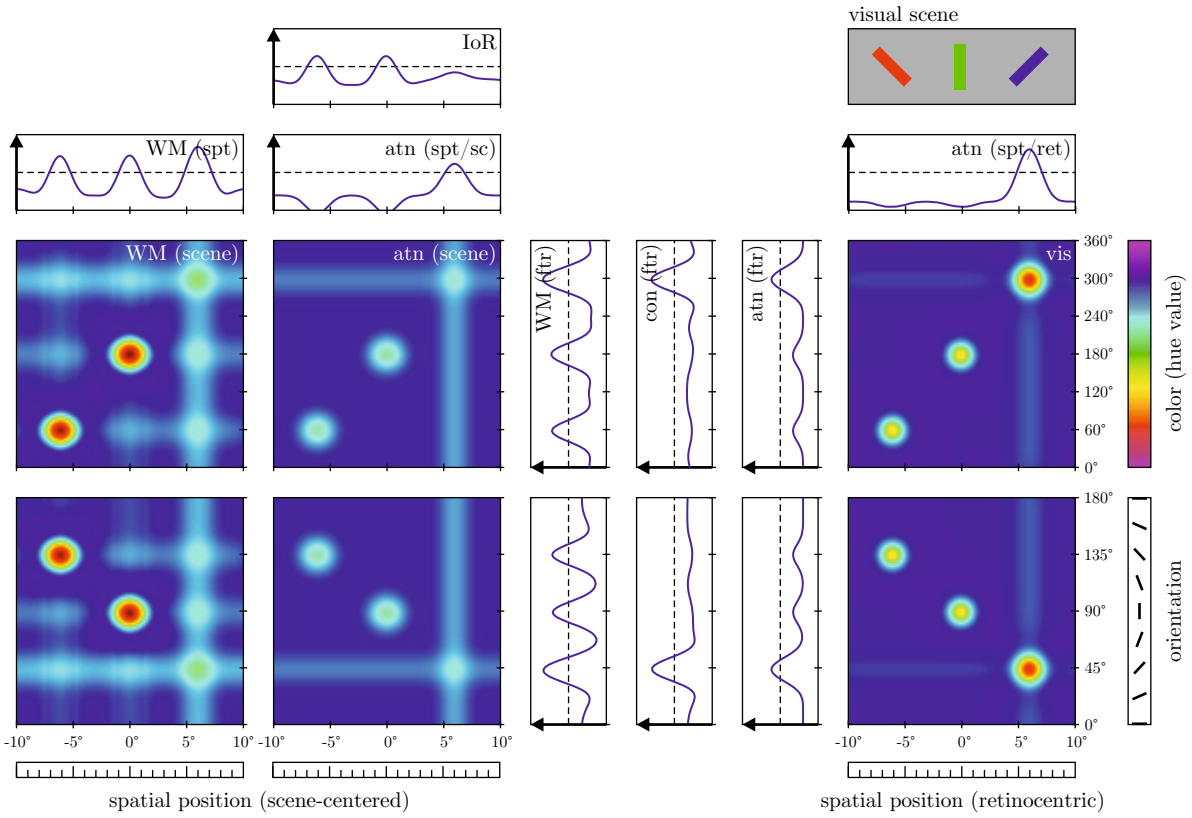


Figure 5.6: Activation patterns in the DNF model at the end of the memorization period. The rightmost item in the scene has now been selected by attention, and a working memory representation for this item is about to form. The two other items are already represented by peaks in the working memory fields, and peaks in the IoR field indicate that the corresponding locations have been inspected.

conjunction mutually support each other, the loss of an activation peak in one of the fields generally leads to the subsequent decay of the other coupled peaks and thereby to the forgetting of the item as whole.

The memorization process can also be performed with overt gaze shifts toward each item when it is memorized. To this end, a simulated gaze change is executed whenever a non-foveated item is selected by spatial attention. A saccadic motor system is currently not integrated into this model, but the generation of the saccade signal may be assumed to occur in the same way as in the DNF model of biased competition, driven by the retinocentric spatial attention field. This is emulated here as follows: For a fixed duration, the visual input is turned off, and then re-activated at new locations to reflect the shift in the retinal image. The gaze signal used in the reference frame

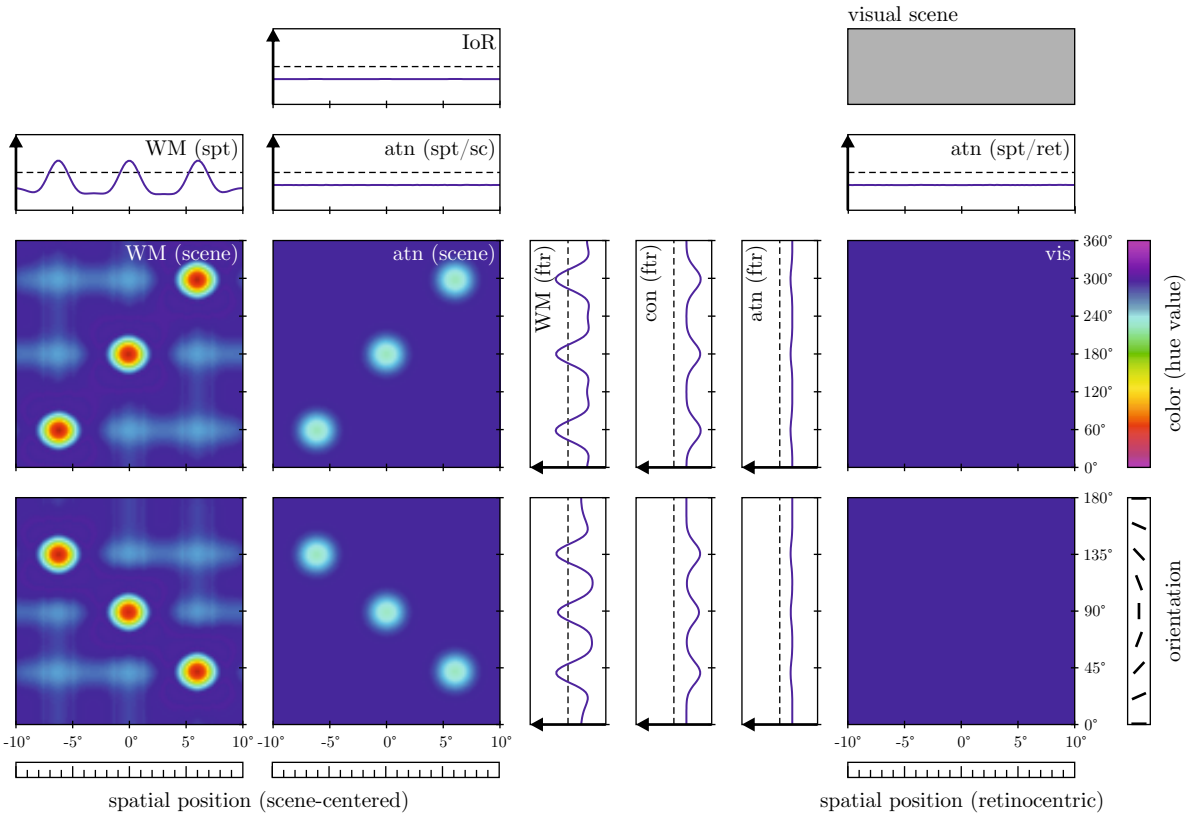


Figure 5.7: Activation patterns in the DNF model after a stimulus array has been memorized. The visual input has been turned off, the activation in the sensory fields has returned to the resting level, but the distributed working memory representation persists in the form of coupled sustained peaks.

transformation is updated according to the size of the simulated saccade. In addition to this, the retinocentric spatial attention field is globally suppressed, since its activation pattern is made obsolete by the gaze shift. In contrast, the scene-centered spatial attention field and the feature attention field are globally activated, so that they retain their activation peaks over the duration of the saccade (with no visual input). This ensures that after the completion of the gaze shift, the same visual stimulus is still attentionally selected. A peak in the retinal spatial attention field forms at the new retinal location of the stimulus, driven by feedback from the scene-centered spatial attention field and the biasing inputs from the feature attention fields, and the now foveated stimulus is memorized in the same fashion as before. Since all working memory representations are in the scene-centered reference frame, they remain unaffected by the gaze change.

After all items in the scene have been inspected and peaks have formed

in the IoR field for all item locations, the sequential deployment of attention effectively stops. When the sample array is then turned off in the visual input, the peaks in the IoR field decay. What remains is the scene representation in working memory, consisting of the coupled representations in the separate spatial and feature WM fields and the scene WM fields. This final state after the memorization of the sample array is shown in Figure 5.7.

5.3.2 Parallel detection of feature changes

In the simplest form of change detection experiment, a single feature of one item in the sample array is changed to a novel feature value in the test array. An experiment of this kind was used for instance by Treisman and Zhang (2006). They used stimulus arrays consisting of three colored shapes or three colored letters, and in the change trials either the color or the shape/letter identity of one item in the test array was set to a feature value that had not been present in the sample array. According to the Feature Integration Theory, such changes in individual feature values can be detected in parallel, without attending to items sequentially. This is also implemented in the DNF model through the parallel projections to the contrast fields. Here, I will first review the role of the contrast fields during the memorization phase, which is directly related to the parallel change detection, and then demonstrate how the model can perform a feature change detection task.

When a new sample stimulus array is first presented for memorization, all of its feature values are novel, in the sense that they are not yet represented in working memory. This is reflected in the activation of the feature contrast fields, as can be seen in Figure 5.5. Due to the direct excitatory input from the visual sensory fields, three peaks form in each feature contrast field. (For the one item that is just being memorized, the peak is kept stable by input from the attention field, which will be discussed later.) As the memorization of items progresses, peaks in the feature contrast fields are suppressed by inhibitory input from the feature WM fields. This is visible in Figure 5.6. Here, two of the items have already been memorized successfully, and the corresponding peaks in the feature contrast fields have disappeared due to inhibition from the feature WM fields. This inhibition can be seen directly when the visual stimuli are removed after the scene is completely memorized (Figure 5.7). Here, without any excitation from the visual stimuli, each memorized feature value in the feature WM fields induces a distinct suppression in the activation profile of the corresponding feature contrast field.

This inhibition from the feature WM fields now also serves to distinguish between already memorized and novel feature values when the test array is presented. This is shown in Figure 5.8. The test array differs from the memorized sample array in the color of the leftmost item (orange instead of red). As soon as the test array is presented, peaks form in the visual sensory

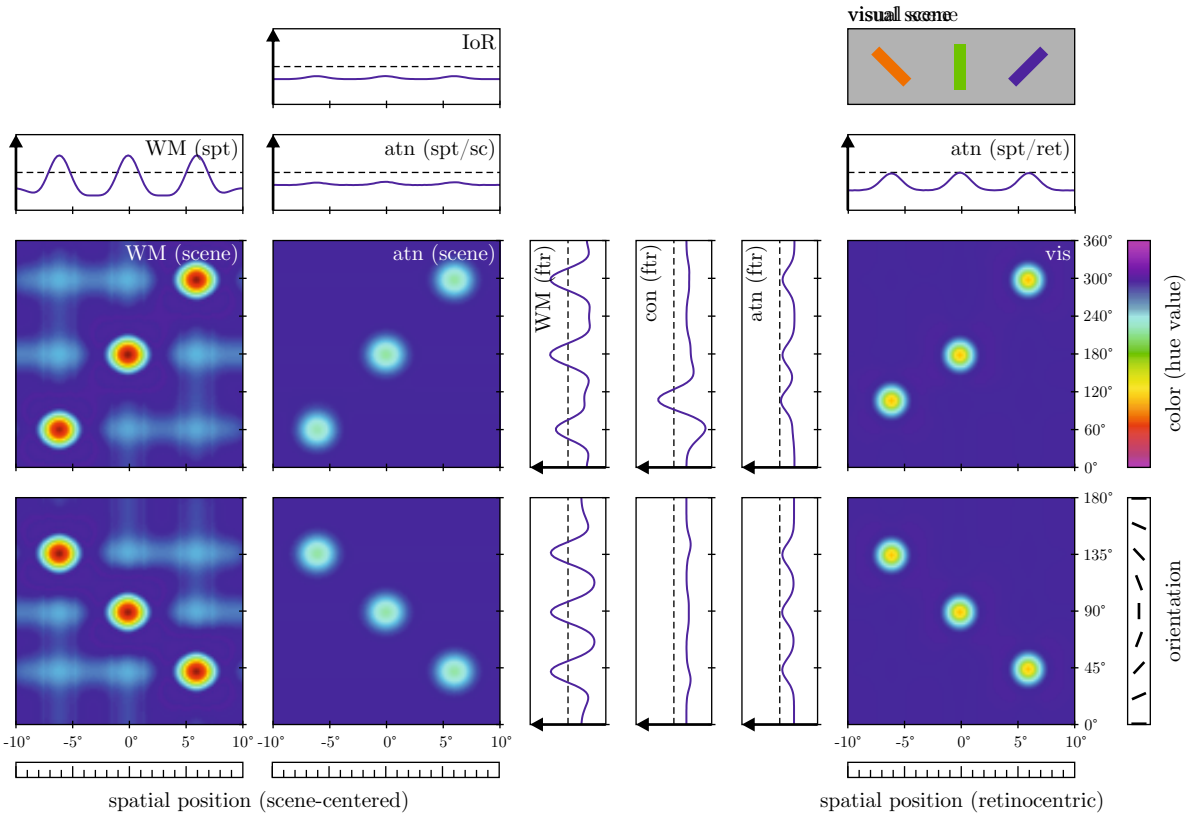


Figure 5.8: Detection of feature changes in the DNF model of scene representation. A visual scene with one novel feature value (orange instead of red bar) is presented to the model, and an activation peak forms in the color contrast field to indicate the change.

fields and project to the feature contrast fields. For the orientation dimension (bottom fields), all peak positions in the visual sensory field match those in the feature WM field. The excitatory and inhibitory inputs to the feature contrast field therefore cancel each other out, and the resulting activation profile is almost flat.

In the color dimension, however, there is a mismatch for the changed item. The corresponding peak in the visual sensory field for color (bottom left in Figure 5.8) provides input to the color contrast field at a position where it is not canceled out by inhibitory input from the feature WM field. Consequently, activation in the contrast field can rise to reach the output threshold, and a stabilized activation peak forms. This peak signals that there is a change in the test array. In some previous models, an additional set of nodes was defined to turn the presence or absence of a peak in the contrast field into a binary response (change or no change; Johnson et al., 2009a).

This response system is omitted here for simplicity, and the response that the system would give is simply deducted from the presence of the contrast field peak itself.

There are few notable properties of this change detection system. First, the detection of a feature change depends on the distance in feature space between the new stimulus and the memorized features. Since both excitatory and inhibitory inputs extend over a certain region in feature space, very small mismatches will not allow a peak to form in the contrast field. The reliability and speed of peak formation increases (up to a certain level) with distance in feature space between sensory and working memory peaks. Second, the detection of feature changes is carried out before an attentional selection of an individual item has taken place. The competition process for attentional selections begins about at the same time at which peaks for novel features form in the contrast fields, as both processes are driven by direct projections from the visual sensory fields. And third, the parallel change detection process is not sensitive to the locations of features or their conjunctions among each other. As long as the individual feature values that are present in the test array have also been present in the sample array, no change is detected in this process.

Finally, the detection of a novel feature also influences the further processing of a scene. The feature contrast fields project to the corresponding feature attention fields in an excitatory fashion. If a change is detected through a peak in one of the contrast field, this creates a bias to direct feature attention to this novel feature value, and thereby makes it likely that the corresponding item in the stimulus array is selected first as the focus of attention. This is consistent with the experimental results on attention capture by novel features mentioned above (Hyun et al., 2009a).

5.3.3 Change detection for space-feature binding

A second type of change detection task requires participants to detect whether the same features are still present at the same locations (Wheeler and Treisman, 2002; Johnson et al., 2008). In these experiments, a change in the test array is introduced by swapping the values for one feature dimension between two items in the sample array. The results of Wheeler and Treisman clearly indicate that additional processes are necessary to detect such changes in location binding, and performance is significantly decreased compared to a pure feature change detection (with novel feature values introduced in the test array) under the same conditions.

The DNF model can solve this task as well, without any changes to the architecture or the parameter values. The memorization of the sample scene is performed in the same way as before, yielding the scene representation as shown in Figure 5.7 as basis for the change detection. The test array is then presented, as shown in Figure 5.9. If no novel features are detected in

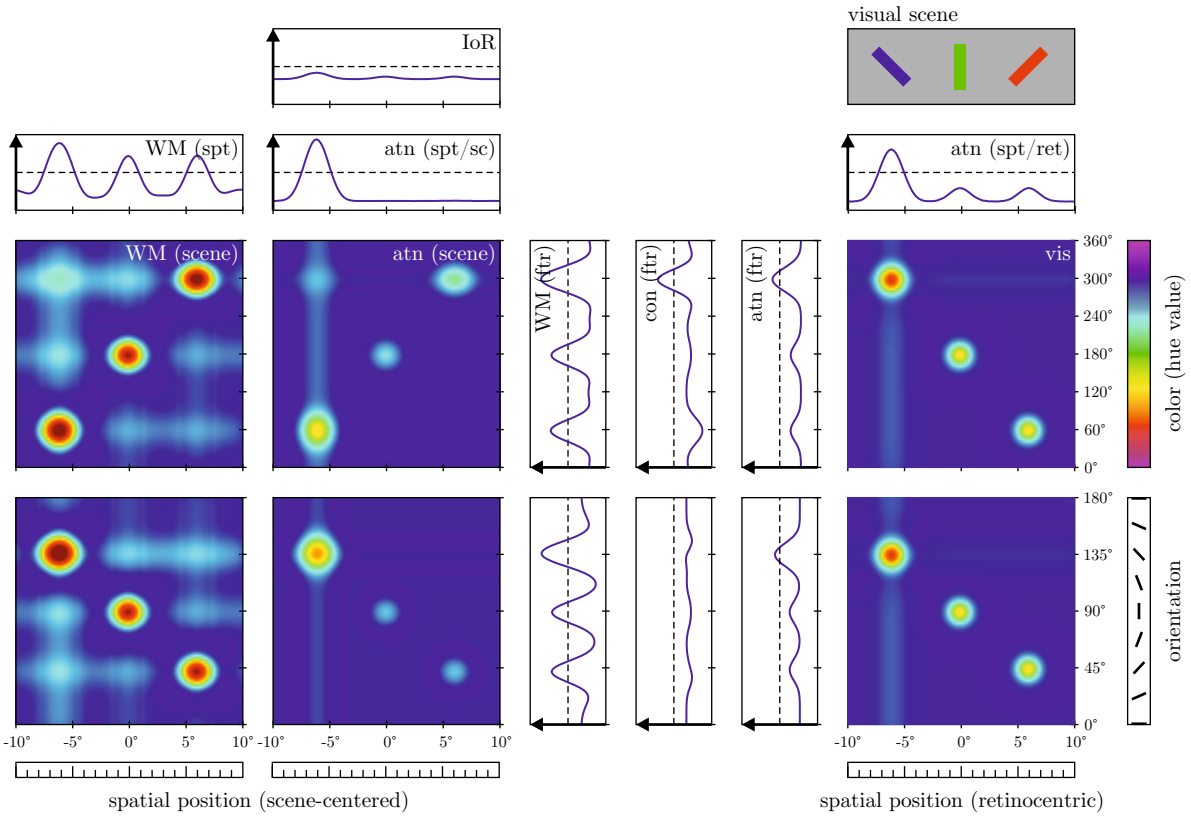


Figure 5.9: Detection of changes in space-feature binding. A test scene is presented in which two items have swapped their colors (the outer red and blue bars). When one of these items is selected by attention, the change in its color is detected through a peak forming in the color contrast field.

the scene by the parallel change detected mechanism, the model proceeds autonomously with a sequential inspection of visual items in the test array. This is driven by the competitive interactions within the attention fields and the mutual coupling between them, just as during the memorization phase.

In Figure 5.9, the leftmost item in the test array has been selected as the first focus of attention. Note that the color of this item has been swapped with the rightmost item compared to the memory sample array (Figure 5.5), while its orientation has remained unchanged. The one-dimensional attention fields that now reflect the surfaces features and the location of the selected item project ridge inputs into the scene attention fields. This projection is stronger along the spatial pathway, reflecting the special role of space in the model to bind surface feature dimensions together.

In the scene attention fields, there are already localized hills of activation induced by the peaks in the scene WM field. Since the items in the test

array occupy the same locations as the items in the sample array in this experimental paradigm, the vertical input ridge from the spatial attention field will always overlap with one of the localized hills of activation in each scene attention field. This combination of inputs induces activation peaks in the scene attention fields. Note that the horizontal ridge inputs from the feature attention field likewise overlap with hills of activation in the scene attention fields (assuming that the same features are present in sample and test array). In the case of a feature-location match between test and sample array, these inputs converge to select the same item from working memory (bottom scene attention field in Figure 5.9). If they do not converge, however, the stronger spatial input dominates in the selection of a working memory item, and the competitive lateral interactions in the scene attention fields ensure that only a single peak for this spatially matching item can form (top scene attention field in Figure 5.9).

The system has now accomplished a simultaneous selection of one item in the current visual scene and of the item at the same location from the working memory representation. This allows the direct comparison of the feature values for this single item in the feature contrast fields. The feature values of the selected visual item are represented by the one-dimensional feature attention fields, which project an excitatory input to their corresponding contrast fields. The feature values for the working memory item can be read out from the scene attention fields, which project an inhibitory input to the contrast fields.

If these feature values match, these inputs cancel each other out in the contrast field, and the activation pattern remains largely flat. This is the case for the orientation dimension in Figure 5.9 (bottom feature contrast field). If the feature values do not match, the excitatory input can induce a supra-threshold activation peak, as can be seen in the color dimension in Figure 5.9 (top feature contrast field). Note that the feature contrast fields are also still receiving the parallel inputs from the visual sensory fields and the feature WM fields, but these always cancel each other out in this experimental paradigm since no new feature values are introduced in the test array.

To complete the task of detecting any feature-location changes in the test array, the system must inspect every item in the present visual scene and test whether it matches the memorized item at the same location. This sequential processing of all items for the change detection is achieved by the same processes as during the sequential memorization. The presence of salient activation peaks in the scene attention fields serves as condition of satisfaction, indicating that the processing of the present item is completed. This is permissible here since by the time that sufficiently strong peaks have formed in the scene attention fields, the change detection process for the selected item is complete (meaning that if there is feature mismatch, a peak

in the contrast field will have formed by this time). The attentional selection of the current item is released by the inhibitory actions of the CoS node and the IoR field, and the IoR field retains a memory of the inspected locations. The next item is selected autonomously by the attentional processes, and the system continues until all items in the visual scene have been inspected. Only if no feature mismatch is found in any of the items, sample and test array can be considered to be the same (again, this final response generation is not captured in the model).

The reduced performance of human subjects in this task compared to the detection of simple feature changes can be explained by several factors in the model. First, the detection of feature location changes requires the formation and maintenance of working memory peaks in the scene WM fields, in addition to the peaks in the one-dimensional WM fields that are sufficient for pure feature change detection. This is consistent with recruitment of additional neural populations observed for the memorization of feature bindings in a recent fMRI study (Parra et al., 2014). Failure to form or maintain these peaks makes it impossible to detect changes later. In addition, the sequential process for feature-location changes is significantly more complex than the parallel process for detecting novel features, and requires the organization of many individual processing steps.

Interestingly, error patterns in different change detection experiments indicate that participants may inadvertently mix up the different types of change detection despite being given explicit instructions. The most common error in change detection for feature-location bindings in the study of Wheeler and Treisman (2002) was failure to detect a change (false negatives), indicating that participants often judged two arrays to be the same when they did not detect novel features, without completing an individual comparison of all items. Conversely, in a task where only pure feature changes should be reported, participants often judged arrays as different when the feature locations had changed (Treisman and Zhang, 2006). The authors concluded that "when a previously filled location is reactivated, retrieval of its previous contents (if they survive) is automatic."

These kinds of intermixing the different tasks appear consistent with the proposed DNF model. The system uses shared structures, namely the feature contrast fields, to detect both types of changes. Absence of a change signal in the parallel feature change detection process may be incorrectly judged as indicating an absence of feature-position changes when the sequential inspection of items is not pursued to the end. Conversely, the change detection for feature-position changes is performed automatically when attention is focused on a single item. This may happen inadvertently in pure feature change detection tasks, explaining the false positive responses in this condition. The experimental results thus support the model assumption that the different types of change detection reflect different modes of operation in a

shared system, with a smooth transition between them.

5.3.4 Change detection for feature conjunctions

The third type of change detection task that the model is capable of addresses the core question of how individual features are bound into object representations. In this type of task, participants are asked to determine whether the sample and test array contain the same objects, irrespective of their locations (Treisman and Zhang, 2006). For instance, if the sample array contained a green triangle and a red circle, the test array must also contain a green triangle and a red circle to be considered the same (rather than, e.g., a red triangle and a green circle). The locations of items in the test array are typically scrambled, either by switching locations between items or by using novel locations for all items. The locations should be irrelevant for the same/different judgment, although the shared locations of features that belong to a single item naturally still define the feature conjunctions within each stimulus array.

Compared to the feature-location change detection, this task is much more open in terms of the strategy to be used in the comparison process. One possible approach would be to sequentially select each item in the test array, and then in turn go through all the items in working memory to see if one of them matches. Obviously, this is not very efficient, and would require additional control structures to organize the nested sequences of comparisons. More efficient approaches can be implemented by using the parallel processing capabilities of the system where possible, namely in the selection of a candidate item from working memory for the comparison with the currently selected visual stimulus. One way to do this would be to select an item from working memory based on one feature of the attended visual stimulus (e.g., color), then check whether the other feature (orientation) matches. This may not work, however, when two or more items share a feature value in one surface feature dimension.

The strategy I propose here is to select a candidate item for the comparison based on *all* features of the currently selected visual stimulus. This approach builds on the assumption that if there is an item in working memory that exactly matches the selected stimulus, it will be selected as candidate for the comparison; otherwise, a partially matching item will be selected. Thus, if the comparison to this candidate yields a mismatch in one feature, it can be concluded that no working memory item matches the selected stimulus. The global response can then be determined by the same simple rule as used in the feature-location change detection: If the contrast field indicates a change for any attended item, the test array is judged as 'changed', otherwise it is 'same'.

To implement this strategy (and, in fact, any of the strategies mentioned here), it is necessary that retinal representations and the working memory

scene representation are spatially de-coupled. The coupling of the spatial attention fields via the reference frame transformation ensures that corresponding locations are always selected in the two reference frames, but for the present task, it is necessary that items at *different* locations can be selected for comparison (reflecting the task instruction that changes in item location between sample and test array are irrelevant for the response). The de-coupling is achieved by tuning down the connection strength between retinal and scene spatial attention fields, as well as the connections strengths for most other projections that are mediated by the reference frame transformation. This may be achieved in a neural system by modulated synaptic connections. To compensate for the resulting loss of input, the resting levels of the scene spatial attention field and the scene attention fields are increased.

Note that the reference frame transformation is still employed in this mode for the IoR field, to support an efficient sequential processing of the perceptual items even when gaze changes are made. It is known from experimental observations that the inhibition-of-return effect acts on attentional selection in a gaze-invariant, rather than a retinal reference frame (Posner and Cohen, 1984), so it does require some form of reference frame transformation. It is less clear, however, how the inhibition-of-return system is related to scene working memory, and whether these systems use shared resources. If there were in fact a separate system to achieve gaze invariance for the inhibition-of-return effect, then the spatial decoupling in the model for this task could be achieved simply by suppressing the gaze signal in the reference frame transformation, without requiring modulation of synaptic connection strengths.

Each trial of the feature conjunction change detection task starts with the formation of a working memory representation for the sample array, in the same way as before. During this period of the trial, the spatial coupling between retinal and scene-level representations is still intact, but is then tuned down when the test array is presented. The events after the presentation of the test array for a 'same' trial are shown in Figures 5.10 and 5.11. Note that the positions of the outer items in this test array are switched compared to the sample array in Figure 5.5, but all feature conjunctions are maintained.

One item in the test array is autonomously selected by the attentional mechanism at the retinal level (the rightmost item in Figure 5.10), yielding its location in the retinal spatial attention field and its surface feature values in the feature attention fields. Due to the spatial de-coupling, activation is now only propagated along the feature pathway. The feature attention fields project ridge inputs to the scene attention fields. Each of these ridges overlaps with one localized hill of activation from the scene WM field (since the feature values present in the test array are always the same as in the sample array for this task). Peaks form from these activation hills in the

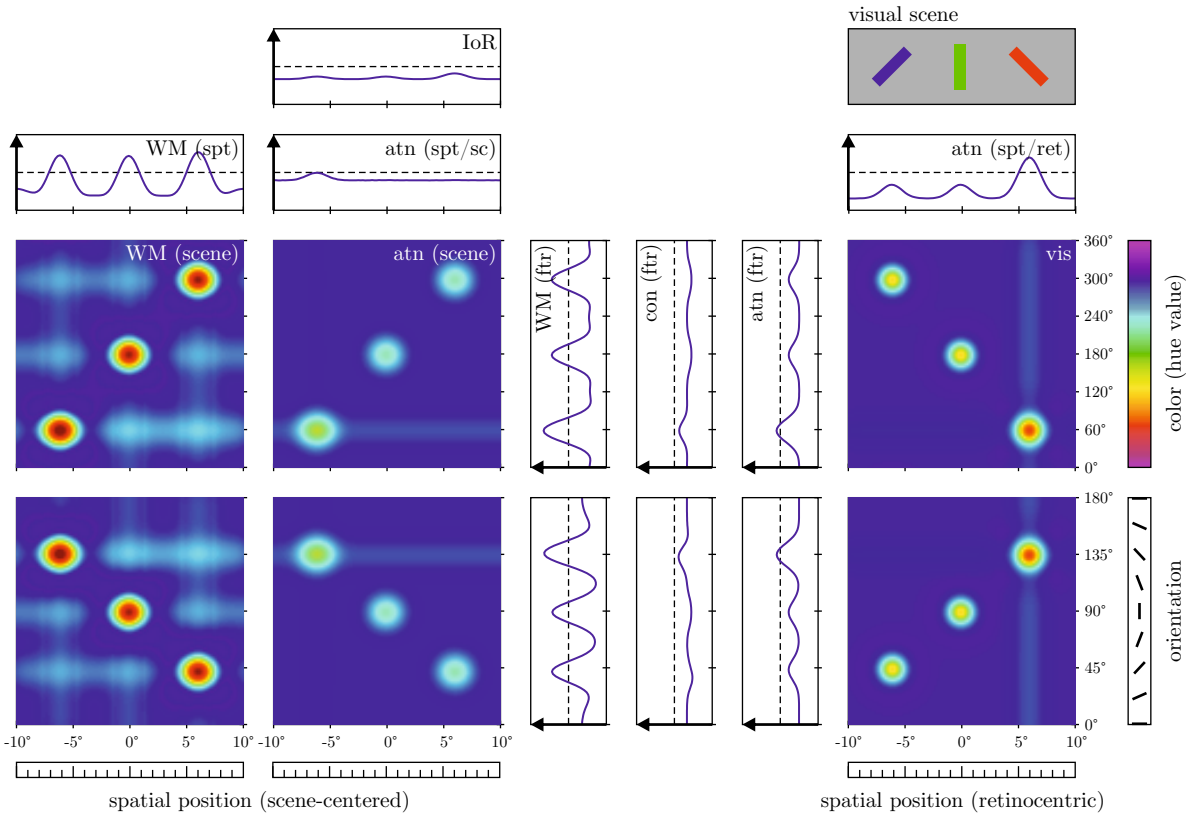


Figure 5.10: Detection of feature conjunction changes, early phase in the processing of one perceptual item in a 'same' trial. The rightmost item has been selected by attentional processes at the perceptual level, and the best matching working memory item is being selected in the scene attention field.

scene attention fields, and project input to the spatial attention field in the scene reference frame.

In the case depicted in Figure 5.10, an exact match for the retinally selected stimulus exists in working memory. Consequently, the peaks that begin to form in the scene attention fields are spatially aligned, since the activated feature values belong to a single item in working memory. The peaks in the scene attention fields both project input to the same location in the spatial attention field at the scene level, and induce a single activation peak here. Note that the position of this peak is different from the peak position in the retinal spatial attention field, reflecting the fact that the item's location has changed between sample and test array. The spatial attention field at the scene level projects a vertical input ridge back into the scene attention fields and thus stabilizes the selection of the matching item from working memory (Figure 5.11).

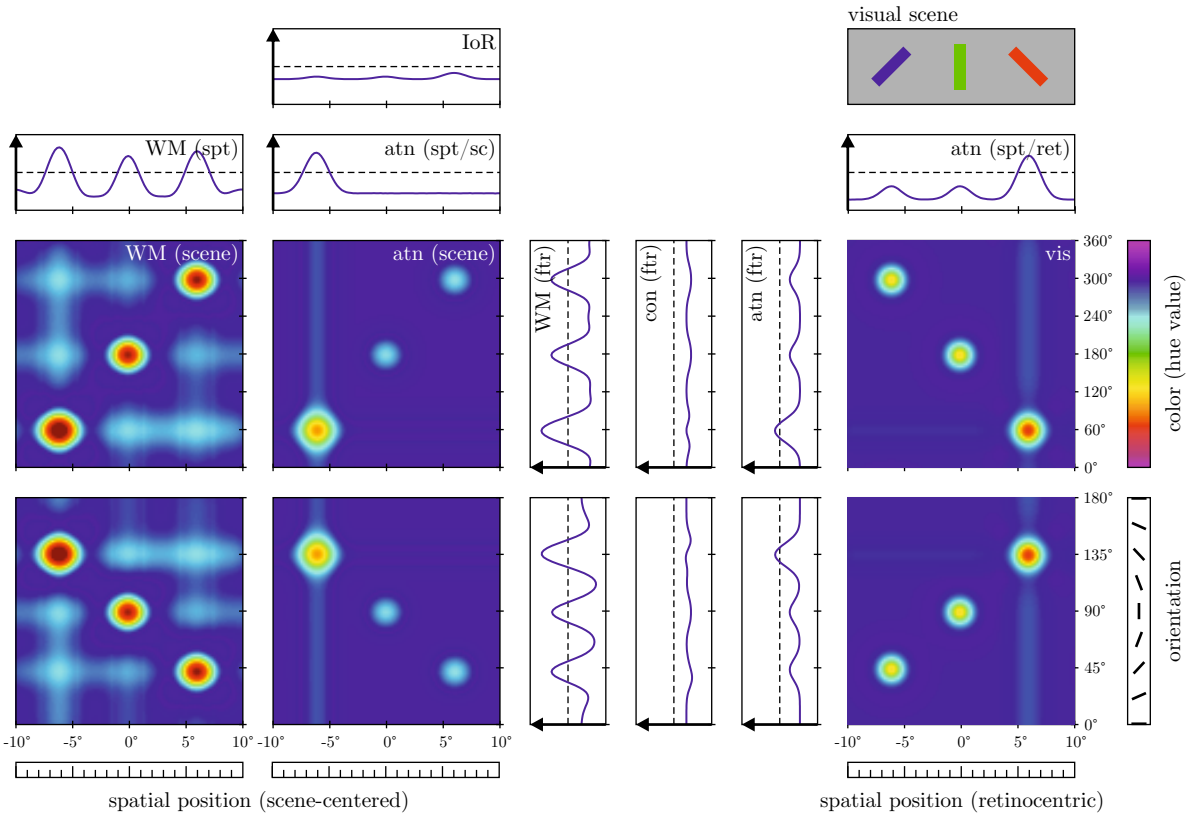


Figure 5.11: Detection of feature conjunction changes, late phase in the processing of one perceptual item in a 'same' trial. A perceptual item and a comparison candidate from working memory have been selected, and the comparison yields no differences between them.

With an item selected in the scene attention field that matches all features of the attended visual stimulus, no change signal is generated by the feature contrast fields. Although local activation levels in these fields briefly rise when the visual stimulus is first selected at the retinal level (Figure 5.10), inhibition from the (initially rather weak) activation peaks forming in the scene attention fields are sufficient to keep activation levels below the output threshold. When the selection at the scene level is completed, the activation profile in the feature contrast fields is again almost flat, since excitatory and inhibitory inputs cancel each other out. The salient peaks in the scene attention fields are also the signal that the comparison process for the currently selected visual item is complete, and triggers the release of attention and transition to the next item.

How does this system work, however, when there is no exactly matching item in working memory? This case is shown in Figures 5.12 and 5.13. Here,

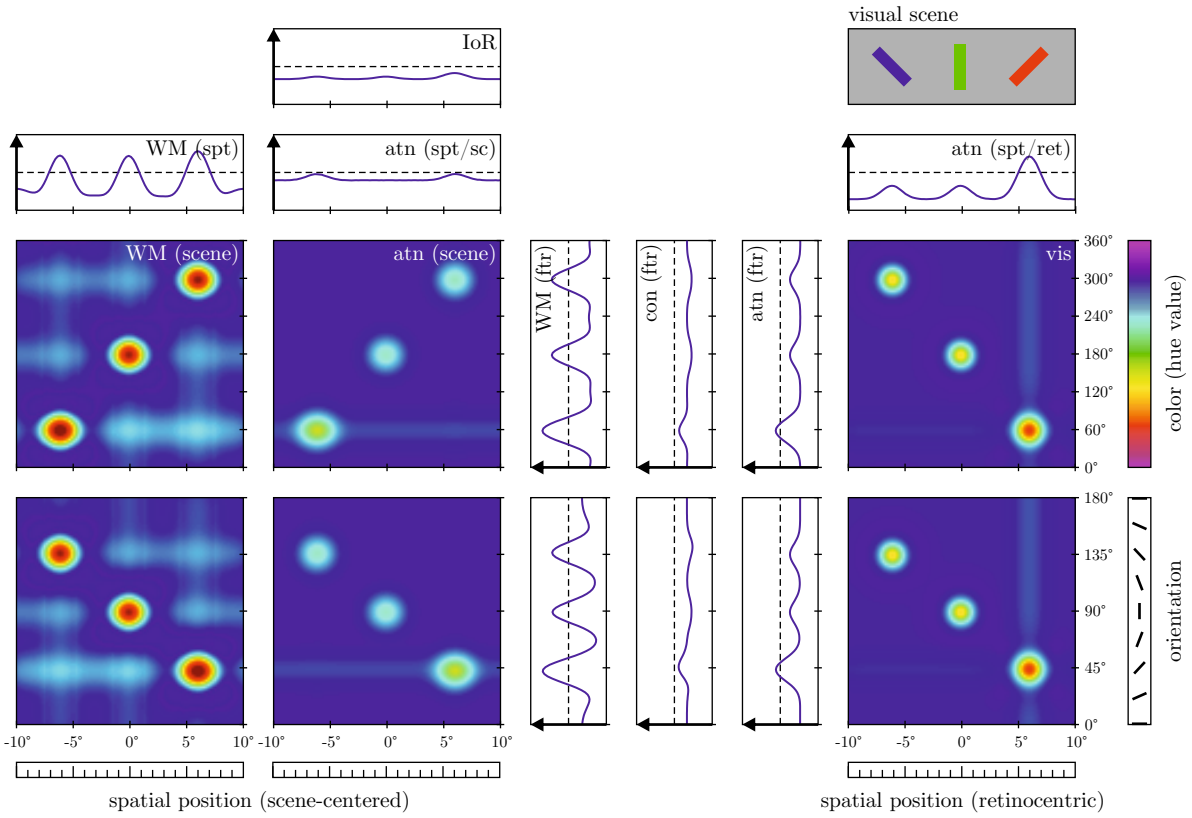


Figure 5.12: Detection of feature conjunction changes, early phase in the processing of one perceptual item in a 'different' trial. The rightmost item has been selected by attentional processes at the perceptual level, and based on its surface features, a comparison candidate is being selected in the scene attention fields.

the colors of the two outer items have been swapped between test array and sample array (Figure 5.5), but the orientations remained the same. Thus, the feature conjunctions for these two items are changed.

Again, the rightmost item in the visual scene is selected first as the focus of attention. The feature attention fields form peaks to reflect the surface features of this visual item, and project ridge inputs into the scene attention fields. Weak activation peaks form in these fields at the points where the ridges overlap with localized input from the scene WM fields (Figures 5.12). Critically, the peaks in the two scene attention fields are not spatially aligned in this case, since the surface features that characterize the currently attended visual item belong to two different items in the working memory representation.

The scene attention fields consequently provide input to two different

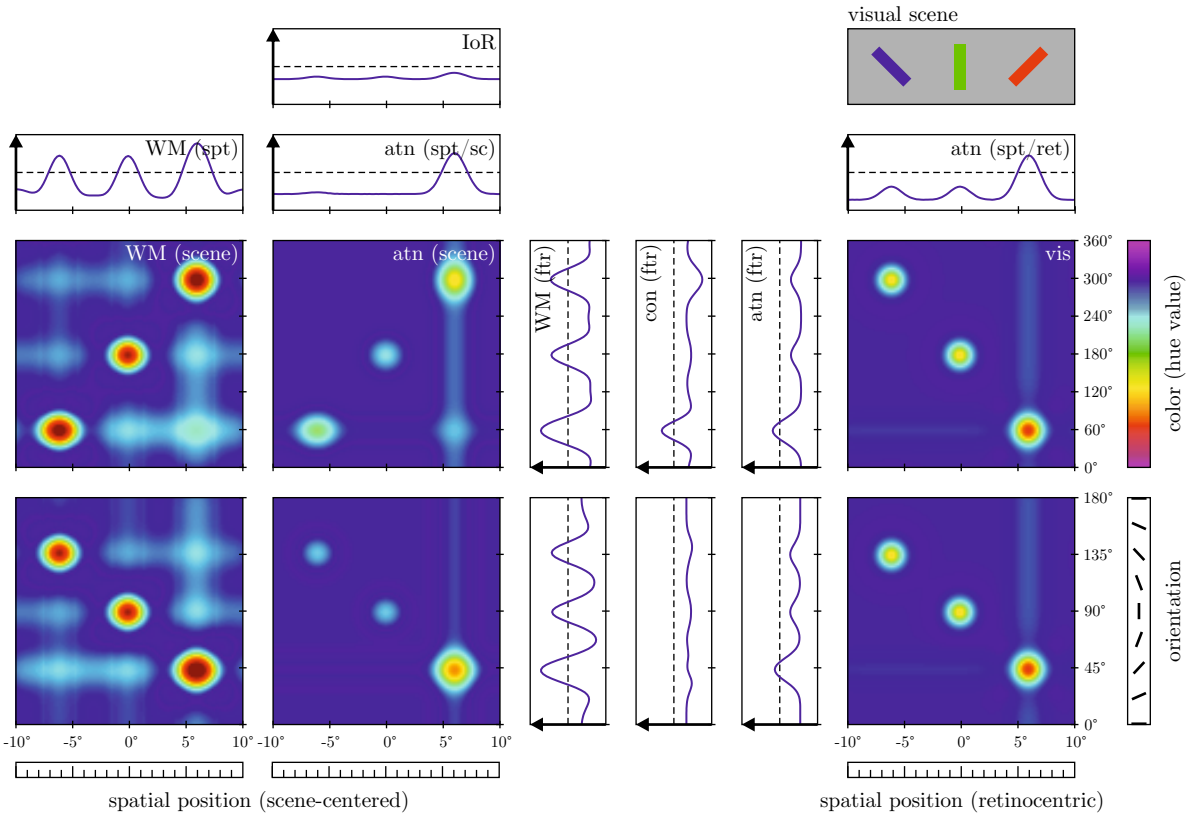


Figure 5.13: Detection of feature conjunction changes, late phase in the processing of one perceptual item in a 'different' trial. A single working memory item has been selected as comparison candidate in the spatially coupled scene attention fields, even though non of the memorized items matches the presently selected perceptual item exactly. The color contrast field indicates the mismatch between the two selected items.

locations in the spatial attention field at the scene level. The competitive lateral interactions in the spatial attention field generate a selection decision, and only a single peak forms from the two inputs (Figure 5.13). The decision which one is selected is largely random, since the inputs are generally equally strong. Here, a peak forms on the right, based on the feature match in the orientation dimension and a remaining weak input from the peak in the retinal spatial attention field at the same location (the latter reflects a weak bias to compare the visual item to the working memory item at the same location, even though the two reference frames are largely de-coupled now).

The peak in the spatial attention field then projects back to the two scene attention fields. In the orientation dimension, it simply strengthens the peak that already exists at this location. In the scene attention field for

color, however, this new spatial input overrides the original selection of a working memory item, and induces a new activation peak that suppresses the old one. In effect, the coupling via space ensures that a single memorized item is selected in both scene attention fields, rather than the features of two different items.

With this consistent selection of a single item at the scene level, the mismatch to the attended visual item can be detected. The new peak in the scene attention field for color does no longer match the selected color in the corresponding feature attention field. Consequently, excitatory and inhibitory inputs to the feature contrast field for color do not cancel each other out, and an activation peak forms in this field to signal the mismatch (Figure 5.13). It is then also clear that no other working memory item would provide a better match for the attended visual stimulus, since the selection in the scene attention fields was already performed based on feature match. Thus, any peak forming in the feature contrast fields unambiguously signals a change of feature conjunctions between sample and test array. Only if all visual items are inspected without such a change signal, the feature conjunctions in the two arrays are the same.

5.4 Discussion

In this chapter, I have presented a DNF architecture for the representation of visual scenes in working memory, and have described the autonomous processes that evolve in this architecture to achieve the memorization of a scene and to solve different classes of change detection tasks. At the core of this theoretical work lies the problem of feature binding. Behavioral evidence shows that humans cannot form memory representations of bound object features in parallel, and instead employ a sequential strategy in which each object has to individually selected by focused attention. The extant models of this process are largely conceptual in nature, and employ language from computer science to describe the underlying representations. Working memory representations of bound features are conceptualized as object files, and these object files and the locations of objects in the world are referenced by a form of pointer. The neural implementation of such conceptual models has been left entirely unaddressed.

The DNF model achieves this implementation in a fully neural architecture, and helps to clarify several aspects of humans scene working memory. In particular, it illustrates where and how the problem of feature binding arises, namely in the recombination of feature and location information from separate processing streams. The recombination is realized in the DNF model as intersection of ridge inputs into a multi-dimensional neural field, and the simultaneous application of multiple ridges in each dimension would lead to spurious intersections that do not match the actual feature con-

junctions of the perceived objects. Moreover, the DNF model explains why separate processing pathways are necessary in the first place: The operations applied on the spatial representations, namely the reference frame transformation, would be highly inefficient when applied to a combined space-feature representation. Analogous reasonings likely hold for the feature pathway, although these are not explicitly explored in the model.

The core binding mechanism in the model's working memory representations is binding via space. Separate feature maps over a shared spatial dimension exist for the different surface feature dimensions. This reflects a natural criterion for feature binding, namely that two features are likely to belong to the same object if they are perceived at the same location. This approach is also biologically plausible since it is directly analogous to the organization of feature maps in early visual cortex. The special role of space in feature binding is furthermore supported by results of behavioral studies, both at the perceptual level (Nissen, 1985) and for working memory (Pertzov and Husain, 2013, , discussed in detail below).

The spatial representation, coupled over different reference frames, fills the role of the pointer in the Object File Theory, and of the visual index in the conceptual theory of Pylyshyn (1989, 2001). Through mutual projections in the DNF model, it can act in two directions: On the one hand, activating a spatial location can bring up the object features at that location from working memory. This was used explicitly in the model demonstrations of the change detection task for feature locations. On the other hand, the spatial memory can also be used to highlight the objects location in the retinal image, via the reference frame transformation, and the memorized location of a specific object may also be queued by a surface feature of this object. This mutual coupling of object features to locations is clearly useful if one assumes that most objects in the world are stationary most of the time, and it is even more useful if moving object can be tracked to some extent, as can be done in the DNF model.

So on the one hand, the spatial representation in the model directly reflects location as a concrete feature of visual objects. But it is at the same time used to bind object features together when the location is irrelevant. This is most clearly demonstrated in the change detection task for feature conjunctions. Here, the locations of objects are not task relevant, and due to the de-coupling between retinal and scene reference frames, the memorized locations lose their link to any locations in the current scene. But still, the shared location of different features in working memory is what indicates that they belong to the same object. This is explicitly used in the change detection process for this task, since the coupling via space and competition in the spatial dimension is what ensures the selection of a single, coherent object from working memory for the comparison operation. Notably, this more abstract use of space for feature binding, even when location is irrele-

vant, emerges directly from the concrete use of representing the locations of features and objects.

The binding mechanism has major impacts on the processing of visual scenes in the model. Most importantly, it requires that each item is selected individually in the coupled attention fields to memorize its features in a bound form, and to compare these bound features to new items during change detection. This is consistent with the central claim of the Feature Integration Theory, derived from human behavioral data in various experimental paradigms. The DNF model implements an autonomous mechanism for the sequential processing of visual items, based on previous DNF models of sequential order and behavior generation (Sandamirskaya and Schöner, 2010; Richter et al., 2012).

While these previous models addressed the learning of arbitrary sequences of elementary actions, the form implemented in the model for scene representation constitutes a special-purpose system specialized for the sequential processing of visual items. It does not explicitly control every step in the inspection of each item, like the selection in each attention field and the peak formation in each working memory field. Instead, it relies on certain assumptions about the sequence of events and the relative timing of projections along different routes. Only one event is explicitly detected: the formation of a peak in the scene attention field, which indirectly indicates that working memory peaks have been created or that a comparison operation has been performed. This gives the mechanism a great deal of flexibility, for instance allowing the attentional selection to be biased toward certain items by cognitive input to the attention fields without requiring any adjustments in the sequencing mechanism. The use of a condition of satisfaction to trigger the transition to the next step also allows for significant differences in the duration of each processing step, for instance due to variability in the time it takes to resolve the competition for attentional selection. This is the key feature adopted from the sequential order models.

I consider the use of this special-purpose mechanism to be appropriate for the task of sequentially processing visual items, since the scanning of a visual scene constitutes a basic behavior that is frequently employed. The fact that errors in change detection tasks are often consistent with unintentional task switching (such as reporting feature location changes in a pure feature change detection task, as described earlier) also indicates that the sequential processing is done in a relatively automatic fashion, without explicit cognitive control in each step. This does not rule out that a more detailed cognitive control may be employed as an alternative strategy under certain circumstances, for instance when task requirements are more complex and time pressure is low.

5.4.1 Alternative approaches to scene working memory and feature binding

Forming internal representations of visual scenes for the planning of goal-directed actions is a highly relevant task for robotics. A robotic version of the DNF model presented here has been shown to work with real-world camera input and is capable of answering queries about visual scenes and guide grasping movements toward objects (Zibner et al., 2011b; Knips et al., 2014). A large number of algorithmic approaches have been developed for the same tasks. Many of these make use of a sequential processing of visual scenes through a series of fixations (with a moving camera system) or attentional selections, an approach often running under the label of *active vision* (Aloimonos et al., 1988; Rasolzadeh et al., 2010). This sequential processing serves both to simplify the task by focusing the scene analysis on a limited set of behaviorally relevant locations, and it provides additional information for the reconstruction of three-dimensional scenes when the images from different fixations are combined (Mishra et al., 2009). Naturally, these algorithmic approaches do not face the same challenges of feature binding in neural systems as the DNF model.

Among explicitly neural models, the DNF architecture presented here is to my knowledge the first work that provides a complete account for the problems of multi-item scene working memory and change detection. Other neurodynamic models have dealt with related, but more limited problems, in particular with visual search. Fix et al. (2011) have presented a model of visual attention that is capable of sequentially searching a visual scene (with or without gaze changes) and that keeps a memory of inspected locations in a gaze-invariant reference frame. The theoretical work of Hamker (2005b; 2006; discussed in the previous chapter) likewise addressed visual search and deployment of visual attention. These models are generally compatible with the present architecture, although they differ in implementation details, but neither of them contains structures to address change detection or feature binding in working memory.

One influential, but also controversial explanation for feature binding in neural systems is based on synchrony of neural firing. This explanation was popularized through the theoretical work of von der Marlsburg (reviewed in Von der Malsburg, 1999) and supported by experimental work of Singer and others (reviewed in Singer, 2001). The theory starts from the problem that different properties of visual objects are represented by separate neural populations, and proposes synchronization of neural firing between these neural populations as a means to signal that the represented features belong the same object. Whereas the DNF model only captures neural activity in the form of a mean firing rate, the synchrony-based approaches assume that the exact timing of individual action potentials matters.

Some models also explicitly employ neural synchrony as a means of fea-

ture binding in working memory (e.g. Shastri, 1999). The synchronization is assumed here to take place within an oscillatory cycle, often equated with neural oscillations in the gamma range. Within each cycle, a certain time window would be assigned to each object, and the simultaneous neural firing (distributed over different populations of neurons) within this time would reflect the features of that one object. In this way, a single neuron can contribute to representing the features of two objects (e.g., if these share the same color), but each action potential from that neuron can be clearly assigned to representing one object. This means that no binding dimension (like space in the present model) would be needed, and a single neural representation over each individual feature space would be sufficient to represent the bound features of multiple objects. Capacity limits for the number of represented objects arise in this framework from the limited number of distinct time slots within each oscillatory cycle.

While this approach appears to require few neural resources, it makes very strong assumptions about special properties of the neural populations involved, namely that they can create, maintain, and transmit a pattern of synchronous firing, likely across different cortical areas. Both the theoretical approach and the experimental support for it have been called into question (Shadlen and Movshon, 1999). Many empirical findings show that correlation between the firing times of different neurons contribute little additional information beyond what is contained in the firing rates, and the functional relevance of this correlation is dubious (Golledge et al., 2003; Palanca and DeAngelis, 2005). And while the advantage of signaling the binding through synchrony is obvious in theoretical models, it is often not explained how this synchrony is created in the first place (unless it is already contained in the input to the system).

The present DNF approach demonstrates that such assumptions of additional representational powers of neural populations are not necessary to account for human performance in scene representation and change detection. The sequential processing of objects and the use of space as a binding dimension are sufficient to account for human capabilities, and are consistent with the limitations in these capabilities that can be observed in different experimental paradigms. Note that the sequential attention to objects can be viewed as a form of macroscopic synchrony—during the time window that an object is attended, the activity of certain populations reflect selectively the properties of that object—but this synchrony can only be used to establish the feature binding, not to maintain in it working memory after attention is shifted to another object.

There is yet another alternative to the use of space as binding dimension. In several studies, it has been suggested that the properties of visual working memory are best explained by a fixed number of slots for integrated object representations (Luck and Vogel, 1997; Zhang and Luck, 2008; but see Bays

et al., 2009 for an opposing view). This view has only been expressed in verbal theories, but it is relatively straightforward to imagine how it could be implemented in a neural architecture. For each feature dimension (both surface features and location), there would be a fixed number of copies of the neural representation of that feature, corresponding to the proposed number of slots (typically around three to five, to match estimates of human working memory capacity). Each of these copies would be capable of holding one feature value, and corresponding slots would be coupled across different feature dimensions to provide a bound representation of one object.

Replacing the spatial dimension by this abstract and discrete "slot dimension" would reduce the required neural resources. It is hard, however, to find any direct evidence either supporting or refuting the existence of such a mechanism using behavioral or neurophysiological methods, since the slot assigned to an item would not be linked to any observable features of the visual object nor to any overt behavior. To some degree, this abstractness of the slot dimension may in itself serve as an argument against this mechanism—in particular from an embodied cognition perspective, which views cognitive capabilities as emergent from sensori-motor processes.

More importantly, there is support from psychophysical experiments for a special role of space for feature binding in working memory, which appears inconsistent with this slot mechanism. Pertzov and Husain (2013) found that memory performance for sequentially presented stimuli (colored oriented bars) is significantly impaired when all stimuli are presented at the same location compared to a condition with a separate location for each stimulus. No analogous impairment was found when the stimuli matched in surface features (color or orientation). Moreover, the memory impairment was specific to the binding of features (characterized by an increased proportion of mis-bindings) and did not affect memory accuracy for the individual feature values. This is consistent with the DNF model, where object location mediates binding of surfaces features, but unbound surface features can still be memorized independent of their location. It cannot readily be explained by a slot mechanism, where object location would be expected to be treated in the same way as surface features, while binding is achieved only via the slot dimension.

It should also be noted that the mechanism in the DNF model—with working memory peaks coupled through space—can by itself account for several experimental observations that have been viewed as support for the existence of discrete working memory slots (Zhang and Luck, 2008). In particular, it has been observed that memory for one visual object appears to be formed in an all-or-non fashion, and memory capacity cannot not be re-distributed to memorize a larger number of items but with less precision. This is consistent with the formation of individual self-sustained activation peaks in DNFs, whose number is limited by the increasing lateral inhibition

produced by additional peaks.

5.4.2 Open issues

One significant limitation in the present model is the greatly simplified feature pathway. Unlike in the brain (and in many models aimed at object recognition and related tasks), the complexity of the represented features does not increase along the pathway. Instead, the elementary features of color and orientation are retained in the same format throughout all feature fields. I consider this an acceptable simplification in the present approach to explain the general mechanisms for creating a scene representation in working memory, focusing on the interplay between feature and spatial pathways and not on the detailed processing within each pathway. It is also sufficient to model the key conditions in change detection experiments, where artificial stimuli composed of simple features are used. The findings from these experiments also confirm the basic assumptions underlying the feature representations in the DNF architecture, namely that elementary features like color and orientation are memorized, and that they are memorized in separate representations, requiring specific mechanisms to bind the features of one object together.

There are several questions, however, that the model does not address due its limitation to elementary features. It does not explain what types of feature representations are necessary to memorize natural objects, how complex these representations have to be, and how they interact with each other. The question of feature complexity is relevant in particular because the proposed mechanism requires feature maps over space (simplified to one-dimensional space in the implementation presented here, but at least two-dimensional space in the biological system). If the feature representations themselves are very complex and require a large amount of neural resources, such feature maps may become unfeasible.

The architecture offers a potential way around this problem due to the fact that it still contains the separate working memory representations for features only, which do not cover the spatial dimension. The role of the scene WM fields, implemented as feature maps over space, is only to provide the binding of features to space (and, indirectly, to each other). These fields do not necessarily have to reflect the full details of the feature representations, and may potentially be coarse in the spatial representation as well, given that a separate purely spatial working memory representation also still exists. Such a coarser spatial representation may, in fact, approximate the abstract "slot dimension" discussed above. The question remains whether a coarser representation in the scene WM fields can still provide effective binding, or if it would automatically be prone to misbindings between similar feature values. But it would certainly provide a way to greatly reduce the required neural resources.

Another open question in the present architecture is how objects can be memorized that occupy the same spatial location. This can be the case when objects are presented sequentially, as in the study of Pertzov and Husain (2013). While this study found a significant impairment of memory performance in this case, it still indicates an ability to memorize more than one item per location. This is not possible in the model as described so far: Multiple working memory peaks for different features cannot be induced at the same location in the scene WM fields, and even if they could, there would be no way to determine which of these features belong together to one object. One possible way around this problem is to assume that long term memory is invoked in such instances, which works in a different fashion and is not constrained by the same limitations as working memory.

But the issue may also be solved within the present framework if the spatial dimension in the scene reference frame can be used very flexibly, and in particular is not linked to allocentric space in a fixed fashion. A certain amount of flexibility in how the scene space is defined is clear from behavioral results: Shifting a whole stimulus array has little influence on change detection performance, unlike randomly scrambling the array (Hollingworth, 2007). Such a behavior is quite straightforward to achieve in the DNF architecture as well, if the gaze signal in the reference frame transformation does not directly depend on actual gaze alone, but can be adjusted to compensate for frame shifts. The required shift value can be determined from a spatial alignment process as described in the previous chapter.

A similar mechanism could be used to map the sequentially presented objects onto different locations in the scene representation. Of course, this requires even greater flexibility in the adjustment of the spatial reference frames, and additional cognitive control in order to shift the reference frame at the right times without any external signal to drive this shift. The observed decrease in memory performance in this condition is consistent with such a more complex process, which would be more prone to errors. Moreover, such mapping from temporal order to spatial locations does not appear implausible based on findings from other fields: In language, in particular, spatial metaphors are used ubiquitously for temporal relations, and it has been proposed that analogies to space form the basis not only for talking about time, but for temporal reasoning in general (Boroditsky, 2000; Gentner, 2001; Casasanto and Boroditsky, 2008). With such generalizations, the model of scene representation might provide the basis not only for memory representations of what is where, but for reasoning in both concrete and abstract spaces.

Chapter 6

Modeling Spatial Language Behavior

6.1 Introduction

The topic of this chapter is relational spatial language. Expressions like "the keys are to the right of the monitor" offer a highly flexible way to communicate about object locations, and are used frequently in everyday language. Such an expression uses one object in the world—ideally one that is salient and easily localized, or whose position is known—to establish a spatial reference frame, and then uses this frame to describe the location of another object. It is the freedom to establish new reference frames that makes relational spatial descriptions so flexible.

The motivation to address spatial language behavior in neurodynamic models is two-fold. On the one hand, spatial language has a practical use in the field of human-robot interaction. It can be used to give instructions to a robot, and provides a natural means to specify an object in a visual scene as the target of an action or a movement. Conversely, robots endowed with the ability to generate relational spatial expressions can describe visual scenes and answer questions about locations in a way that is easy to understand for humans. Existing approaches in robotics employ algorithmic methods to determine object locations in space, assess their spatial relationship, and map this onto a verbal spatial description (Stopp et al., 1994; Skubic et al., 2004).

On the other hand, the field of spatial language has also attracted considerable attention in psychology and cognitive science. The reason is that it offers a test case to investigate the link between the abstract, discrete and symbolic representation in language and the metric sensory representations of the visual world (Regier and Carlson, 2001). This is also the core issue to be addressed by the DNF model described in this chapter: How are links established between certain elements of a verbal phrase and aspects of a visual

scene? I will refer to this as the *grounding* of the verbal description in the scene (Roy, 2005). This grounding is the foundation to determine truth values for verbal statements about a concrete visual scene, to complete partial statements and answer questions, and to generate behavior based on verbal descriptions.

The individual steps necessary for apprehending a spatial relation have been analyzed by Logan and Sadler (1996; see also Logan, 1994, 1995). A spatial phrase of the type "The keys are to the right of the monitor" consists of three elements: The target object ("the keys") whose location is described by the phrase, the reference object ("the monitor") that serves to anchor the spatial description, and a spatial relation ("to the right"). To resolve such a phrase, Logan and Sadler propose the following steps: (1) spatial indexing, in which the target and reference descriptors in the verbal phrase are bound to objects in a perceptual representation; (2) adjustment of the reference frame to center it on the reference object; (3) alignment of a spatial template (e.g., for the term "to the right") with this reference frame; and (4) the assessment of the match between that aligned spatial template and the location of the target object. These steps can be assembled in different combinations and orders to solve different tasks and answer questions, for example of the type "Where is the pencil relative to the cup?" or "What is to the right of the plate?"

Notably, in some instances the processing steps proposed by Logan and Sadler do not appear to be applied in a strictly sequential fashion. In a relatively open-ended task, Carlson and Hill (2008) asked participants to provide a description of an object location in a visual scene, and subjects had to choose both a spatial term and an appropriate reference object. The results indicated that both the visual saliency of potential reference objects and the goodness of fit to a relational spatial term influenced the choice of the reference object. Thus, participants could neither have chosen the reference object first and then selected an appropriate spatial term (because then the fit of the spatial term would not have influenced reference object selection), nor could they have selected the spatial term first (because then they would have no certainty that a potential reference object, let alone a salient one, was available to go with that term). This indicates that reference object and spatial term are selected in parallel to some extent, as will be discussed in detail later in this chapter.

However, there is also a specific reason why certain processing steps in these tasks should be executed in a sequential fashion. A binding problem—specifically, the problem of variable binding—occurs in the grounding of spatial language, analogous to the problem of feature binding that necessitated sequential processing in the scene representation model. To ground a relational spatial phrase in a visual scene, the locations and identities of two objects must be bound to their specific semantic roles of target and referent

in the phrase. They are not interchangeable—the statement "the key is to the right of the monitor" is not the same as "the monitor is to the right of the key." The solution proposed here follows the same principle as in the scene representation model: Different visual items are focused sequentially by spatial and feature attention, and activation peaks in spatial representations serve as a form of pointer to track different visual items and their semantic roles.

In the following, I will present a DNF model of spatial language behaviors that captures the underlying neural processes, roughly following the processing steps proposed by Logan and Sadler (1996). The model in particular aims to emulate the flexibility of human spatial language use by solving different tasks in a single unified architecture. The double aim of flexibility and neural realism contrasts with previous models of spatial language. The models proposed in the psychological literature primarily aim to capture behavioral signatures in specific spatial language tasks as a function of stimulus properties. While some of them are inspired by neural principles, such as the AVS model of Regier and Carlson (2001), they do not describe the actual neural processes that produce behavior. Existing neural network models are typically aimed at specific, narrow tasks (Denève and Pouget, 2003; Coventry et al., 2005).

The concrete scenario used to test the model is as follows: An image of a visual scene with several distinct objects is presented to the model, either a camera image of a table top scene or an array of artificial stimuli for statistical evaluation of the response behavior. Based on such images, the system can then solve three types of tasks: (1) extract the spatial relation between two specified items in the scene and select an appropriate spatial term; (2) select and identify an object in the visual scene based on a spatial description; and (3) generate a spatial description for an item by choosing an appropriate reference object and a relational spatial term. The system determines spatial relations in the two-dimensional image plane, and covers the projective spatial terms "left", "right", "above", and "below".

The verbal questions posed to the model are represented as activations of discrete nodes that reflect the semantic roles and semantic content for different elements of each phrase. Issues of speech recognition or syntactic analysis are not addressed in the model, since the focus is on the grounding of spatial language. The model employs a relatively simple visual system with color as the only represented surface feature, and the objects in the visual scenes used to test the model are chosen in such a fashion that each of them can be uniquely identified by its salient color.

The version of the model presented here is the one originally published in Lipinski et al. (2012). This model is based upon previous DNF models of spatial language behaviors in robotics (Lipinski et al., 2009; Sandamirskaya et al., 2010), but the DNF architecture is re-structured to achieve a greater

degree of neural realism and behavioral flexibility. It should be noted that this model was designed before most of the work presented in the previous chapters. It employs the same general mechanisms for space-feature binding and spatial transformations as discussed so far, but some aspects are implemented in a simpler form. In particular, the model in its original form features only limited autonomy, with individual processing steps induced by providing a fixed sequence of external control inputs. Nonetheless, the system architecture is compatible with a more autonomous mode of behavior generation, as has been demonstrated in subsequent work (Richter et al., 2014a,b). In another work based on this model, it has been shown how the repertoire of spatial terms can be expanded, and how rotations of the reference frame can be included in addition to pure shifts as in the original model (van Hengel et al., 2012).

In the following sections, I will first give a detailed description of the DNF architecture and its general function. I will then demonstrate the generation of spatial language behaviors in this architecture in a variety of tasks. In one class of tasks, the basic capabilities of the system will be shown using natural images as visual inputs. In another class of tasks, the response statistics of the system will be tested under controlled stimulus conditions using artificial visual inputs. The results are compared to the results of psychophysical experiments in humans, and support the notion that the DNF model captures not only neural principles of visual processing, but also the behavioral characteristics of humans in applying spatial descriptions to visual scenes.

6.2 DNF architecture for relational spatial language

6.2.1 Model description

An overview of the DNF architecture is shown in Figure 6.1. The architecture can be divided into two functional parts: The first one provides a simple representation of the visual scene. It enables the system to form associations between color (as an object identifier) and object location in the visual image, using the mechanism described in Chapter 3. Instead of representing color information in a continuous field, it contains a set of dynamic nodes for discrete colors to serve for verbal input and output. The second part of the architecture processes spatial information to determine spatial relationships between objects. It performs a reference frame transformation into an object-centered reference frame, using the mechanism introduced in Chapter 4. The object-centered representation is linked to discrete nodes standing for spatial terms as interface to verbal expressions. I will describe the individual elements of this architecture in detail below. A formal description of the model with field equations and parameter values is given in

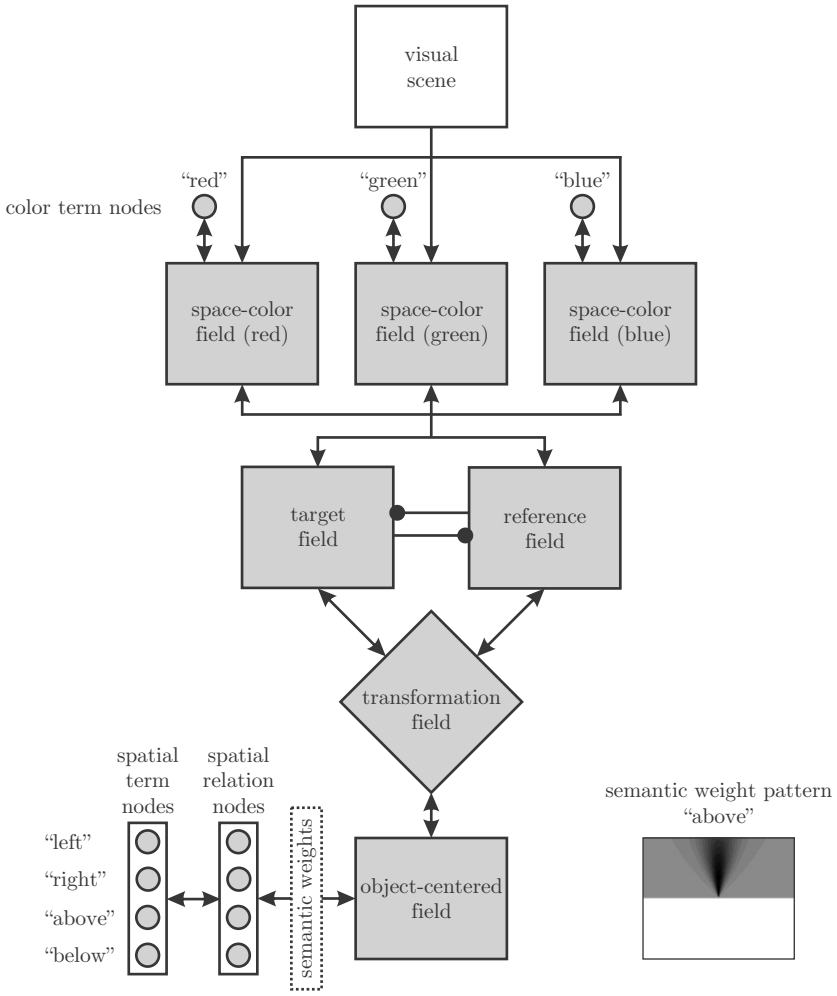


Figure 6.1: DNF architecture for spatial language behaviors. Two-dimensional DNFs over the image space are shown as gray rectangles, discrete dynamic nodes as gray circles. The four-dimensional transformation field is depicted as a gray diamond. Arrows indicate excitatory projections between elements, lines ending in circles indicate inhibition. The inset in the bottom right shows an exemplary semantic weight pattern (black standing for highest, white for lowest connection weights).

the next section.

The visual input for the system is supplied to a stack of *space-color fields* (shown side-by-side in the top part of Figure 6.1). These fields collectively take a role analogous to the two-dimensional visual sensory field in the biased competition model, but with some adjustments to the implementation: They cover the two-dimensional space of the image, while the color dimension is

reduced to three discrete hue values (red, green, and blue). This results in a stack of three fields over the same two-dimensional space. Visual input is preprocessed by a thresholding operation to determine regions of salient color for each of the three hue values. Every pixel in the image that exceeds a threshold for saturation and value in the hue-saturation-value color space provides localized input to the space-color field that best matches the pixel's hue value. This input induces active regions in these fields for colored objects in the scene. Lateral interactions consisting of local excitation and local surround inhibition are implemented within each color-space field, forming stabilized peaks from the visual input.

A *color term node* is coupled bi-directionally to each space-color field. These nodes stand for the verbal terms (or, more abstractly, the concepts) of the three colors "red", "green", and "blue". Input from each node globally excites the corresponding space-color field and strengthens any activation peaks in the field. In a reverse projection, the total output of each space-color field, integrated over space, is fed back as input to the corresponding color node. The nodes are coupled among each other in a competitive fashion, with each node exciting itself and inhibiting all others. The nodes can be activated by external inputs which are provided at certain times during a task to reflect elements of the verbal questions. Node activation at the end of a task can be read out as a response of the system regarding object identity.

The space-color fields furthermore provide input to two spatial fields, the *target field* and the *reference field*, defined over the same two-dimensional space in the reference frame of the visual image. The input is computed by summing the output of the three space-color fields at every location. This input is not sufficient, however, to induce activation peaks in the target and reference field. An additional "boost" input is necessary to lift the field activation beyond the output threshold. The boost input homogeneously raises the activation level over a whole field, and is given as an external control signal during certain phases of each task. Since this external input is provided at different times for the target field and the reference field, the two fields can form different peaks, even though the input they both receive from the space-color fields is always the same.

The lateral interactions in the target and reference fields are set up to create a selection regime, with only a single peak forming even in the presence of multiple inputs. As long as a weaker external boost input is provided, these peaks remain stable (even though no new activation peaks can form). When the boost signal is turned off completely, the peaks decay. Existing activation peaks in the target and reference fields provide feedback to the stack of space-color fields. They activate the same spatial region in each space-color field, such that peaks in these regions that were induced by the visual input are strenghened. In addition, target and reference fields mutually inhibit each other, such that peaks cannot form at the same location in

both of them.

The target field and the reference field provide two functionally distinct inputs to the reference frame transformation mechanism. This mechanism works in the same way as described in Chapter 4. But while the transformation previously combined retinal stimulus position and gaze direction to determine the body-centered position of a stimulus, now a combination of target object location and reference object location (both in the image reference frame) is used to compute the relative location of the target to the reference object. This can be expressed arithmetically as a vector subtraction (position of the target object in the image minus position of the reference object). The result is fed as input into another field defined over two-dimensional space, the *object-centered field*.

The connectivity required to implement the transformation for the case of one-dimensional inputs is shown in Figure 6.2. The target field projects a vertical ridge input into the two-dimensional *transformation field*, the reference field projects a horizontal ridge input, and an activation peak forms at the intersection of these ridges. For the projection to the object-centered field, the transformation field is then read out along the diagonal. Two things are noteworthy about the implementation here compared to the one used in Chapter 4. First, the orientation of the reference field is flipped compared to the gaze field in the model from Chapter 4. This reflects the fact that the reference object position has to be subtracted from the target object position rather than added to it. Second, the projections between the transformation field and the three other fields are fully bi-directional, so the reference frame transformation can be used in all possible directions. The robust implementation of this mechanism is made easier by the fact that all fields here support only a single activation peak, so no spurious intersections between multiple ridge inputs can occur.

In the full implementation of the transformation mechanism in the spatial language model, the transformation field spans a four-dimensional space to combine the two-dimensional inputs from both target and reference field. Lateral interactions in the transformation field are restricted to global inhibition to limit growth of activation in response to the external inputs. The object-centered field is defined over a two-dimensional space, and covers twice the range of the target and reference field. This way, it can capture all possible relative positions between objects represented in these two fields. As in the transformation field, the lateral interactions within the object-centered field consist only of global inhibition. This enables the field to hold broadly distributed activation patterns that are needed in certain tasks, and not only narrow peaks.

The different positions in the object-centered field now directly correspond to different spatial relations between target and reference object. By construction, the center of this field corresponds to the location of the refer-

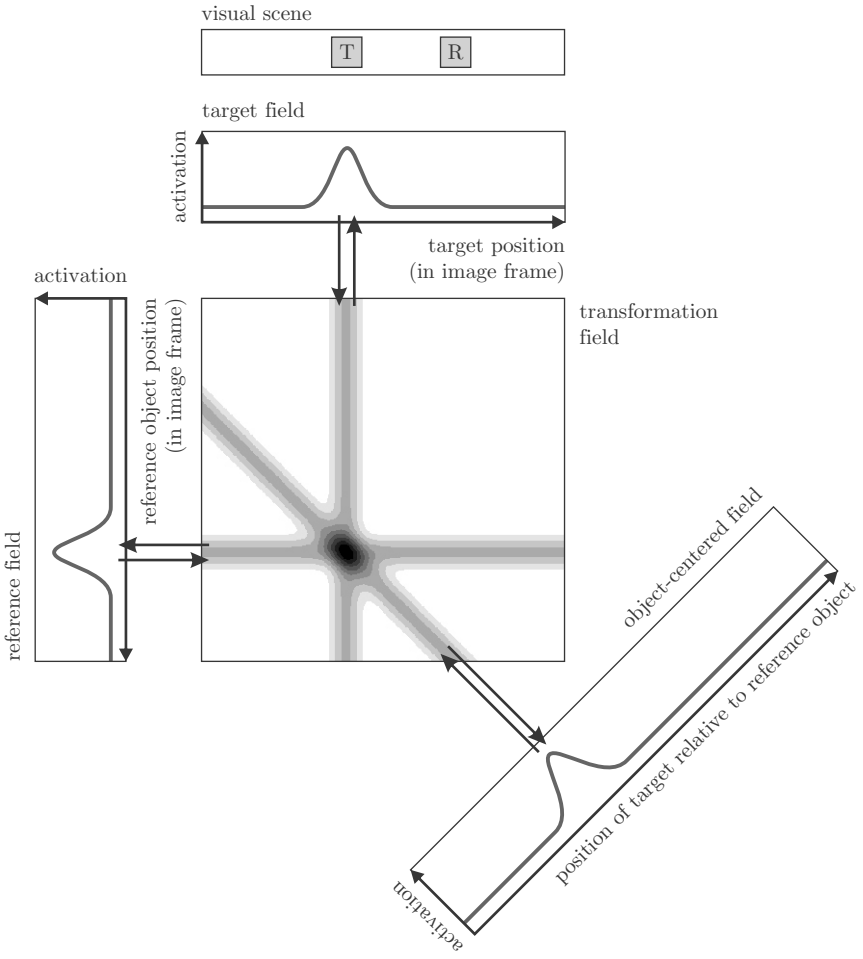


Figure 6.2: Connectivity for the transformation mechanism in the spatial language model (simplified version for a single spatial dimension). A visual scene with a target object (T) and reference object (R) is shown on top, the DNF architecture to determine the relative position between them is shown below. Arrows indicate excitatory projections between fields. The activation pattern in the transformation field is coded by gray values (with black standing for highest activation).

ence object. An activation peak forming in the right half of the field indicates that the target location is somewhere to the right of the reference object, a peak in the left half indicates that it is to the left, and so on. Due to these properties of the field, it is possible to map directly from field positions to a representation of discrete spatial relations through fixed synaptic connections.

These spatial relations are represented by a set of discrete dynamic nodes, the *spatial relation nodes*. There is one node for each of the relations "above", "below", "left", and "right". Each of these nodes is bidirectionally coupled to a portion of the object-centered field by a *semantic weight pattern* (see inset in Figure 6.1). These patterns contain graded connection weights, such that for the term "left", for instance, the region directly to the left of the field center receives a higher weight than regions diagonally to the left and above the center. In the forward projection, the strength of the input that each spatial relation nodes receives therefore reflects how well a peak position in the object-centered field matches the meaning of a certain relational spatial term. Through the reverse projection, the nodes can induce graded activation patterns in the object-centered field. The spatial relation nodes feature lateral interactions consisting of self-excitation and mutual inhibition that create a weak selection effect.

A second set of nodes, called *spatial term nodes*, is used to provide verbal input to the system and to read out a verbal response. There are four spatial term nodes matching the four spatial relation nodes, and each spatial term node is coupled bidirectionally to its corresponding spatial relation node. The spatial term nodes feature stronger competitive lateral interactions so that only one node can be strongly activated at any time. The reason to have these two separate sets of nodes for spatial relations and spatial terms lies in the different types of tasks that the model is applied to. On the one hand, the model should be able to select a single spatial term as a response to a question about the spatial relation between two objects, and it has to do so even if no term fits perfectly. On the other hand, I will also show tests of the model in an emulation of a rating task, where the model should give a graded response about the applicability of a term for a gradually varied arrangement of objects. This latter response is read out from the spatial relation nodes, while the spatial term nodes are used for the former type of task.

6.2.2 Field equations and parameters

The DNF model of spatial language can be formally described by a set of differential equations. A two-letter index is used in these equations to identify each field of the architecture, listed in Table 6.1. The spatial dimensions of the fields are defined over the space of input image positions, spanning 152×120 pixels. The object-centered field covers twice this range to be able to represent all possible relative positions within the input image. The color representation in the model uses a set of three discrete color hue values $V = \{\text{red}, \text{green}, \text{blue}\}$, and a set of four spatial relations $S = \{\text{left}, \text{right}, \text{above}, \text{below}\}$ is supported.

A set of space-color fields is defined over the space of image positions,

field name	field index	h	β	c^{exc}	c^{inh}	c^{gi}
space-color fields	sc	-2	4	2.5	10	0
color term nodes	ct	-4	4	2.5	0	2
reference field	rf	-4	4	10	0	0.02
target field	tg	-4	4	10	0	0.02
transformation field	tn	-2	4	0	0	0.0075
object-centered field	oc	-1	2.5	0	0	0.000175
spatial relation nodes	sr	-1.95	2	0.25	0	1.25
spatial term nodes	st	-4	4	2.5	0	4

Table 6.1: Field indices, field parameters, and lateral interaction parameters.

with one field for each color $v \in V$, governed by the field equations

$$\begin{aligned}\tau \dot{u}_{\text{sc}}(x,y,v) = &- u_{\text{sc}}(x,y,v) + h_{\text{sc}} + i_{\text{sc}}(x,y,v) \\ &+ [k_{\text{sc,sc}} * f(u_{\text{sc}}(\cdot,\cdot,v))](x,y) + c^{\text{exc}}_{\text{sc,ct}} f(u_{\text{ct}}(v)) \\ &+ [k_{\text{sc,tg}} * f(u_{\text{tg}})](x,y) + [k_{\text{sc,tg}} * f(u_{\text{rf}})](x,y) \\ &+ q_{\text{sc}} \xi(x,y,v).\end{aligned} \tag{6.1}$$

The lateral interactions described by the difference-of-Gaussians kernel $k_{\text{sc,sc}}$ act only within each space-color field, and not across different colors. The external input for each space-color field is determined directly from a visual image (camera image or artificial visual scene) by determining saliently colored pixels within a certain range of hue values. The input at the locations of these pixels is set to a fixed value $i_{\text{sc}}(x,y,v) = 2$, and is zero everywhere else.

The behavior of the color term nodes is specified by the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{ct}}(v) = &- u_{\text{ct}}(v) + h_{\text{ct}} + i_{\text{ct}}(v) + b_{\text{ct}} \\ &+ c^{\text{exc}}_{\text{ct,ct}} f(u_{\text{ct}}(v)) - c^{\text{gi}}_{\text{ct,ct}} \sum_{v' \in V} f(u_{\text{ct}}(v')) \\ &+ c_{\text{ct,sc}} \iint f(u_{\text{sc}}(x,y,v)) dx dy + q_{\text{ct}} \xi(v).\end{aligned} \tag{6.2}$$

Here, i_{ct} is a specific input for a single node (set to $i_{\text{ct}} = 5$ to specify a color in a verbal task), and b_{ct} is global boost of all color nodes (set to $b_{\text{ct}} = 4$ to obtain a color response).

The target field and the reference field are defined with symmetric con-

nectivity, and they use the same kernels for projections from other fields:

$$\begin{aligned}\tau\dot{u}_{\mathrm{tg}}(x,y) = &- u_{\mathrm{tg}}(x,y) + h_{\mathrm{tg}} + b_{\mathrm{tg}} + [k_{\mathrm{tg,tg}} * f(u_{\mathrm{tg}})](x,y) \\ &+ [k_{\mathrm{tg,tn}} * F_{\mathrm{tg}}(u_{\mathrm{tn}})](x,y) - [k_{\mathrm{tg,rf}} * f(u_{\mathrm{rf}})](x,y) \\ &+ \sum_{v\in V}[k_{\mathrm{tg,sc}} * f(u_{\mathrm{sc}}(\cdot,\cdot,v))](x,y) + q_{\mathrm{tg}}\xi(x,y)\end{aligned} \tag{6.3}$$

$$\begin{aligned}\tau\dot{u}_{\mathrm{rf}}(x,y) = &- u_{\mathrm{rf}}(x,y) + h_{\mathrm{rf}} + b_{\mathrm{rf}} + [k_{\mathrm{rf,rf}} * f(u_{\mathrm{rf}})](x,y) \\ &+ [k_{\mathrm{tg,tn}} * F_{\mathrm{rf}}(u_{\mathrm{tn}})](x,y) - [k_{\mathrm{tg,rf}} * f(u_{\mathrm{tg}})](x,y) \\ &+ \sum_{v\in V}[k_{\mathrm{tg,sc}} * f(u_{\mathrm{sc}}(\cdot,\cdot,v))](x,y) + q_{\mathrm{rf}}\xi(x,y)\end{aligned} \tag{6.4}$$

The input from the transformation field to these two fields is obtained by integrating the transformation field output over the disregarded dimensions, as specified below. In order to generate an activation peak in one these fields, the global boost inputs b_{tg} or b_{rf} are set to a value of 4 at different times during task execution, and then set to a value of 2 to actively maintain that activation peak.

The transformation field is defined over a four-dimensional space, with the first two dimensions (variables x and y in the equation) reflecting the location of the target object in the image, the following two dimensions (r and s) reflecting the position of the reference object. The field equation is given by

$$\begin{aligned}\tau\dot{u}_{\mathrm{tn}}(x,y,r,s) = &- u_{\mathrm{tn}}(x,y,r,s) + h_{\mathrm{tn}} \\ &- c^{\mathrm{gi}}_{\mathrm{tn,tn}} \iiiint f(u_{\mathrm{tn}}(x',y',r',s'))dx'dy'dr'ds' \\ &+ [k_{\mathrm{tn,tg}} * f(u_{\mathrm{tg}})](x,y) + [k_{\mathrm{tn,tg}} * f(u_{\mathrm{rf}})](r,s) \\ &+ [k_{\mathrm{tn,oc}} * f(u_{\mathrm{oc}})](x-r,y-s) + q_{\mathrm{tn}}\xi(x,y,r,s).\end{aligned} \tag{6.5}$$

For the projections to the target, reference, and object-centered field, the following integrals are defined:

$$F_{\mathrm{tg}}(u_{\mathrm{tn}})(x,y) = \iint f(u_{\mathrm{tn}}(x,y,r,s))drds \tag{6.6}$$

$$F_{\mathrm{rf}}(u_{\mathrm{tn}})(r,s) = \iint f(u_{\mathrm{tn}}(x,y,r,s))dxdy \tag{6.7}$$

$$F_{\mathrm{oc}}(u_{\mathrm{tn}})(x,y) = \iint f(u_{\mathrm{tn}}(x+r,y+s,r,s))drds \tag{6.8}$$

The dynamics of the object-centered field is described by the field equation

$$\begin{aligned}\tau\dot{u}_{\mathrm{oc}}(x,y) = &- u_{\mathrm{oc}}(x,y) + h_{\mathrm{oc}} + b_{\mathrm{oc}} - c^{\mathrm{gi}}_{\mathrm{oc,oc}} \iint f(u_{\mathrm{oc}}(x',y'))dx'dy' \\ &+ [k_{\mathrm{oc,tn}} * F_{\mathrm{oc}}(u_{\mathrm{tn}})](x,y) \\ &+ c_{\mathrm{oc,sr}} \sum_{s\in S} W_{\mathrm{s}}(x,y) f(u_{\mathrm{sr}}(s)) + q_{\mathrm{oc}}\xi(x,y).\end{aligned} \tag{6.9}$$

The field receives a global boost input $b_{\text{oc}} = 1$ during specific periods of some tasks. The connections between the object-centered field and the spatial relation nodes are mediated by the semantic weight patterns W_s, which are generated as a superposition of a Gaussian function in polar coordinates (following O'Keefe, 2003), and a sigmoid function along either the vertical (for relations "above" and "below") or the horizontal axis ("left" and "right").

The spatial relation nodes for the relations $s \in S$ are governed by the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{sr}}(s) = &- u_{\text{sr}}(s) + h_{\text{sr}} + c^{\text{exc}}_{\text{sr,sr}} f(u_{\text{sr}}(s)) - c^{\text{gi}}_{\text{sr,sr}} \sum_{s' \in S} f(u_{\text{sr}}(s')) \\ &+ c^{\text{exc}}_{\text{sr,st}} f(u_{\text{st}}(s)) - c^{\text{gi}}_{\text{sr,st}} \sum_{s' \in S} f(u_{\text{st}}(s')) \\ &+ c^{\text{exc}}_{\text{sr,oc}} \iint W_{\text{s}}(x,y) f(u_{\text{oc}}(x,y)) dxdy + q_{\text{sr}} \xi(s).\end{aligned} \tag{6.10}$$

The differential equations for the corresponding spatial term nodes are given as

$$\begin{aligned}\tau \dot{u}_{\text{st}}(s) = &- u_{\text{st}}(s) + h_{\text{st}} + i_{\text{st}}(s) + b_{\text{st}} \\ &+ c^{\text{exc}}_{\text{st,st}} f(u_{\text{st}}(s)) - c^{\text{gi}}_{\text{st,st}} \sum_{s' \in S} f(u_{\text{st}}(s')) \\ &+ c^{\text{exc}}_{\text{st,sr}} * f(u_{\text{sr}}(s)) + q_{\text{st}} \xi(s).\end{aligned} \tag{6.11}$$

Analogous to the color term nodes, the spatial term nodes can receive an individual input $i_{\text{st}}(s) = 5$ to specify a relational term in a verbal task, or a global boost input $b_{\text{st}} = 4$ to generate a response.

For numerical simulations, the two-dimensional fields are sampled with one unit per pixel in the input image. In the four-dimensional transformation field, the space is downsampled by a factor of 8. The width σ^{exc} of lateral excitation in all fields and for all projections between fields is 4 pixels of the input image, except for the broad feedback projection from the target and reference fields to the space-color fields, where it is 15 pixels. The width σ^{inh} of lateral inhibition in the space-color fields is 10 pixels. The noise level of all fields is set to $q = 0.1$ except for the two sets of response nodes, where it is reduced to $q_{\text{ct}} = q_{\text{st}} = 0.025$. The remaining parameters of fields and lateral interactions are given in Table 6.1, the parameters for projections between fields are given in Table 6.2.

6.3 Demonstrations

I will show the function and capabilities of the DNF architecture in five demonstrations (previously described in Lipinski et al., 2012). In the first one, the system has to determine the spatial relation between two objects in

projection index	c^{exc}	c^{gi}
ct, sc	0.01	0
sc, ct	1	0
tg, sc	6	0
sc, tg	4	0.0005
tn, tg	5	0
tg, rf	1.5	0
tg, tn	0.175	0
oc, tn	0.75	0
tn, oc	1.5	0
sr, oc	0.00215	0
oc, sr	1.0	0
st, sr	2	0
sr, st	4	2.5

Table 6.2: Connection strengths (excitatory and global inhibitory) for projections between fields and/or nodes.

a natural scene, effectively answering the question "Where is the green item relative to the red item?" The second demonstration is a variant of this task applied to artificial stimuli, in which the model is used to reproduce human rating data from Regier and Carlson (2001) by judging the applicability of the spatial term "above" for varied arrangements of objects. In the third demonstration, the system has to select and identify an object from the visual scene given a spatial description, answering a question of the type "What is to the right of the blue item?" In the fourth demonstration, the system has to solve a more open ended task, answering a question of the type "Where is the red object?" Here, the system has to select an appropriate reference object from the visual scene and a matching spatial term to generate a response like "To the right of the blue object." In the fifth and final demonstration, the same behavior is applied to a set of artificial stimuli to show that the model can reproduce human reference object selection behavior.

All tasks are solved by the same DNF architecture with identical parameters. Different sequences of inputs that reflect the components of the verbal tasks are supplied to this architecture, as well as additional control inputs to structure the behavior of the system for the different types of tasks. These inputs are given with a fixed timing for each type of task, although the details of the timing (and in some cases even the order of the inputs) are not critical as long as the system is given sufficient time to settle into a stable state. The responses in all tasks are read out from the states of the dynamic nodes for colors and spatial terms at the end of the task. I will give a detailed description of the sequence of inputs and the resulting evolution

of activation patterns in the model for each demonstration.

6.3.1 Spatial term selection

In this task, the model is provided with a camera image of a table top scenario showing three colored objects (Figure 6.3a), and has to answer the question "Where is the green item relative to the red item?" Note that the correct response, "to the right", can only be generated for this scene if the red object is correctly chosen as spatial referent. The green object is neither in the right part of the image nor to the right of the other object in the scene.

The visual input induces weak activation peaks in the stack of space-color fields, with one peak in each of the three fields to reflect the three differently colored items (Figure 6.3b); activation in all other fields is initially at the resting level. The first input to the system in this task now specifies that the green item should be the target object. The "green" color node is activated, and at the same time the activation level of the target field is globally increased (indicated as a "boost" input in Figure 6.3c). The active color node raises the activation level in the "green" space-color field, and significantly strengthens the activation peak present in that field. When the target field is boosted, it receives the strongest input from the position of this amplified peak. The target field forms a peak at this position, which reflects the location of the green item in the image. This peak remains stable when the color input is turned off and the boost input is reduced to an intermediate level.

The next step is to specify the red object as the referent in the task. Analogously to the first step, the "red" color node is activated and the reference field is homogeneously boosted (Figure 6.3d). The reference field forms a peak at the location of the red item, which is highlighted by the color input. The color input is then turned off again, and the boost input for the reference field is reduced to an intermediate level to retain the activation peak.

As soon as peaks are present in both the target and the reference field, the relative position of the target to the reference object is determined autonomously through the reference frame transformation mechanism. In the present scenario, a peak appears in the object-centered field to the right of the midpoint. This region in the field is strongly coupled to the "right" spatial relation node, and the node is activated by input from the field. (Being located near the vertical midline, the peak also provides weak input to both the "above" and "below" spatial relation nodes.) As the spatial relation node for "right" becomes sufficiently activated, it provides input to the corresponding spatial term node, while also projecting its semantic weight pattern back to the object-centered field.

After a brief delay that allows these dynamics to unfold, all spatial term

nodes are globally activated by another boost input. The node for "right" that receives significant input from its associated spatial relation node becomes fully active and suppresses the other spatial term nodes. It is thereby selected as the response of the system, providing the correct answer for the given question.

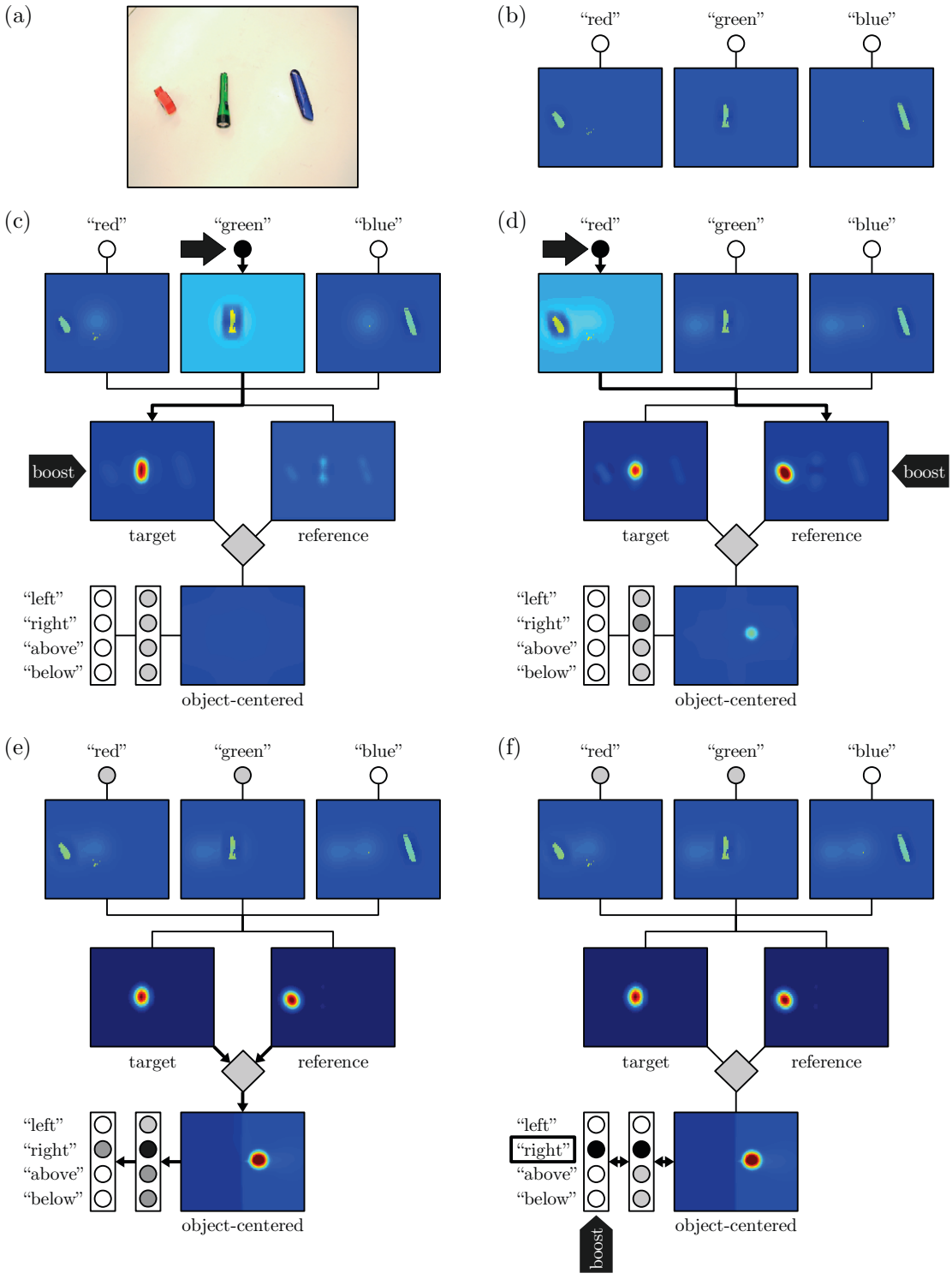
To illustrate the relevance of this last step for the response selection, and to show that the model can also deal with more ambiguous inputs, a second instance of the same task with a different visual input is shown in Figure 6.4. Here, the central green object is shifted such that it is now located diagonally to the right and above the red object (Figure 6.4a). Target and reference object are specified as before and the relative position of the two objects is autonomously determined by the model, yielding the state shown in Figure 6.4b directly before the response generation.

Since the peak location in the object-centered field matches the semantic weight patterns for "above" and "right" about equally well, the two corresponding spatial relation nodes are activated, and both provide input to their associated spatial term nodes. When the spatial term nodes now receive a boost input, the two active nodes compete with each other by means of lateral interactions. One node is selected as a response (the "right" node in Figure 6.4), while all others are suppressed. This selection of a single term reflects the fact that in speech production, one word has to be produced at a time. This does not rule out that a second matching term is still added at a later time (which could be achieved in the model by repeating the selection process while suppressing the already produced term), but this is not covered in the present implementation.

6.3.2 Rating spatial term applicability

Rating tasks constitute one common method to experimentally explore human spatial language behavior. In this type of task, participants are shown an arrangement of two or more stimuli, and are asked to provide a rating (e.g., as a numerical value from 0 to 9) for the applicability of a certain spatial description to this scene, such as "The small item is above the large item." Tasks of this type help to specify the precise semantics of individual spatial terms. For instance, the results show that not all target objects whose vertical position is higher than the reference object are equally judged to be "above" that object. While relative vertical position is one factor that influences the "above" rating, only a certain region relative to the reference object receives very high ratings. The present demonstration shows that the DNF model, using appropriately defined semantic weight patterns, can reproduce these experimental data.

Regier and Carlson (2001) have furthermore used rating tasks to investigate how object positions are determined in judging spatial relations. Specifically, they asked whether the position of the reference object is deter-


(a)
(b)
"red"
"green"
"blue"
(c)
"red"
"green"
"blue"
boost
target
reference
"left"
"right"
"above"
"below"
object-centered
(d)
"red"
"green"
"blue"
boost
target
reference
"left"
"right"
"above"
"below"
object-centered
(e)
"red"
"green"
"blue"
target
reference
"left"
"right"
"above"
"below"
object-centered
(f)
"red"
"green"
"blue"
target
reference
"left"
"right"
"above"
"below"
boost
object-centered

mined as the object's center of mass, its geometrical midpoint, or the point within the reference object closest to the target object. To answer this question, they used a rating task with an oblong reference object. This referent was oriented horizontally or vertically, and the position of the target object was adjusted to keep either the center-of-mass vector between the two object constant (Figure 6.5a and b), or to keep the proximal vector constant, which connects the closest points between the two objects (Figure 6.5c and d). They tested both versions in separate experiments as part of an array of possible target object locations around the reference object. An additional experiment was used by Regier and Carlson to distinguish between the effect of the reference object's center of mass and its horizontal midpoint for the "above" ratings. To this end, they tested ratings for different target positions in the region above either an upright or an inverted triangle (Figure 6.6).

The results showed that both the center of mass and the closest point in the reference object affect the assessment of the "above" relationship. The horizontal midpoint does not seem to play a significant role. Regier and Carlson proposed a mathematical model to explain the results, inspired by principles of neural processing. The attentional vector sum (AVS) model proposes that a vector is determined from every point within the reference object to the target object (which is modeled as a single point). Each vector is then weighted according to an "attentional beam", which is centered on the point in the reference object that is closest to the target. A single direction vector is determined by averaging over these weighted vectors, and the estimated rating is determined from the orientation of this vector (in combination with a measure of the vertical offset between the target and the highest point in the reference object).

To show that the DNF model is able to capture these behavioral characteristics, the tasks were emulated in simulation. Artificial images of 152×120 pixels were generated. For the first two experiments, an (invisible) grid with 5×5 cells of 24×24 pixels was laid over the central part of the image. A

Figure 6.3 *(preceding page)*: Evolution of activation patterns in the spatial language model for the task "Where is the green item relative to the red item?" DNFs and dynamic nodes are shown in the same arrangement as in Figure 6.1, activation states of nodes are indicated by their gray value (black indicating highest activation). Large black arrows show specific input to nodes, block arrows show global boost of fields or nodes. Relevant projections between fields in each step are shown as thicker lines. (a) Visual input for the task. (b) Initial activation state of the space-color fields. (c) Specification of the green item as target object. (d) Specification of the red item as reference object. (e) Automatic assessment of spatial relation. (f) Response generation.

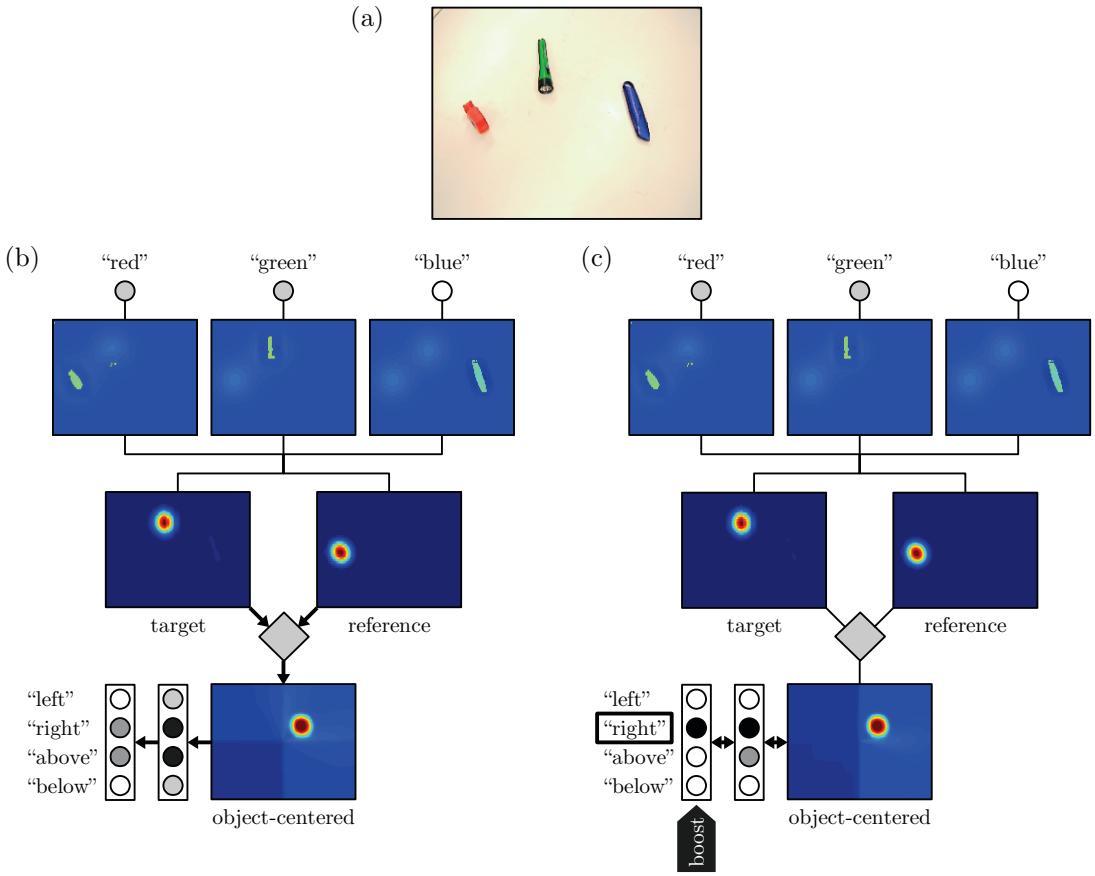


Figure 6.4: Response generation for the same task as in Figure 6.3 in the case of ambiguous spatial relations. (a) Visual input. (b) Activation state after target and reference object have been specified. (c) Response generation, with selection of a single spatial term through competition between spatial term nodes.

green rectangle of 24×8 pixels was placed as reference object in the central cell, and a red square of 8×8 pixels as target object placed in each of the surrounding cells in subsequent trials. The placement of the target within the cell is modified in the different versions of the trials as illustrated in Figure 6.5. In the first experiment, the center-of-mass vector is held fixed over different orientations of the reference object (corresponding to Figure 6.5a and b). In the second experiment, the proximal vector is held constant (corresponding to Figure 6.5c and d). For the third experiment, the reference object is a right-angle triangle with edge lengths of 36 and 12 pixels. The target object takes the same form as before, placed in three positions as illustrated in Figure 6.6.

The task is performed by the model in the same fashion as the spatial

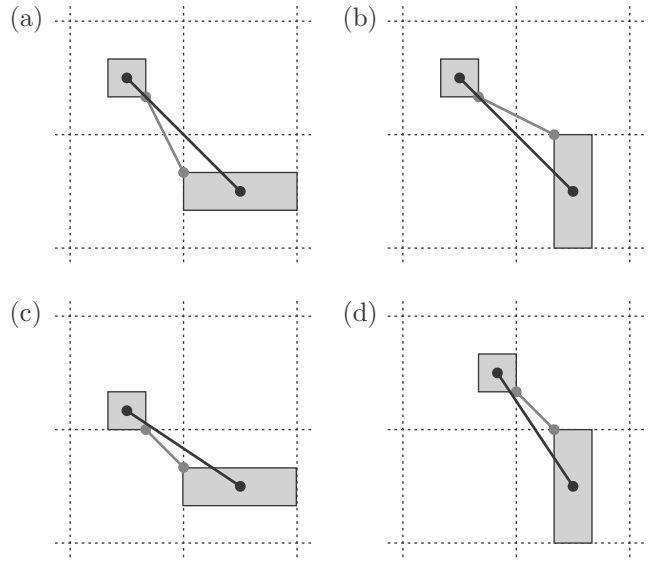


Figure 6.5: Proximal vectors (gray) and center-of-mass vectors (black) between a small square target object and a larger rectangular reference object. (a, b) The center-of-mass vector remains the same when the reference object is rotated, the proximal vector changes. (c, d) By adjusting the position of the target object, the proximal vector is held constant over rotations of the reference object, the center-of-mass vector now changes.

term selection task described above, with one difference: The final boost input to the spatial term nodes that effects the selection of a single response is omitted. Instead, the output of the "above" spatial relation node is read out at the end of the task, and scaled from its original range $[0, 1]$ to the range $[0, 9]$ used in the rating task. Note that the generation of a rating responses in humans, using an arbitrarily defined scale, certainly involves more complex processes, but attempting to derive the properties of these processes from the rating results would be purely speculative. Deriving the ratings in the model directly from the neural activation of the relevant nodes appears as the most parsiminous approach.

The rating results for the first task is shown in Table 6.3, for the second task in Table 6.4, with empirical results from Regier and Carlson (2001) given in parentheses for comparison. The results show that the model qualitatively captures the pattern of rating responses given by human participants, with highest rating when the target is placed in the two cells directly above the reference object, decreasing for the more diagonally placed targets, and hitting zero when the vertical position of the target is lower than the referent (model fit in first experiment, vertical reference object: $R^2 = 0.98$, $RMSD = 0.55$;

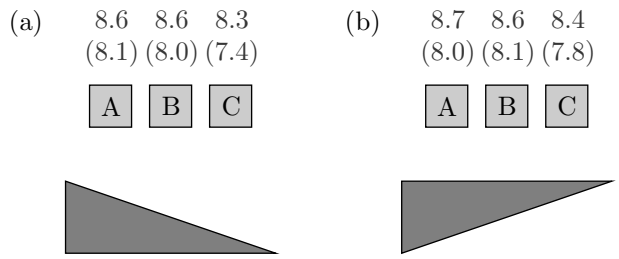


Figure 6.6: Stimulus arrangement and rating results for the spatial term "above" in the third rating experiment. The target object (small square) is placed at different locations above a larger, triangular reference object, either in an upright (a) or inverted orientation (b). Experimental rating results from Regier and Carlson (2001) are given in parentheses.

horizontal reference object: $R^2 = 0.97$, $RMSD = 0.60$; model fit in second experiment, vertical reference object: $R^2 = 0.99$, $RMSD = 0.65$; horizontal reference object: $R^2 = 0.96$, $RMSD = 0.92$).

The model also captures the key effects of proximal vector and center-of-mass vector between target and referent. Regier and Carlson evaluated these effects by comparing the mean ratings for the oblique (left and right above) target locations between the horizontally and vertically oriented reference objects. In the first experiment, only the proximal vector changes between the two conditions (Figure 6.5a and b). It is steeper for the horizontally oriented reference object, and the authors found an increase in rating of 0.093 for this condition. In the simulations of the DNF model, this effect is reproduced, with a difference of 0.075 in mean ratings between the two conditions. These results show that both in the experiments and in the model, the proximal vector between the two objects does have an influence on spatial term applicability, independent of the center-of-mass vector.

In the second experiment, the proximal vector remains the same when the reference object is rotated, but the center-of-mass vector changes (Figure 6.5a and b). In the empirical data, rating results for the oblique target locations in the vertical condition (with steeper center-of-mass vector) were increased by 0.11 compared to the horizontal condition. In the model the mean ratings change in the same fashion, although the amplitude of this change is an order of magnitude larger (1.64). The rating in both humans and DNF model is thus sensitive to center-of-mass orientation. The third experiment confirms that it is indeed the center of mass of the reference object, and not just its horizontal midpoint, that is relevant for judging spatial relations (Figure 6.6). Target location A in the figure (located over the wide end of the triangle, closer to its center of mass) received higher "above" ratings than target location C (over the narrow end of the triangle). The

6.0 (6.7)	8.3 (7.4)	8.8 (8.9)	8.3 (7.4)	6.0 (6.8)
5.4 (5.6)	7.3 (6.6)	8.6 (8.9)	7.3 (6.2)	5.4 (6.0)
0.9 (0.9)	0.9 (0.9)		0.9 (1.0)	0.9 (1.3)
0.0 (0.6)	0.0 (0.3)	0.0 (0.6)	0.0 (0.4)	0.0 (0.6)
0.0 (0.4)	0.0 (0.4)	0.0 (0.3)	0.0 (0.6)	0.0 (0.3)
5.9 (6.5)	8.3 (7.3)	8.8 (8.9)	8.3 (7.0)	5.9 (6.9)
5.5 (6.2)	7.6 (6.4)	8.6 (8.4)	7.6 (6.9)	5.5 (6.2)
0.9 (0.7)	0.9 (0.8)		0.9 (0.7)	0.9 (0.8)
0.0 (0.4)	0.0 (0.5)	0.0 (0.3)	0.0 (0.4)	0.0 (0.3)
0.0 (0.4)	0.0 (0.4)	0.0 (0.4)	0.0 (0.3)	0.0 (0.3)

Table 6.3: Effect of proximal vector orientation on spatial term ratings in the DNF model. Ratings for the term "above" are shown for different placements of the target object in a 5×5 grid around the central reference object. The reference object is oriented either vertically (top part of the table) or horizontally (bottom). The center-of-mass vector remains fixed between these conditions, the proximal vectors differ. Empirical rating results from Regier and Carlson (2001) are given in parentheses.

difference was comparable for empirical results (0.45) and DNF model (0.28, both averaged over the two orienations of the triangle).

The effect of the reference object's center of mass on spatial term ratings in the model is straightforward to explain. The position of the peak in the reference field is determined primarily by the input from the space-color fields, smoothed with a Gaussian kernel. For contiguous and convex stimuli, the resulting location of the activation peak in the reference field is approximately at the center of mass of the stimulus. The cause for the observed effect of the proximal vector is more subtle. Both the reference field and the target field project feedback to the space-color fields, in the form of broad Gaussians, that strengthen the stimulus representations at matching locations. If the two objects are relatively close to each other, the feedback from the target field also overlaps with the location of the reference object and raises the activation levels for those parts of the object that are closest to the target. Due to this attentional modulation, the peak in the reference field is pulled slightly toward the target object location. This is largely consistent with the explanation used in the AVS model, where attentional weighting is used to give points within the reference object that are closer to the target a higher impact on the computed average vector.

The significantly greater influence of the center-of-mass vector in the model as compared to behavioral data may be an effect of the very simple visual system that is used in the model. Here, every saliently colored point

7.4 (6.6)	8.6 (7.3)	8.8 (8.7)	8.6 (7.7)	7.4 (6.9)
6.3 (6.3)	8.4 (6.7)	8.6 (8.6)	8.4 (7.0)	6.3 (6.3)
0.9 (1.2)	1.1 (1.1)		1.1 (1.5)	0.9 (1.2)
0.0 (0.3)	0.0 (0.4)	0.0 (0.5)	0.0 (0.4)	0.0 (0.4)
0.0 (0.5)	0.0 (0.4)	0.0 (0.3)	0.0 (0.3)	0.0 (0.5)
6.3 (6.7)	8.0 (7.0)	8.8 (9.0)	8.0 (7.4)	6.3 (7.1)
3.9 (5.9)	5.7 (6.8)	8.4 (8.9)	5.7 (6.7)	3.9 (6.4)
0.9 (1.1)	0.9 (1.2)		0.9 (1.2)	0.9 (1.6)
0.1 (0.6)	0.0 (0.6)	0.0 (0.4)	0.0 (0.7)	0.0 (0.7)
0.0 (0.6)	0.0 (0.5)	0.0 (0.9)	0.0 (0.9)	0.1 (0.9)

Table 6.4: Effect of center-of-mass vector orientation on spatial term ratings in the DNF model. The locations of the target objects are adjusted between the condition with a vertically oriented (top) and horizontally oriented reference object (bottom) to keep the proximal vector fixed. Empirical rating results from Regier and Carlson (2001) are given in parentheses.

in the input image simply contributes an equal input to the initial visual representation. A more detailed model of the biological visual system (with edge detection and only a weak response to uniformly colored areas) would likely allow for greater effects of attentional modulation. It is also possible that the shape of the reference object influences spatial term ratings in the experiment by defining an axes to divide space—for instance, a horizontally oriented rectangle may be viewed as a boundary that separates space into clearly distinct "below" and "above" regions, whereas a vertically oriented rectangle is less likely to be interpreted in that way. Such effects may counteract the effect of the center-of-mass vector in the second experiment, but are not accounted for in the current model.

6.3.3 Object selection based on spatial description

The task described above determines the spatial relationship between two given objects. To make actual use of spatial language, the system should also be capable of performing the inverse operation, that is, use a relational spatial description to select an object in a visual scene. This corresponds to answering a question of the type "What is above the blue item?" Here, the reference object (the blue item) and a relational spatial term ("above") are given. The system has to select a matching target object in the scene, and responds by giving the color of that object.

Figure 6.7 shows the visual scene that is used here and the sequence of activation states for solving the task. First, the reference object is selected, in the same fashion as in the previous task: The node for "blue" is activated

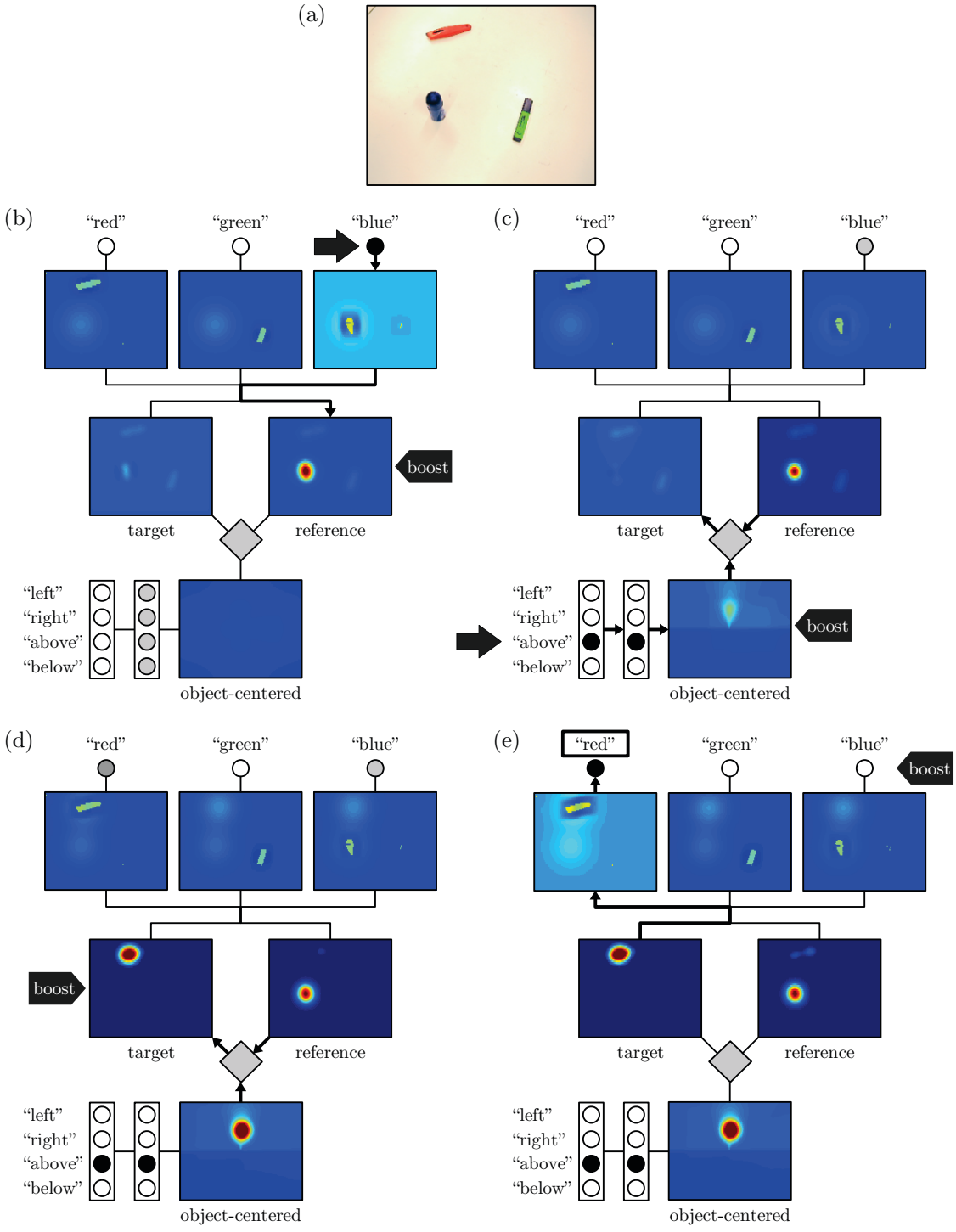
by an external input, reflecting the identification of the reference object in the verbal phrase, and the reference field is boosted simultaneously. This induces an activation peak in the reference field that reflects the location of the blue object in the image (Figure 6.7b), and that remains stable when the color input is turned off and the boost is reduced.

Now, an external input is applied to activate the spatial term node for "above", and the object-centered field is boosted simultaneously. The spatial term node activates the corresponding spatial relation node. That node then projects to the object-centered field through its specific synaptic connection pattern, which implements the spatial semantic template for "above". This pattern then shapes the activation distribution in the object-centered field, so that the activation at each point corresponds to how well this point matches the meaning of "above" (Figure 6.7c). The additional boost input to the object-centered field globally lifts the activation level so that it pierces the output threshold and the spatial pattern is projected to the transformation field. Note that due to the absence of strong local interactions in the field and the use of a soft sigmoid function (with low value for the steepness parameter β), the object-centered field can project a broad and graded output pattern rather than a single localized peak.

The transformation field now receives input from two sources: From the reference field, where a peak reflects the reference object location in the image, and from the object-centered field, reflecting the area relative to the reference point that matches the term "above". These two pieces of information are combined, and the transformation field projects a shifted version of the spatial semantic pattern to the target field. This input moderately activates the regions in the target field that are located above the reference object.

In the next step, the target field is boosted to spatially select a target object from the input image. The target field receives localized input from the stack of space-color fields, reflecting the locations of salient objects in the input image. (The location of the selected reference object, however, is suppressed by inhibitory input from the reference field.) The additional broad input from the transformation field strengthens those object locations that are positioned in the area above the reference object. In the competition process, these locations have a clear advantage, and in the present example, a peak forms in the target field at the location of the red object (Figure 6.7d).

In the final step, the selected target object has to be identified by determining its color. The peak in the target field projects spatial feedback to the stack of space-color fields. This strengthens the stimulus-induced peak in the "red" space-color field, and thereby increases the total output of this field. When the set of color nodes receives a global boost input to induce a selection decision, the node for "red" has the highest activation level and prevails in the competition. It suppresses the other color nodes, and the resulting


(a)
(b)
"red"
"green"
"blue"
boost
target
reference
"left"
"right"
"above"
"below"
object-centered
(c)
"red"
"green"
"blue"
target
reference
"left"
"right"
"above"
"below"
boost
object-centered
(d)
"red"
"green"
"blue"
boost
target
reference
"left"
"right"
"above"
"below"
object-centered
(e)
"red"
"green"
"blue"
boost
target
reference
"left"
"right"
"above"
"below"
object-centered

activation pattern of these nodes yields the response "red" for the given task (Figure 6.7e). Note that in this last step, the reference field likewise projects feedback to the color-space fields, which strengthens the activation in the "blue" field. However, the target field is still boosted, so that the peak in this field is strenghened and its influence dominates over the influence of the reference field.

As a variant of this procedure, the system is also capable of solving tasks of the type "The red object is above which other objects?" In this form—which is less common in natural language—the target object is specified and the reference object has to be determined. Since the target and reference fields are symmetric in the architecture, and since the spatial transformation mechanism between target, reference, and object-centered field can act in all possible directions, this task variant can be solved in a fashion entirely analogous to the process described here. The roles of target and reference fields are switched, and the system will produce the color of the most appropriate reference object as a response.

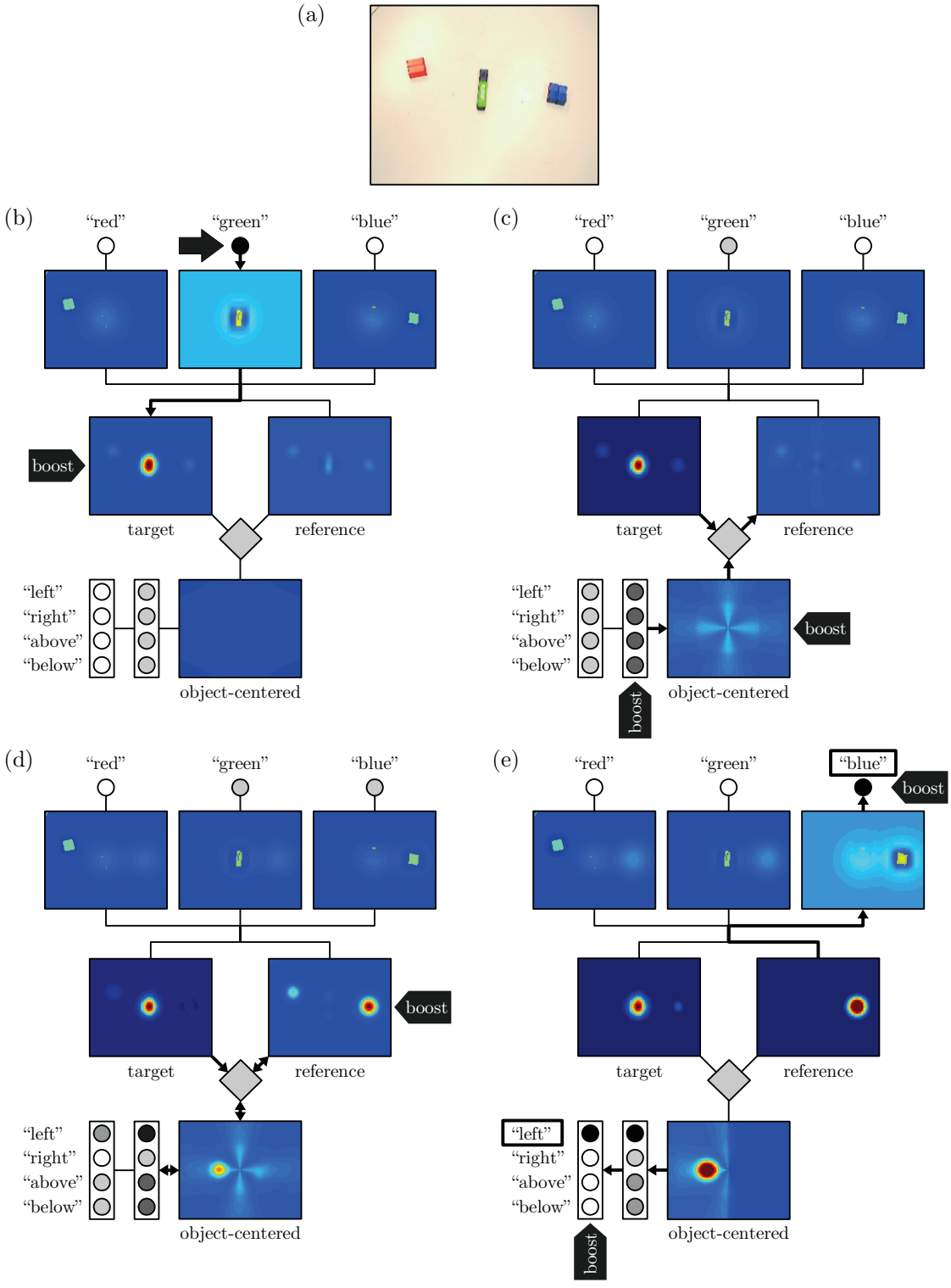
6.3.4 Generating spatial descriptions

In the last type of task, the system has to generate a relational spatial description of an object in a given scene. It has to answer a question of the type "Where is the green item?" The system's response in this task should take the form "to the left of the blue item", specifying both a reference object and a matching spatial term. This task is more open ended, in that there can be multiple combinations of reference object and spatial term in a scene that yield a valid description of an object's location. I will describe the general model mechanism for solving this task here, and provide comparisons to experimental data in the next section to show that the model captures key properties of human reference object selection behavior in ambiguous conditions.

Figure 6.8 shows the steps to process this task in the model. The input image shows a green highlighter near the center of the image, with a red stack of blocks to its left and a blue stack of blocks to its right, and the verbal task is "Where is the green item?" This task specifies the green item as target in a relational spatial expression. This information is given to the DNF model in the first step (Figure 6.8b), by simultaneously activating the

---

Figure 6.7 *(preceding page)*: Evolution of activation patterns for the task "What is above the blue item?" (a) Visual input. (b) Specification of the blue item as reference object. (c) Specification of the spatial term "above". (d) Spatial selection of a target object. (e) Response generation, yielding the color of the selected target.


(a)
(b)
"red"
"green"
"blue"
boost
target
reference
"left"
"right"
"above"
"below"
object-centered
(c)
"red"
"green"
"blue"
target
reference
"left"
"right"
"above"
"below"
boost
boost
object-centered
(d)
"red"
"green"
"blue"
boost
target
reference
"left"
"right"
"above"
"below"
object-centered
(e)
"red"
"green"
"blue"
boost
target
reference
"left"
"right"
"above"
"below"
boost
object-centered

color node for "green" and boosting the target field. In the same way as in the first task, this combination of inputs causes a selection decision to take place in the target field, and a peak forms for the location of the green item.

Next, the spatial relation nodes are globally activated, and the object-centered field is boosted (Figure 6.8c). The spatial relation nodes project through their semantic weight patterns to the object-centered field, and the boost input to that field ensures that the resulting activation pattern is reflected in the field output and is projected to the transformation field. This is largely analogous to the second step in the previous task, with the difference that no single spatial term is specified here. Instead, a superposition of all spatial semantic patterns is induced in the object-centered field.

Now, the reference field is boosted. The reference field receives localized input from the space-color fields that reflects salient object locations, and broad input from the object-centered field via the transformation mechanism. The location of the selected target object is inhibited by input from the target field. The boost input initiates a competition within this field between the possible reference object locations. On the one hand, the outcome of this competition depends on the relative saliency of the visual stimuli. On the other hand, it is also biased by the input from the object-centered field, in such a way that those object locations that provide a good fit with the semantic pattern of any spatial term gain an advantage. Note that in the coupling between reference field and object-centered field via the transformation mechanism, the directions are inverted. For instance, a reference field peak to the left of the target object induces activation in the right part of the object-centered field (because it indicates that the target is to the right of the referent).

In the present scenario, the potential reference objects are nearly equal in saliency, but the blue stack of blocks provides a better spatial term fit (the green highlighter is almost exactly to the left of it, while it is to the right and slightly below the red stack of blocks). Consequently, a peak forms in the reference field for the location of this blue stack of blocks (Figure 6.8d). As soon as this peak begins to form, it also projects input back to the object-centered field, which increases the activation levels in the left part of that field, and consequently strengthens the spatial relation node for "left". This in turn adds further support for the selection of the blue stack of blocks in the reference field. Effectively, a coupled selection takes place for an object

---

Figure 6.8 *(preceding page)*: Evolution of activation patterns for the task "Where is the green item?" (a) Visual input. (b) Specification of the green item as reference object. (c) Activation of all spatial relations. (d) Coupled selection of reference object location and spatial relation. (e) Response generation.

location and a matching spatial relation, with both parts of the selection reinforcing each other. This process is analogous to the coupled selection of object location and surface feature in the biased competition model, although the connection patterns that mediate the coupling are more complex in the present model.

After a reference object location has been selected and the matching spatial relation node has been activated, the system's response is generated in the final step. To this end, both the set of color nodes and the set of spatial term nodes are boosted (Figure 6.8e). The boost of the color nodes produces the object identification in the same fashion as in the previous task. However, in the present case, the color of the reference object is produced as a response (because the reference field is still boosted), while in the previous task the color of the selected target was determined. The boost of the spatial term nodes simultaneously selects a single spatial term, based on the activation levels of the spatial relation node. The selection of the "blue" color node and the spatial term node for "left" yields the response "to the left of the blue item", which is one of the two possible valid descriptions for the green item's location in the given visual scene.

6.3.5 Statistics of reference object selection

In the above task, the system has the freedom to choose between different potential reference objects to generate a spatial description. It does so on the basis of both object saliency and match of the relative position to the available spatial terms. The reference selection behavior in humans in comparable situations has been investigated by Carlson and Hill (2008). In one of their experiments, the authors presented their participants with different arrangements of two or three items (images of real-world objects) on a uniform background (Figure 6.9). One of the objects was designated as the target for a spatial description, or *located object* in the authors' terminology (abbreviated L). The participants were asked to provide a spatial description for this object by completing the phrase "The *located object* is ...."

The second item that was present in the array was always larger and of a different shape than the designated target object. In most trial conditions, there was a third item that was of similar size and shape as the target. Carlson and Hill referred to the larger item as the reference object and to the other item as distractor, based on their roles in a previous experiment. I will use the abbreviations R and D based on these designations to refer to these objects in the following. Note, however, that participants in this experiment were in no way instructed or encouraged to use the larger object R as the reference object in their response.

The experiment was emulated in the model using arrangements of differently colored squares instead of real-world images as stimuli, in order to accommodate for the model's simple visual system. The smaller objects L

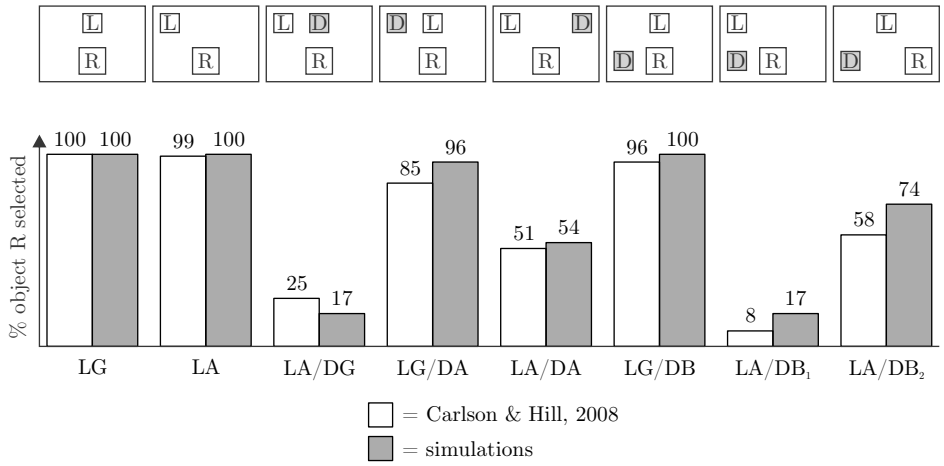


Figure 6.9: Reference object selection in an experiment of Carlson and Hill (2008) and in the DNF model. The top row shows different spatial arrangements between an object L, whose location is to be described, and two possible reference objects R and D. The bar plot shows the frequency with which the more salient object R was selected as reference object for the generation of a spatial description for each arrangement.

and D were represented by squares of 10×10 pixels, the more salient object R by a square of 14×14 pixels, approximating the size relations in the experimental study. The input image was divided into an invisible 5×3 grid, with objects centered in the grid cells of the top and bottom rows, and the leftmost, center, and rightmost column. The stimulus arrangements for the different experimental conditions are shown in Figure 6.9. Conditions were labeled according to the placement of object L in the good (LG) or acceptable (LA) "above" regions relative to the larger reference object (R), and the placement of object D in the good (DG), acceptable (DA), or bad (DB) "above" regions. This naming convention is again based on an earlier experiment in the same study, and the use of the term "above" in the response was not required or encouraged.

The model performed 100 trials for each condition in which it produced a description of the target object by selecting both a reference object (identified by its color) and a relational spatial term in the fashion described above. Random noise was added to all field and node activations in each step to obtain stochastic results. In all cases, the response of the model was a valid description of the target object location. Note that in the case of a target diagonally displaced from the selected reference object, two spatial terms were considered correct, for instance both "above" and "left" in condition LA. The key experimental measure was the proportion of trials in which

the more salient object R was used as reference object in the response. The results from both the model and the experiment are shown in Figure 6.9.

The model provides a good fit of the results from the experimental study in all conditions. In the first two conditions (LG and LA) with only two objects in the scene, the results are unsurprising. The only available object beside the target is always selected in the model's reference field, and in turn drives the selection of a matching spatial term. In the next condition, LA/DG, the target object L is directly to the left of object D, and diagonally to the left and above object R. Object R does not provide a good match to the semantic template of any available spatial term in this case, and consequently it is only chosen in a small proportion of trials despite its greater visual saliency (25% in the experiment, 17% in the model). The model produces the same result in condition LA/DB$_1$, where object L is exactly above object D and diagonally above and to the right of object R. In the experimental results object R is chosen less frequently as referent here (8% of trials), which may indicate an asymmetry in choosing horizontal versus vertical relational spatial terms in humans which is not captured in the model.

In condition LG/DA, objects D and R provide equally good fits for one spatial term ("right" and "above", respectively). In this case, the more salient object R is chosen in the large majority of trials in both the experiment (85%) and the model simulations (96%). Condition LA/DA is similar to condition LA/DG, but now the distance between objects L and D is larger. This is relevant for the model behavior because the semantic weight pattern is distance dependent, with slightly decreasing weights for larger distances between objects. This is consistent with the boundary cell semantic distributions proposed by O'Keefe (2003). Object D still yields a better spatial term match than object R, but since object R has greater visual saliency, the selection behavior is largely balanced in the model (54% selection of object R). This is consistent with experimental data for this condition (51%).

In condition LG/DB, the target object L is located directly above object R and diagonally above and to the right of object D. Both spatial term match and visual saliency favor object R to be selected as reference object in this condition, explaining the results in the model (100%) and the experimental study (96%). Finally, in condition LA/DB$_2$, the target object L is located diagonally above both D and R (to the right and left, respectively). As in condition, LG/DA, object R gets chosen more often than D due to its greater visual saliency (74% in the model, 58% in the experiment). Note, however, that the advantage of R over D is significantly smaller here than in condition LG/DA, even though within each of these conditions, the spatial term match is equally good for R and D. The difference is that the overall spatial term match is higher in condition LG/DA. This leads to higher overall activation values in the reference field in that condition, and a faster

and more stimulus-driven selection. In condition LA/DB$_2$, both objects only match poorly with the spatial terms, leading to lower overall activation levels in the reference field during the selection process and a stronger influence of random noise. This larger influence of noise produces a more balanced selection behavior.

It should be mentioned that the overall level of random noise in field activations was treated as a free parameter in these simulations, and was chosen to fit the experimental data. It was not, however, varied between different conditions. As such, it only determines a global level of stochasticity. Without any noise, the outcome in each condition would be deterministic without any graded selection preferences. For increasing noise levels, the impact of activation differences induced by the properties of the stimuli becomes less relevant compared to the random fluctuations in activation, and the probability for different objects to be selected becomes more similar. The pattern of *relative* selection proportions that is visible in Figure 6.9 cannot be produced by changing the noise level. Instead, it reflects how similar the activation levels for different potential reference objects are during the selection process, which, in turn, is determined by the properties of the stimuli and the internal parameters, such as the definition of the semantic weight patterns.

6.4 Discussion

In this chapter, I have presented a DNF model of relational spatial language behaviors. This model demonstrates that the same basic mechanisms that were introduced in the context of visual processing can also form the basis of more cognitive tasks, such as the generation and assessment of spatial phrases for a given visual scene. The model combines the basic mechanism for space-feature association with a reference frame transformation to determine the relative position of one object to another, and it then maps this relative position onto a discrete symbolic representation that stands for a specific spatial term.

The DNF model can solve a variety of tasks without any changes in its parameters or connection patterns. The type of operation that is performed is determined by a series of control inputs that globally activate certain elements of the architecture. These control inputs determine the effective flow of information in the dynamical system. They build on the fact that all projections within the model are implemented bi-directionally, and that the core mechanism employed here—space feature association and reference frame transformation—can flexibly work in different directions depending on present activation patterns.

The core problem that is addressed by the model is an instance of a central problem in cognitive science: The *grounding* of abstract concepts, that

is, linking symbolic representations for concepts like "above" to concrete sensory or motor representations (Roy, 2005). The model follows an approach that is consistent with the concepts of embodied cognition: It does employ discrete nodes (that may be classified as a symbolic representation) at the input and response stages, but the actual processing underlying the assessment of a spatial relation is performed on metric representations of space and visual features, which are directly linked to sensory input. This is an important contrast to typical symbolic models of processing, where the complete sensory information is assumed to be transformed into a purely symbolic representation, on which the cognitive operations are then executed (e.g., Shastri, 1999).

The DNF model also contrasts with mathematical models of spatial language such as the AVS model of Regier and Carlson (2001). Several aspects of this earlier model are inspired by neural processes, namely the mechanisms of attentional weighting and the computation of an estimated direction vector as a weighted sum over a large number of individual vectors. This latter approach reflects the estimation of metric values in neural population codes, as described for instance by Georgopoulos et al. (1986) for determining the movement vector in the planning of reaching movements. However, the AVS model does not aim to capture the actual neural processing underlying the assessment of spatial relation, and instead computes spatial relation ratings in a simple algorithmic formulation.

The DNF model goes significantly beyond the scope of this earlier work by proposing a detailed neural process, covering task components from the initial representation of the visual image all the way to the selection or assessment of a spatial term. Importantly, the transition from the sensory representation to the verbal representation is not achieved in a single step, but is an active and composite process that requires the combination of several basic operations.

6.4.1 Sequential processing steps

In each of the tasks performed by the model, I have described a series of distinct processing steps, such as the localization of the target object, localization of the reference object, and selection of a spatial term. It is important to keep in mind that these steps still emerge from a continuous change of activation patterns in the neural fields, as described by the differential equations. Their macroscopically discrete nature arises from instabilities in the field dynamics—transitions to a new stable state, typically through the formation of an activation peak—and not from the use of discrete processing steps at the microscopic level.

Moreover, the processing is not strictly sequential in the sense that each operation is always applied to one object at a time. In many instances, the system makes use of the inherently parallel representations provided by the

neural fields. For instance, when selecting a target object based on a given reference object and a spatial relation, the system assesses the match of all possible target locations to the spatial description in parallel during the central step of this task. The parallel processing is even more striking in the open-ended task of generating a spatial description. Here, different combinations of reference object and spatial term form implicitly in the coupled activation patterns and they directly compete with each other, without any need to assess each combination individually. The sequential and selective processing is used when it is required to solve a specific problems, namely different instances of the binding problem (see below). This combination of parallel and sequential processing matches the approach used in the scene representation model.

The sequence of steps is induced in this model by a series of global control inputs, synchronized with the corresponding content-carrying inputs (colors and spatial terms from the verbal task). The different sequences of these inputs form a kind of program executed in the architecture. They all use the same basic operations, but by using them in different combinations and in different temporal orders, they solve qualitatively different task. For instance, the selection of the target objects is always achieved by boosting the target field (which will then form a single peak due to the competitive interactions). In two of the tasks, the selection is biased by specifying an object color and thereby modulating the visual input that the target field receives. In another task, however, the target object selection is only done after both a reference object and a spatial term have been set, and no object color is specified. In this case, the selection is biased toward a certain spatial region. This approach strongly depends on a bidirectional coupling between the elements in the architecture, a feature that is also prevalent in the form of reciprocal connections in biological neural systems (Lamme et al., 1998).

The system of different input sequences contrasts to some extent with the approach used in the scene representation model. That model also executed sequences of processing steps to execute change detection tasks on visual scenes, but the underlying mechanism to drive these sequences was more strongly integrated into the architecture itself. This reflects the somewhat more stereotyped mode of operation in that model. Even though the scene representation architecture solved a variety of tasks, it always performed the same basic operations to sequentially inspect items. Different behaviors within the change detection task were produced by higher-level adjustments, such as tuning down the spatial coupling between the retinal and the scene level. This more integrated and stereotyped form of creating sequences of operations appears more appropriate for the comparatively low-level and highly automated task of scanning a scene, whereas the spatial language behaviors can be considered more cognitive tasks that require greater variation in the sequential organization of elementary operations.

A significant limitation of the sequential mechanism in the spatial language model compared with the scene representation architecture is a lack of autonomy. The global control inputs that drive the transition to the next step are applied externally with a fixed order and timing, chosen in such a way that the transition of the system to a new attractor state is almost certainly completed before the next input is applied. This makes the system slow and inflexible. Later extensions of the model have addressed this problem and added neural mechanisms for the autonomous control of the processing steps. Van Hengel et al. (2012) have combined the spatial language architecture with a neurodynamic model of serial order (Sandamirskaya and Schöner, 2010). In this system, a series of dynamic nodes provides the control inputs to drive the processing steps. Connections between the nodes implement the sequential order of operations, and a condition-of-satisfaction mechanism is used that detects when each processing step has been completed (e.g., when a peak has formed to select a target object location). This triggers the transition to the next processing steps, so that the whole sequence of steps can be traversed with a timing adjusted to the requirements of the specific task.

In a further development of this model, a variant of the spatial language architecture presented here was combined with a neurodynamic model of behavior organization (Richter et al., 2014a,b), which is itself an extension of the serial order model mentioned above. This system adds a working memory representation for the content of the verbal task, and employs a system of rules implemented in specific connections between dynamic nodes to organize the processing steps. The ordering of the processing steps then emerges autonomously based on these rules, and may vary between different instances of the same task. These extensions show that the basic mechanism of generating complex and varied behaviors through a series of instabilities, driven by control inputs, is a viable route to create behavioral flexibility in an autonomous neurodynamic system.

6.4.2 Variable binding

The core reason that makes sequential processing necessary in this model is the binding problem. This appears here both in the form of feature binding, which was a key issue in the previous chapter, and in a new form known as variable binding. The problem of feature binding comes up in the spatial language tasks when performing a visual search for an object, where a given color has to be associated with a spatial location. But this alone does not fully explain the need for sequential processing. In a task like "Where is the green item relative to the red one?", it is still conceivable that a visual search could be performed for both red and green objects simultaneously, and would yield two locations as a result.

But critically, the two objects referenced in the verbal task are not inter-

changeable. They take different semantic roles—those of target and reference object—and these roles are reflected in the different connection patterns of the corresponding spatial representations in the reference frame transformation mechanism. This is the problem of variable binding: The two objects, specified in the verbal phrase through their surface features, must be bound to their semantic roles in that phrase. In previous robotic DNF models of spatial language (Lipinski et al., 2009; Sandamirskaya et al., 2010), this problem was avoided by using separate processing paths to select the target and the reference object in the visual scene. This does not seem biologically plausible, however, since the localization for both objects must be achieved by a single visual system in the brain.

The approach used here to solve the problem of variable binding is in principle the same as employed in the previous chapter to address the problem of feature binding. On the one hand, there is a form of conjunctive coding in the form of the separate target and reference fields. These fields form an explicit representation of the combination between semantic role and object location. On the other hand, the system uses the sequential processing of individual items to establish the binding between object identities and their semantic roles in the first place. This allows the model to use only a single general-purpose visual system that provides the space-feature association for both objects in a spatial relation, without explicitly dealing with the objects' semantic roles. The assignment of an object localized by the visual system to a specific semantic role is then accomplished by the timing of the control inputs that boost activation levels in either the target or the reference field.

There is experimental evidence for the notion that humans likewise employ a sequential processing of target and reference object when assessing spatial relations. Franconeri et al. (2012) had subjects perform a task in which they had to determine the spatial relation between two simple colored stimuli, and used EEG measurement to estimate where the subjects' attention was directed. They found a clear shift of attention during the task, and explain this with the need to select one object location at a time, based on the Feature Integration Theory and consistent with the approach presented here. The authors suggest that the spatial relation is determined directly from the direction of this attentional shift, but do not propose a model for this process (and it is not clear that determining the shift direction is actually different from determining the relative position between start and end point of the shift). Other experiments, using both eye tracking data (Burigo and Knoeferle, 2011) and behavioral cuing (Roth and Franconeri, 2012) provide further support for a sequential processing of objects to determine their spatial relation.

Notably, the spatial language model also uses the spatial representations in the target and reference fields as a form of pointer to objects in the

visual representation, in the same way as used in the scene representation model. While any objects provided in the verbal phrase are identified by their surface features, only their spatial locations are actively retained once it has been found in the visual scene. But this location can be used to retrieve the features of an item when required, as shown in several of the tasks. In the present model, this reduction to object location is motivated in particular by the fact that the location is the relevant aspect of an object for judging spatial relations. But it also provides an additional example of the concept introduced in the previous chapter: A spatial representation can act as a form of pointer in neural architectures to select and individuate objects from more complex representations that capture detailed object properties.

One may ask whether a mechanism of variable binding that uses a separate spatial field for each semantic role is feasible beyond the scope of relational spatial language, or would require an excessive number of such fields. There is reason to believe, however, that the roles of reference and target objects for a spatial relation are truly essential for many tasks, and that there is only a limited number of other comparable semantic roles that would require a separate treatment. For instance, relative positions and relative movements between two objects are central for classifying object-oriented actions (such as touching, pushing, hitting). A neurodynamic model of action recognition by Fleischer et al. (2013) uses an object-centered spatial representation analogous to the object-centered field in the present model as a central element. Moreover, neural representations of space in an object-centered frame of reference have been found in the parietal cortex (Chafee et al., 2007), further supporting the need for specialized representations to determine spatial relations.

6.4.3 Limitations and future extensions

An obvious limitation of the present model is the very simple visual system that only captures the distribution of one surface feature, namely color. For use in real-world robotics tasks, for instance, this system would have to be capable of actual object recognition. In principle, it is easy to incorporate a more elaborate visual system into the spatial language model. It only has to support two basic operations: It must be able to localize an object in a visual scene, given an object identifier (a label or a feature description); and, conversely, it must be capable of identifying an object at a location selected in the target or reference field. A basic segmentation of a scene is also needed (to determine candidate objects that can be selected as target or referent), but may be provided by a separate system. An object recognition architecture based on the DNF framework that fulfills all these requirement has been presented by Faubel and Schöner (2009).

Another critical limitation in the visual system is that it is purely input driven (with weak stabilization from lateral interactions), and that it

uses a retinal or image-based reference frame. It is therefore not capable of compensating for image shifts due to gaze changes, which frequently occur during spatial language tasks (Burigo and Knoeferle, 2011), and it does not retain any information about the objects in a scene if the visual input is occluded or turned off.

A more robust system could be constructed by integrating the spatial language model into the scene representation architecture. The scene representation in working memory can provide a stable, gaze invariant substrate to locate objects and determine spatial relations between them. The working memory fields over space and surface features have the same basic structure as the space-color fields used in the spatial language model, and it has already been shown for the robotic version of the scene representation model how "queries" can be processed on these fields—a form of visual search in working memory to localize objects that match specific surface feature values (Zibner et al., 2011b). Initially, the spatial language model did not make use of gaze invariant working memory representations because the scene representation model was only implemented after the spatial language model. But integrating these two models has the potential to produce a more powerful architecture with an extended behavioral repertoire.

Chapter 7

General Discussion

In this thesis, I have presented a series of neurodynamic models that address different problems from the field of active vision, and that cover the range from perceptual processing to cognitive operations. The models are formulated within the theoretical framework of Dynamic Field Theory, and their implementation is consistent with principles of embodied cognition (Wilson, 2002; Riegler, 2002). Cognitive operations in these models emerge from sensorimotor processing, and both build on the same representations and neural mechanisms.

The DNF models are based directly on neural population dynamics, and they aim to capture the real-time evolution of neural activation patterns that underly behavior generation. Their mode of operation is inherently parallel, with activation values in different DNFs and at different positions in feature space within each DNF changing simultaneously in response to external stimuli and internal interactions. All interactions in the models can be described by fixed excitatory or inhibitory connection patterns, reflecting synaptic connectivity in biological neural systems.

The results that I have presented further support the biological plausibility of the DNF models, and highlight how these models establish a link form neural processes to overt behavior. I have shown that the DNF models can account for many characteristics of human behavior—often in quantitative detail—with respect to the planning and execution of saccadic eye movements, human working memory and change detection performance, and use of spatial language. Processes in the models can be directly linked to behavioral measures in psychophysical experiments. In the case of saccadic eye movements, for instance, different reaction times in the model result from differences in the time it takes to resolve a competition between possible movement plans within neural representations. Moreover, for the case of peri-saccadic remapping, I have shown how the DNF model can directly account for neural activity patterns obtained in experiments with macaque monkeys (for a DNF model addressing both behavioral and electrophysio-

logical results in a combined fashion, see Klaes et al., 2012).

All theoretical models presented in this thesis are process models, that can actually perform their specific tasks and generate responses based on the sensory inputs they receive. As a restriction, it must be noted that the models were for the most part operated only in simulated environments. They did receive sensory input in a format consistent with biological sensory systems (rather than abstract symbolic input), but it was often simplified and idealized. These simplifications were adopted in order to allow an easier exploration of the general principles that may be used for autonomous processing in neural architectures. An extrapolation to real-world applications may therefore still be a challenging step, although several robotic models that were developed in parallel with the work presented here demonstrate that this step is indeed feasible (Zibner et al., 2011a; Lipinski et al., 2009).

The larger theoretical questions addressed by these models are the following: How can a parallel, continuously operating neural system perform complex cognitive operations that comprise distinct processing steps? And how can such a system, that is characterized by fixed synaptic connections, apply cognitive operations flexibly to different objects (either perceived sensory stimuli or internal object representations)? These questions are made concrete in the examples treated in the last two chapters of this thesis. In the model of scene representation and change detection, the system sequentially attends to individual items in a visual scene to memorize them or compare them to previously memorized items. In the spatial language model, different stimuli in a visual scene are sequentially assigned to different grammatical roles in order to ground a verbal spatial phrase.

Both of the core questions—how to generate processing steps and how to specify targets for cognitive operations—arise in the DNF models to a large part due to their autonomous mode of operation. In the majority of neural models in the literature, the different arguments for the operation to be performed are simply provided via separate input channels at the time of initialization. The neural model then processes these inputs, either in a simple feed-forward pass or through an iterative process with recurrent interactions, and produces a result. Examples for such models are found in Zipser and Andersen (1988) for the problem of spatial transformations, and Denève and Pouget (2003) in the field of spatial language.

The DNF models, in contrast, aim to describe the continuous processing that is required in biological neural systems, rather than isolated operations. In the earlier examples of the biased competition model and the model of peri-saccadic remapping, this is expressed in a largely reactive mode of operation. These models produce behaviors (namely saccadic eye movements) and update their activation states in response to external stimuli, which can appear at arbitrary times. In contrast, the DNF model of spatial language has to generate a more cognitive operation that results in a specific response.

But the arguments for this operation—such as the objects whose spatial relation should be assessed—are not simply provided through separate channels. Instead, they must be extracted from the continuously present visual input. This brings up the different forms of binding problems in the DNF models, and generates the need for sequential processing.

7.1 Organization of discrete processing steps

The starting point for generating ordered sequences of processing steps from continuous neural dynamics is provided by a core principle of Dynamic Field Theory, namely its focus on attractor states and instabilities between them. The basic instabilities in DNFs mark discrete points in time at which qualitative changes occur in the continuously changing activation distributions (Schneegans and Schöner, 2008). These qualitative changes in activation patterns can be linked to the completion of elementary mental operations—the detection of a new stimulus, a selection decision between multiple alternatives, or the formation of a working memory representation.

The DNF model of biased competition directly builds on this mechanism. Its primary behavioral output—the saccadic motor signal—is at its core the result of a selection decision between different visual stimuli. But the model already goes beyond the basic concept of an instability in a single DNF. It demonstrates how a coupled selection decision can take place in a representation that is distributed over multiple interconnected DNFs, and how such a distributed decision forms the basis for new behavioral mechanisms. The architecture allows contributions from the different fields to influence the competition process—namely spatial biases and color biases—and integrates them to generate a simple form of goal-directed behavior.

The DNF model of peri-saccadic remapping has a more complex architecture, and during saccadic eye movements, it shows an intricate sequence of state transitions. New activation peaks form, some transiently and some sustained, and existing activation peaks decay or are actively suppressed, first in the fields of the gaze update system, then in the remapping system. This sequence of state transitions in the continuous field dynamics ultimately results in the macroscopically discrete event of remapping, that is, a shift in the retinocentric representation of object locations.

The stabilizing effects of lateral interactions in the fields play a critical role in this process. They drive the activation patterns within each field toward a stereotypical shape and project a relatively uniform signal to downstream fields, filtering out any fluctuations and irregularities in the activation patterns. This greatly simplifies the construction of large interconnected architectures, since the input to each field can be assumed to be a set of uniform activation peaks at all times, independent of how these peaks themselves were induced. And it makes it possible for the system to process

different contents—different stimulus positions and saccade metrics—while still reliably following a fixed pattern of state transitions.

While the model of saccadic remapping produces relatively intricate series of state transitions in this way, these transitions are still triggered by an external input (the saccade signal), and progress through the architecture in a mostly linear fashion (albeit with strong interactions and back-coupling). The scene representation model introduces a new mechanism that enables it to sequentially inspect the items in a visual scene. For each individual item, a series of state transitions occurs in a similar way as in the remapping model, beginning with the coupled attentional selection of an item and leading to the formation of sustained activation peaks in the working memory fields. But now additional structures and connections are introduced that serve specifically to trigger a re-initialization of this process with the selection of a different item.

The central concept here is the specification of a *condition of satisfaction* (CoS) that indicates the completion of a processing step (or a series of steps). This idea has been employed in previous DNF models to produce sequences of arbitrary actions with a flexible timing (Sandamirskaya and Schöner, 2010) and to organize sequences of overt behaviors for goal-directed action (Richter et al., 2012). In the scene representation model, this CoS is the formation of an activation peak in the scene attention field which signals that the memorization or comparison for the currently selected item is complete. Again, the lateral interactions play a crucial role here, since they ensure that a qualitative change occurs in the field's activation state, which can be detected unambiguously as CoS. The concept of a CoS allows a new level of description, in which a processing step is defined as the state transition or series of state transitions required to trigger a specific CoS. The scene representations model goes through a series of such processing steps, with the initial trigger given by the presentation of an external stimulus array, but the further progression with its attentional shifts from one item to the next driven only by its internal dynamics.

Finally, the spatial language model demonstrates how such processing steps can be combined flexibly to generate a variety of cognitive behaviors in a single neurodynamic architecture. Due to the strong multi-directional connectivity in this model, the effective direction of activation flow between fields can be determined to a large degree by modulatory inputs. These inputs globally raise the activation levels in certain fields and thereby induce instabilities, typically through the formation of an activation peak. The specifics of each instability, namely where a peak forms, is determined by the latent inputs to the field. The global control input, however, determines the timing when an instability occurs, and can produce qualitatively different final activation states by varying the order of instabilities within the architecture.

In the original DNF architecture for spatial language presented in this thesis (based on Lipinski et al., 2012), the system is implemented with only limited autonomy. The control inputs are applied in fixed order and with predefined timing, and no explicit conditions of satisfaction are defined for the individual processing steps. Subsequent work has shown, however, that an additional layer of dynamic nodes can take over the role of these fixed external inputs, apply them flexibly while ensuring the task-dependent constraints on their order, and detect the completion of individual processing steps from state transitions in the fields of the architecture (Richter et al., 2014a,b). A form of cognitive program (Ballard et al., 1997) can then be defined as a set of processing steps and constraints on their ordering, implemented through sets of interconnected nodes. Each processing step specifies one or more control inputs applied to the DNF architecture, and an associated CoS. This does not explicitly specify the operation to be performed, as it would be done in a step of an algorithm, but it can still ensure implicitly that the desired operation is executed within the neurodynamic system.

7.2 Spatial pointers

The second challenge in implementing complex cognitive operations in a parallel neural system lies in flexibly specifying the targets of those operations. This can be likened to passing arguments to function in a computer program. Here, it is often realized efficiently by passing a pointer or reference to a more complex data structure. One common theme of the models presented in this thesis is the use of spatial representations to take a role comparable to such pointers. Similar approaches have previously been suggested in conceptual models. For instance, Ballard et al. (1997) suggested that spatial attention or overt visual fixation can single out an object in the world, and its perceptual representation can thereby be selected as the target of a cognitive operation.

The starting point for using spatial representation as a form of pointer is again a basic characteristic of DNFs, namely the peak of activation as a stable state in the field dynamics. The activation peak specifies a discrete value within the continuous activation distribution. It is stabilized against fluctuations in the field input, and can serve to represent a discrete object location. The basic field dynamics allow multiple activation peaks to coexist and support tracking of changing inputs, which directly provides a model of visual tracking for moving stimuli (Spencer et al., 2012). Population code representations of space that are consistent with such DNFs are found throughout the cortex, and play important roles in attention and movement planning (Colby and Goldberg, 1999).

The DNF model of biased competition demonstrates the basic association mechanism through which an activation peak in a spatial field can be used to

access non-spatial features of an object. In the model, location and surface features of multiple visual stimuli are represented in a combined fashion in a multi-dimensional visual sensory field. Through the combination of a spatial ridge input and a read-out projection, the surface feature of a specific item can be determined given its location. This mechanism is extended to multiple surface features in the perceptual part of the scene representation model. The extension illustrates why the spatial dimension is a suitable candidate to be used in the role of a pointer, more than any other feature dimension: Features that belong to the same object naturally appear at the same location in visual space. Space can therefore be used as a binding dimension between different surface features, and can mediate the coupled selection over all feature dimensions. Multiple separate feature maps over a shared space avoid the curse of dimensionality that would occur when using a full representation over the combination of all feature spaces. The use of space as a binding dimension is consistent with numerous existing models of visual processing and visual attention, both conceptual (Treisman, 1988) and computational (Hamker, 2005b), and is supported by psychophysical studies (Nissen, 1985).

But the biased competition model also touches on one of the key problems of using spatial pointers. Saccadic eye movements constantly shift the visual image, and make the retinocentric location information from the previous fixation obsolete. The model of peri-saccadic remapping addresses this problem. As solution, it proposes a transformation of location information into a gaze-invariant frame of reference, while still keeping it continuously linked to the current retinal scene. The proposed system can transform the locations of multiple visual stimuli simultaneously, so that is compatible with spatial fields containing multiple activation peaks.

With the spatial transformation mechanism in place, it becomes possible to accumulate visual information over multiple fixations, and to form stable object representation in working memory while still using space—now in a gaze-invariant reference frame—to bind different surface features together. This is implemented in the DNF model of scene representation. The model provides a working neural implementation that realizes the core characteristics of scene working memory proposed in conceptual form in the influential Object File Theory (Kahneman et al., 1992). Surface features and locations for a limited number of visual items can be stored in a bound form in the coupled working memory fields over feature and space, without requiring an addressable general-purpose memory as the term object file may suggest. These bound working memory representations may then be re-activated when the object location is re-attended in the visual scene, due to the coupling between different spatial fields through the reference frame transformation. This has been demonstrated in the task of feature location change detection, and it implements the kind of access to the memory

representation via a spatial pointer as proposed in the Object File Theory.

The coupled spatial representations in the scene representation architecture also fulfill the key requirements of the spatial indexes in the theoretical work of Pylyshyn (2001). They individuate a limited number of objects, keep track of them over gaze changes, and can be used to find them again in the visual scene. But importantly, the spatial representations at the scene level can also act in a more abstract role when they are decoupled from the retinocentric reference frame and thereby loose their fixed relationship to any physical space. In that condition, demonstrated in the task of feature conjunction change detection, the spatial dimension serves as an abstract binding dimension for the surface features of an object. This more abstract use of space emerges directly from its concrete use to reflect object locations, simply by dampening certain connections in the DNF model.

The spatial language model has then demonstrated the use of space in another form of binding, namely variable binding. Separate fields over space are used to associate an object in a visual scene to its semantic role, namely target or reference object in a spatial phrase. These spatial representations are then used to assess the spatial relations between objects, but they can also serve in the role of spatial pointers again, to direct attention to a specific object in the visual scene and retrieve its surface features for identification. This was demonstrated in the response generation for the task of the type "What is above the blue object?" The assignment of objects to different semantic roles via spatial fields, and the subsequent use of these fields to perform spatial operations, can be viewed as analogous to passing arguments to a function. The flexibility of this system—being able to select any objects in a scene as target or referent—is achieved by the sequential attentional selection. The stable activation states of fields are again critical here, since they allow the association of an object to a semantic role be retained while another object is selected by attention.

The spatial representations thereby fulfill several roles that are analogous to pointers in computer architectures. They do not, however, implement an actual pointer mechanism that allows access to a location in an addressable general-purpose memory. The function of the spatial representations is more restricted, and they require special neural structures to act in a pointer-like fashion. They can only be used to access perceptual or working memory items in neural representations that directly reflect spatial position, or that are linked to such spatial representations (e.g., via a space-feature association mechanism). This reflects an adherence to neural principles, in particular modality-specific representations through population codes and fixed synaptic connections among them.

There are other theoretical models that have proposed implementations of actual pointer mechanism in neural architectures to address the problem of variable binding. The model of Barrett et al. (2008) includes neural repre-

sentations that hold an address to an item in another, larger representation. This mechanism is used to assign specific content to different semantic roles in a system of knowledge representation and reasoning. A different approach is the theory of binding through synchrony (Von der Malsburg, 1999), which was discussed in some detail in Chapter 5. This theory proposes a special neural mechanism which allows object features that are represented in a distributed fashion across different neural populations to be linked to each other directly.

While these models do provide elegant solutions to avoid most problems associated with feature binding and variable binding, their biological plausibility is questionable. No actual pointer representations have ever been found in neural systems, and the evidence for a key role of spike synchrony in neural processing is much disputed (Palanca and DeAngelis, 2005). In contrast, the limitations of the spatial pointer approach advocated in this thesis can in fact be taken as an argument for the plausibility of this approach, since they match limitations observed in human behavior. One core property in both the mechanism of feature binding and of variable binding presented here is the need for sequential processing of individual items. This is consistent with a host of findings both in scene perception and working memory (Treisman, 1988) and in the assessment of spatial relations (Franconeri et al., 2012). The DNF models explain how the need for this sequential processing arises, and also provide solutions to control the sequential processes through neural mechanisms.

Moreover, the mechanism of spatial pointers is made plausible by the fact that it illustrates how the functions of feature and variable binding can emerge from basic principles of sensorimotor processing, without requiring novel and special-purpose mechanisms like neural synchrony. The role as pointer-like elements for spatial representations is derived directly from the fact that space is a common dimension between neural feature maps. And besides their role as pointers, the spatial representations in the DNF models always retain their original function of representing concrete content, namely the locations of objects. This exemplifies one core tenet of the embodied cognition stance, that abstract and cognitive functions emerge from sensorimotor processes.

7.3 Outlook

The models discussed in this thesis have shown how spatial representations can be used in a generalized fashion, acquiring additional functionality that is not strictly linked to representing locations. This can be taken further in several directions. As a first step, spatial representations can be used in a fashion that is no longer linked to a concrete physical space in the world, while still retaining its spatial meaning. To some extent, the spatial

dimension for the working memory fields in the scene representation model was already interpreted in this way—as not being strictly allocentric, but tied to a more abstract scene reference frame. This would allow the model, for instance, to compare a sample scene viewed on one computer screen with a test scene viewed on another screen, or one viewed in the real world. This was not fully worked out in the model, but with the mechanism for variable reference frame shifts (that can also be used to determine the alignment between two scenes), it is relatively straightforward to imagine how this may be implemented.

A direct extension of this approach is then to use the same structures also for scenes that are not at all linked to any space in the world, but arise in the imagination alone. This is the idea behind an envisioned combination of the scene representation system and the spatial language model. Such a combination was already proposed in the previous chapter to stabilize the visual representation for the assessment of spatial relations, but it may also be used in another way: to generate an internal scene representation purely on the basis of a verbal descriptions (such as "there is blue square to the right of a green triangle") without any visual input. Due to its separate spatial and feature pathways, the model is already set up to integrate location and feature information for new objects given separately in a verbal statement, and the core of the spatial language system can be used to select appropriate locations for the objects. To operate autonomously, the model would again have to generate complex sequences of individual processing steps that can be flexibly combined. This could be based on the same mechanisms as used in the later versions of the spatial language model, but may have to be even more elaborate.

Such a model would then offer the opportunity to also address processes of mental reasoning. When a scene representation has been built, the system is able to answer questions about spatial relations that were not given explicitly in the verbal description. The empirical and theoretical work of Knauff (2013) has provided evidence that humans do indeed use such an approach for reasoning about spatial relations, even though it can be susceptible to errors. For instance, humans tend to make certain default assumptions when first constructing the scene in their minds, and do not review these assumptions when reasoning based on that scene. Knauff proposed an algorithmic model of the steps to build up a scene in the mind, but the DNF approach may provide an account of how the neural system actually achieves this kind of reasoning.

This approach to reasoning is in line with a number of theories of *grounded cognition*, such as the perceptual symbol systems of Barsalou (1999). This theory emphasizes the importance of simulations in sensorimotor systems for reasoning. Numerous experimental findings support the notion that even tasks that could be solved on a purely abstract level are influenced

by such sensorimotor simulations, for instance in language processing, memory and perceptual tasks (Barsalou, 2008). This approach is in opposition to amodal theories of cognition, which propose that reasoning is performed by operations on abstract symbol systems that are separate and qualitatively different from representations at the perceptual level (Fodor and Pylyshyn, 1988). The DNF model of relational spatial language is already following the grounded cognition approach, in that it performs the assessment of spatial relations on a level that is directly tied to the perceptual processing, without introducing abstract symbols. An extension of this model could extend a neural implementation of this grounded cognition to the level of reasoning.

When the spatial representations can become dissociated from any concrete locations in physical space and are used to construct imaginary scenes, then it is also conceivable that they may be dissociated from the notion of object locations itself, and be used to represent more abstract spaces. Indeed, several theories have suggested that the numerous spatial representations in the brain, which have evolved to represent physical space as a key perceptual dimension, can be co-opted to represent feature spaces that cannot be directly perceived (Anderson, 2010). It has already been mentioned in Chapter 5 that spatial analogies may be the basis for both speaking and reasoning about temporal relations (Casasanto and Boroditsky, 2008). But spatial analogies may be used much more widely, for instance to describe hierarchies in size, value, or importance; to perform approximate arithmetics, using the analogy of a number line (Hubbard et al., 2005; Chen and Verguts, 2012); or to think about abstract classifications as spatially grouping items together. If this is indeed the case, then the models described in this thesis, which address some core problems in active vision, may be the basis for explaining cognitive processes with a much wider scope.

Bibliography

Aloimonos, Y., Weiss, I., and Bandyopadhyay, A. (1988). Active vision. *International Journal of Computer Vision*, 1(4):333–356.

Amari, S.-i. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological cybernetics*, 27(2):77–87.

Andersen, R. A., Essick, G. K., and Siegel, R. M. (1985). Encoding of spatial location by posterior parietal neurons. *Science*, 230(4724):456–458.

Andersen, R. A. and Mountcastle, V. B. (1983). The influence of the angle of gaze upon the excitability of the light-sensitive neurons of the posterior parietal cortex. *The Journal of Neuroscience*, 3(3):532–548.

Andersen, R. A., Snyder, L. H., Bradley, D. C., and Xing, J. (1997). Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual review of neuroscience*, 20(1):303–330.

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and brain sciences*, 33(04):245–266.

Avillac, M., Denève, S., Olivier, E., Pouget, A., and Duhamel, J.-R. (2005). Reference frames for representing visual and tactile locations in parietal cortex. *Nature neuroscience*, 8(7):941–949.

Baddeley, A. D. and Hitch, G. (1974). Working memory. *Psychology of learning and motivation*, 8:47–89.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., and Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20(04):723–742.

Barrett, L., Feldman, J., and Dermed, L. (2008). A (somewhat) new solution to the variable binding problem. *Neural computation*, 20(9):2361–2378.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660.

Barsalou, L. W. (2008). Grounded cognition. *Annuual Review of Psychology*, 59:617–645.

Bastian, A., Schöner, G., and Riehle, A. (2003). Preshaping and continuous evolution of motor cortical representations during movement preparation. *European Journal of Neuroscience*, 18(7):2047–2058.

Bays, P. M., Catalao, R. F., and Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, 9(10):7.

Bicho, E., Mallet, P., and Schöner, G. (2000). Target representation on an autonomous vehicle with low-level sensors. *The International Journal of Robotics Research*, 19(5):424–447.

Blasdel, G. G. (1992). Orientation selectivity, preference, and continuity in monkey striate cortex. *The Journal of Neuroscience*, 12(8):3139–3161.

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, 75(1):1–28.

Bundesen, C., Habekost, T., and Kyllingsbæk, S. (2005). A neural theory of visual attention: bridging cognition and neurophysiology. *Psychological review*, 112(2):291.

Burigo, M. and Knoeferle, P. (2011). Visual attention during spatial language comprehension: Is a referential linking hypothesis enough? In Carlson, L., Hölscher, C., and Shipley, T., editors, *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Springer.

Carlson, L. A. and Hill, P. L. (2008). Processing the presence, placement, and properties of a distractor in spatial language tasks. *Memory & cognition*, 36(2):240–255.

Casasanto, D. and Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition*, 106(2):579–593.

Cassanello, C. R. and Ferrera, V. P. (2007). Computing vector differences using a gain field-like mechanism in monkey frontal eye field. *The Journal of physiology*, 582(2):647–664.

Cavanagh, P., Hunt, A. R., Afraz, A., and Rolfs, M. (2010). Visual stability based on remapping of attention pointers. *Trends in cognitive sciences*, 14(4):147–153.

Chafee, M. V., B, A. B., and Crowe, D. A. (2007). Representing spatial relationships in posterior parietal cortex: Single neurons code object-referenced position. *Cerebral Cortex*, 17(12):2914–2932.

Chen, Q. and Verguts, T. (2012). Spatial intuition in elementary arithmetic: a neurocomputational account. *PloS one*, 7(2):e31180.

Cisek, P. and Kalaska, J. F. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: specification of multiple direction choices and final selection of action. *Neuron*, 45(5):801–814.

Colby, C. L. and Goldberg, M. E. (1999). Space and attention in parietal cortex. *Annual review of neuroscience*, 22(1):319–349.

Collins, T., Rolfs, M., Deubel, H., and Cavanagh, P. (2009). Post-saccadic location judgments reveal remapping of saccade targets to non-foveal locations. *Journal of Vision*, 9(5):29.

Coventry, K. R., Cangelosi, A., Rajapakse, R., Bacon, A., Newstead, S., Joyce, D., and Richards, L. V. (2005). Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In Freksa, C., Knauff, B., Krieg-Bruckner, B., and Nebel, B., editors, *Spatial Cognition IV. Reasoning, Action, Interaction*, pages 98–110. Springer.

Deco, G. and Lee, T. S. (2002). A unified model of spatial and object attention based on inter-cortical biased competition. *Neurocomputing*, 44:775–781.

Denève, S., Latham, P. E., and Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nature neuroscience*, 4(8):826–831.

Denève, S. and Pouget, A. (2003). Basis functions for object-centered representations. *Neuron*, 37(2):347–359.

Denève, S. and Pouget, A. (2004). Bayesian multisensory integration and cross-modal spatial links. *Journal of Physiology-Paris*, 98(1):249–258.

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1373):1245–1255.

Deubel, H. (2004). Localization of targets across saccades: Role of landmark objects. *Visual Cognition*, 11(2-3):173–202.

Deubel, H., Bridgeman, B., and Schneider, W. X. (1998). Immediate post-saccadic information mediates space constancy. *Vision research*, 38(20):3147–3159.

Dorris, M. C., Pare, M., and Munoz, D. P. (1997). Neuronal activity in monkey superior colliculus related to the initiation of saccadic eye movements. *The Journal of Neuroscience*, 17(21):8566–8579.

Duhamel, J., Colby, C., and Goldberg, M. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. *Science*, 255(5040):90–92.

Erickson, R. P. (1974). Parallel "population" neural coding in feature extraction. In Schmitt, F. and Worden, F., editors, *The neurosciences: Third study program*, pages 155–169. MIT Press, Cambridge, MA.

Erlhagen, W., Bastian, A., Jancke, D., Riehle, A., and Schöner, G. (1999). The distribution of neuronal population activation (dpa) as a tool to study interaction and integration in cortical representations. *Journal of Neuroscience Methods*, 94(1):53–66.

Faubel, C. and Schöner, G. (2009). A neuro-dynamic architecture for one shot learning of objects that uses both bottom-up recognition and top-down prediction. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 3162–3169. IEEE.

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

Fleischer, F., Caggiano, V., Thier, P., and Giese, M. A. (2013). Physiologically inspired model for the visual recognition of transitive hand actions. *The Journal of Neuroscience*, 33(15):6563–6580.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3–71.

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., and Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, 122(2):210–227.

Fuster, J. M. and Jervey, J. P. (1981). Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. *Science*, 212(4497):952–955.

Gardner, J. L., Merriam, E. P., Movshon, J. A., and Heeger, D. J. (2008). Maps of visual space in human occipital cortex are retinotopic, not spatiotopic. *The Journal of neuroscience*, 28(15):3988–3999.

Gentner, D. (2001). Spatial metaphors in temporal reasoning. In Gattis, M., editor, *Spatial schemas and abstract thought*, pages 203–222. MIT Press, Cambridge, MA.

Georgopoulos, A. P., Schwartz, A. B., and Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419.

Golledge, H. D., Panzeri, S., Zheng, F., Pola, G., Scannell, J. W., Giannikopoulos, D. V., Mason, R. J., Tovée, M. J., and Young, M. P. (2003). Correlations, feature-binding and population coding in primary visual cortex. *Neuroreport*, 14(7):1045–1050.

Golomb, J. D., Marino, A. C., Chun, M. M., and Mazer, J. A. (2011). Attention doesn't slide: spatiotopic updating after eye movements instantiates a new, discrete attentional locus. *Attention, Perception, & Psychophysics*, 73(1):7–14.

Goossens, H. and Van Opstal, A. (2006). Dynamic ensemble coding of saccades in the monkey superior colliculus. *Journal of neurophysiology*, 95(4):2326–2341.

Groh, J. M. (2001). Converting neural signals from place codes to rate codes. *Biological cybernetics*, 85(3):159–165.

Groh, J. M., Trause, A. S., Underhill, A. M., Clark, K. R., and Inati, S. (2001). Eye position influences auditory responses in primate inferior colliculus. *Neuron*, 29(2):509–518.

Hallett, P. E. and Lightstone, A. (1976). Saccadic eye movements to flashed targets. *Vision research*, 16(1):107–114.

Hamker, F. H. (2005a). The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. *Computer Vision and Image Understanding*, 100(1):64–106.

Hamker, F. H. (2005b). The reentry hypothesis: the putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas v4, it for attention and eye movement. *Cerebral Cortex*, 15(4):431–447.

Hamker, F. H. (2006). Modeling feature-based attention as an active top-down inference process. *BioSystems*, 86(1):91–99.

Hayhoe, M., Lachter, J., and Feldman, J. (1991). Integration of form across saccadic eye movements. *Perception*, 20(3):393–402.

Heide, W., Blankenburg, M., Zimmermann, E., and Kömpf, D. (1995). Cortical control of double-step saccades: implications for spatial orientation. *Annals of neurology*, 38(5):739–748.

Hoffman, J. E. and Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & psychophysics*, 57(6):787–795.

Hollingworth, A. (2007). Object-position binding in visual memory for natural scenes and object arrays. *Journal of Experimental Psychology: Human Perception and Performance*, 33(1):31.

Hollingworth, A. and Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 28(1):113.

Hollingworth, A. and Hwang, S. (2013). The relationship between visual working memory and attention: retention of precise colour information in the absence of effects on perceptual selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1628):20130061.

Hollingworth, A., Matsukura, M., and Luck, S. J. (2013a). Visual working memory modulates low-level saccade target selection: Evidence from rapidly generated saccades in the global effect paradigm. *Journal of vision*, 13(13):4.

Hollingworth, A., Matsukura, M., and Luck, S. J. (2013b). Visual working memory modulates rapid eye movements to simple onset targets. *Psychological science*, page 0956797612459767.

Houtkamp, R. and Roelfsema, P. R. (2006). The effect of items in working memory on the deployment of attention and the eyes during visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 32(2):423.

Hubbard, E. M., Piazza, M., Pinel, P., and Dehaene, S. (2005). Interactions between number and space in parietal cortex. *Nature Reviews Neuroscience*, 6(6):435–448.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106.

Hyun, J.-s., Woodman, G. F., and Luck, S. J. (2009a). The role of attention in the binding of surface features to locations. *Visual cognition*, 17(1-2):10–24.

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

Irwin, D. E. (1992). Memory for position and identity across eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2):307.

Irwin, D. E. and Andrews, R. V. (1996). Integration and accumulation of information across saccadic eye movements. *Attention and performance XVI: Information integration in perception and communication*, 16:125–155.

Irwin, D. E., Yantis, S., and Jonides, J. (1983). Evidence against visual integration across saccadic eye movements. *Perception & Psychophysics*, 34(1):49–57.

Issa, N. P., Trepel, C., and Stryker, M. P. (2000). Spatial frequency maps in cat visual cortex. *The Journal of Neuroscience*, 20(22):8504–8514.

Jancke, D., Erlhagen, W., Dinse, H. R., Akhavan, A. C., Giese, M., Steinhage, A., and Schöner, G. (1999). Parametric population representation of retinal location: Neuronal interaction dynamics in cat primary visual cortex. *The Journal of Neuroscience*, 19(20):9016–9028.

Johnson, J. S., Hollingworth, A., and Luck, S. J. (2008). The role of attention in the maintenance of feature bindings in visual short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*, 34(1):41.

Johnson, J. S., Spencer, J. P., Luck, S. J., and Schöner, G. (2009a). A dynamic neural field model of visual working memory and change detection. *Psychological Science*, 20(5):568–577.

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

Kaas, J. H., Nelson, R. J., Sur, M., Lin, C.-S., and Merzenich, M. M. (1979). Multiple representations of the body within the primary somatosensory cortex of primates. *Science*, 204(4392):521–523.

Kahneman, D., Treisman, A., and Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive psychology*, 24(2):175–219.

Keith, G. P., Blohm, G., and Crawford, J. D. (2010). Influence of saccade efference copy on the spatiotemporal properties of remapping: a neural network study. *Journal of neurophysiology*, 103(1):117–139.

Klaes, C., Schneegans, S., Schöner, G., and Gail, A. (2012). Sensorimotor learning biases choice behavior: A learning neural field model for decision making. *PLoS computational biology*, 8(11):e1002774.

Knauff, M. (2013). *Space to reason: A spatial theory of human thought*. MIT Press, Cambridge, MA.

Knips, G., Zibner, S. K., Reimann, H., Popova, I., and Schöner, G. (2014). A neural dynamics architecture for grasping that integrates perception and movement generation and enables on-line updating. In *Intelligent Robots*

and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on, pages 646–653. IEEE.

Kopecz, K. and Schöner, G. (1995). Saccadic motor planning by integrating visual information and pre-information on neural dynamic fields. *Biological cybernetics*, 73(1):49–60.

Krüger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., Rodríguez-Sánchez, A. J., and Wiskott, L. (2013). Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1847–1871.

Lamme, V. A., Super, H., and Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing in the visual cortex. *Current opinion in neurobiology*, 8(4):529–535.

Land, M. F. and Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision research*, 41(25):3559–3565.

Lee, C., Rohrer, W. H., Sparks, D. L., et al. (1988). Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature*, 332(6162):357–360.

Lipinski, J., Sandamirskaya, Y., and Schöner, G. (2009). Swing it to the left, swing it to the right: enacting flexible spatial language using a neurodynamic framework. *Cognitive Neurodynamics*, 3(4):373–400.

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

Livingstone, M. S. and Hubel, D. H. (1984). Anatomy and physiology of a color system in the primate visual cortex. *J Neurosci*, 4(1):309–356.

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, 20(5):1015.

Logan, G. D. (1995). Linguistic and conceptual control of visual spatial attention. *Cognitive psychology*, 28(2):103–174.

Logan, G. D. and Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In Bloom, P., Peterson, M., Nadel, L., and Garrett, M., editors, *Language, Speech, and Communication Series: Language and space*, pages 493–529. MIT Press, Cambridge, MA.

Luck, S. J. and Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657):279–281.

Marino, R. A., Trappenberg, T. P., Dorris, M., and Munoz, D. P. (2012). Spatial interactions in the superior colliculus predict saccade behavior in a neural field model. *Journal of cognitive neuroscience*, 24(2):315–336.

McAdams, C. J. and Maunsell, J. H. (2000). Attention to both space and feature modulates neuronal responses in macaque area v4. *Journal of Neurophysiology*, 83(3):1751–1755.

Miller, E. K., Erickson, C. A., and Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *The Journal of Neuroscience*, 16(16):5154–5167.

Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, 6:414–417.

Mishra, A., Aloimonos, Y., and Fermuller, C. (2009). Active segmentation for robotics. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 3133–3139. IEEE.

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas v1, v2, and v4 in the presence of competing stimuli. *Journal of neurophysiology*, 70:909–909.

Nissen, M. J. (1985). Accessing features and objects: Is location special. In Posner, M. I. and Marin, O. S., editors, *Attention and performance XI*, pages 205–219. Erlbaum, Hillsdale, JN.

O'Dhaniel, A., Cohen, Y. E., and Groh, J. M. (2005). Eye-centered, head-centered, and complex coding of visual and auditory targets in the intraparietal sulcus. *Journal of Neurophysiology*, 94(4):2331–2352.

O'Keefe, J. (2003). Vector grammar, places, and the functional role of the spatial prepositions in english. In van der Zee, E. and Slack, J., editors, *Representing direction in language and space*, pages 69–85. Oxford University Press, Oxford, England.

Olivers, C. N., Peters, J., Houtkamp, R., and Roelfsema, P. R. (2011). Different states in visual working memory: When it guides attention and when it does not. *Trends in cognitive sciences*, 15(7):327–334.

Olshausen, B. A., Anderson, C. H., and Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience*, 13(11):4700–4719.

O'Regan, J. K., Deubel, H., Clark, J. J., and Rensink, R. A. (2000). Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition*, 7(1-3):191–211.

Palanca, B. J. and DeAngelis, G. C. (2005). Does neuronal synchrony underlie visual feature grouping? *Neuron*, 46(2):333–346.

Parra, M. A., Della Sala, S., Logie, R. H., and Morcom, A. M. (2014). Neural correlates of shape–color binding in visual working memory. *Neuropsychologia*, 52:27–36.

Pasternak, T. and Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, 6(2):97–107.

Perone, S., Simmering, V. R., and Spencer, J. P. (2011). Stronger neural dynamics capture changes in infants' visual working memory capacity over development. *Developmental science*, 14(6):1379–1392.

Pertzov, Y. and Husain, M. (2013). The privileged role of location in visual working memory. *Attention, Perception, & Psychophysics*, pages 1–11.

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

Pouget, A. and Sejnowski, T. (1997). Spatial transformations in the parietal cortex using basis functions. *Cognitive Neuroscience, Journal of*, 9(2):222–237.

Pylyshyn, Z. W. (1989). The role of location indexes in spatial perception: A sketch of the finst spatial-index model. *Cognition*, 32(1):65–97.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, 80(1):127–158.

Pylyshyn, Z. W. (2007). *Things and places: How the mind connects with the world*. MIT press, Cambridge, MA.

Quaia, C., Optican, L. M., and Goldberg, M. E. (1998). The maintenance of spatial accuracy by the perisaccadic remapping of visual receptive fields. *Neural Networks*, 11(7):1229–1240.

Rasolzadeh, B., Björkman, M., Huebner, K., and Kragic, D. (2010). An active vision system for detecting, fixating and manipulating objects in the real world. *The International Journal of Robotics Research*, 29(2-3):133–154.

Regier, T. and Carlson, L. A. (2001). Grounding spatial language in perception: an empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2):273.

Rensink, R. A., O'Regan, J. K., and 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.

Reynolds, J. H. and Chelazzi, L. (2004). Attentional modulation of visual processing. *Annu. Rev. Neurosci.*, 27:611–647.

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

Richter, M., Lins, J., Schneegans, S., Sandamirskaya, Y., and Schöner, G. (2014a). Autonomous neural dynamics to test hypotheses in a model of spatial language. In Bello, P., Guarini, M., McShane, M., and Scassellati, B., editors, *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 2847–2852, Austin, TX. Cognitive Science Society.

Richter, M., Lins, J., Schneegans, S., and Schöner, G. (2014b). A neural dynamic architecture resolves phrases about spatial relations in visual scenes. In *24th International Conference on Artificial Neural Network, ICANN*, pages 201–208, Heidelberg, Germany. Springer.

Richter, M., Sandamirskaya, Y., and Schöner, G. (2012). A robotic architecture for action selection and behavioral organization inspired by human cognition. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 2457–2464. IEEE.

Riegler, A. (2002). When is a cognitive system embodied? *Cognitive Systems Research*, 3(3):339–348.

Roth, J. and Franconeri, S. (2012). Asymmetric coding of categorical spatial relations in both language and vision. *Frontiers in psychology*, 3.

Roy, D. (2005). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1):170–205.

Sandamirskaya, Y., Lipinski, J., Iossifidis, I., and Schöner, G. (2010). Natural human-robot interaction through spatial language: a dynamic neural field approach. In *19th IEEE international symposium on robot and human interactive communication, RO-MAN*, pages 600–607, Viareggio, Italy. IEEE.

Sandamirskaya, Y., Richter, M., and Schöner, G. (2011). A neural-dynamic architecture for behavioral organization of an embodied agent. In *Development and Learning (ICDL), 2011 IEEE International Conference on*, volume 2, pages 1–7. IEEE.

Sandamirskaya, Y. and Schöner, G. (2010). An embodied account of serial order: How instabilities drive sequence generation. *Neural Networks*, 23(10):1164–1179.

Schall, J. D. (2004). On the role of frontal eye field in guiding attention and saccades. *Vision research*, 44(12):1453–1467.

Schneegans, S. (in press). Sensory-motor and cognitive transformations. In Schöner, G. and Spencer, J. P., editors, *Dynamic Thinking: A Primer on Dynamic Field Theory*. Oxford University Press, New York.

Schneegans, S., Lins, J., and Spencer, J. P. (in press a). Integration and selection in multi-dimensional dynamic fields. In Schöner, G. and Spencer, J. P., editors, *Dynamic Thinking: A Primer on Dynamic Field Theory*. Oxford University Press, New York.

Schneegans, S. and Schöner, G. (2008). Dynamic field theory as a framework for understanding embodied cognition. In Calvo, P. and Gomila, T., editors, *Handbook of cognitive science: An embodied approach*, pages 241–271. Elsevier, Amsterdam, Netherlands.

Schneegans, S. and Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biological cybernetics*, 106(2):89–109.

Schneegans, S., Spencer, J. P., and Schöner, G. (in press b). Integrating "what" and "where": Visual working memory for objects in a scene. In Schöner, G. and Spencer, J. P., editors, *Dynamic Thinking: A Primer on Dynamic Field Theory*. Oxford University Press, New York.

Schneegans, S., Spencer, J. P., Schöner, G., Hwang, S., and Hollingworth, A. (2014). Dynamic interactions between visual working memory and saccade target selection. *Journal of vision*, 14(11):9.

Schutte, A. R. and Spencer, J. P. (2009). Tests of the dynamic field theory and the spatial precision hypothesis: Capturing a qualitative developmental transition in spatial working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 35(6):1698.

Schutte, A. R. and Spencer, J. P. (2010). Filling the gap on developmental change: Tests of a dynamic field theory of spatial cognition. *Journal of Cognition and Development*, 11(3):328–355.

Shadlen, M. N. and Movshon, J. A. (1999). Synchrony unbound: a critical evaluation of the temporal binding hypothesis. *Neuron*, 24(1):67–77.

Shastri, L. (1999). Advances in shruti—a neurally motivated model of relational knowledge representation and rapid inference using temporal synchrony. *Applied Intelligence*, 11(1):79–108.

Simmering, V. R., Schutte, A. R., and Spencer, J. P. (2008). Generalizing the dynamic field theory of spatial cognition across real and developmental time scales. *Brain research*, 1202:68–86.

Simmering, V. R., Spencer, J. P., and Schöner, G. (2006). Reference-related inhibition produces enhanced position discrimination and fast repulsion near axes of symmetry. *Perception & Psychophysics*, 68(6):1027–1046.

Simons, D. J. and Levin, D. T. (1997). Change blindness. *Trends in cognitive sciences*, 1(7):261–267.

Singer, W. (2001). Consciousness and the binding problem. *Annals of the New York Academy of Sciences*, 929(1):123–146.

Skubic, M., Perzanowski, D., Blisard, S., Schultz, A., Adams, W., Bugajska, M., and Brock, D. (2004). Spatial language for human-robot dialogs. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34(2):154–167.

Snyder, L. H., Grieve, K. L., Brotchie, P., and Andersen, R. A. (1998). Separate body-and world-referenced representations of visual space in parietal cortex. *Nature*, 394(6696):887–891.

Sommer, M. A. and Wurtz, R. H. (2004). What the brain stem tells the frontal cortex. i. oculomotor signals sent from superior colliculus to frontal eye field via mediodorsal thalamus. *Journal of Neurophysiology*, 91(3):1381–1402.

Sommer, M. A. and Wurtz, R. H. (2006). Influence of the thalamus on spatial visual processing in frontal cortex. *Nature*, 444(7117):374–377.

Sommer, M. A. and Wurtz, R. H. (2008). Brain circuits for the internal monitoring of movements. *Annual review of neuroscience*, 31:317.

Sparks, D. L. (2002). The brainstem control of saccadic eye movements. *Nature Reviews Neuroscience*, 3(12):952–964.

Sparks, D. L. and Nelson, I. S. (1987). Sensory and motor maps in the mammalian superior colliculus. *Trends in Neurosciences*, 10(8):312–317.

Spencer, J., Barich, K., Goldberg, J., and Perone, S. (2012). Behavioral dynamics and neural grounding of a dynamic field theory of multi-object tracking. *Journal of integrative neuroscience*, 11(03):339–362.

Spencer, J. P., Smith, L. B., and Thelen, E. (2001). Tests of a dynamic systems account of the a-not-b error: The influence of prior experience on the spatial memory abilities of two-year-olds. *Child development*, 72(5):1327–1346.

Stopp, E., Gapp, K.-P., Herzog, G., Laengle, T., and Lueth, T. C. (1994). Utilizing spatial relations for natural language access to an autonomous mobile robot. In Nebel, B. and Dreschler-Fischer, L., editors, *Proceedings of KI-94: Advances in Artificial Intelligence, 18th German Annual Conference on Artificial Intelligence*, Berlin and Heidelberg, Germany. Springer-Verlag.

Stricanne, B., Andersen, R. A., and Mazzoni, P. (1996). Eye-centered, head-centered, and intermediate coding of remembered sound locations in area lip. *Journal of Neurophysiology*, 76(3).

Swindale, N. V. (1998). Orientation tuning curves: empirical description and estimation of parameters. *Biological cybernetics*, 78(1):45–56.

Tipper, S. P., Weaver, B., Jerreat, L. M., and Burak, A. L. (1994). Object-based and environment-based inhibition of return of visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, 20(3):478.

Trappenberg, T. P., Dorris, M. C., Munoz, D. P., and Klein, R. M. (2001). A model of saccade initiation based on the competitive integration of exogenous and endogenous signals in the superior colliculus. *Journal of Cognitive Neuroscience*, 13(2):256–271.

Treisman, A. (1988). Features and objects: The fourteenth bartlett memorial lecture. *The quarterly journal of experimental psychology*, 40(2):201–237.

Treisman, A. (1996). The binding problem. *Current opinion in neurobiology*, 6(2):171–178.

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

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136.

Van der Stigchel, S. and Nijboer, T. C. (2011). The global effect: what determines where the eyes land. *Journal of Eye Movement Research*, 4(2):1–13.

van Hengel, U., Sandamirskaya, Y., Schneegans, S., and Schöner, G. (2012). A neural-dynamic architecture for flexible spatial language: intrinsic frames, the term "between", and autonomy. In *21st IEEE international*

symposium on robot and human interactive communication, RO-MAN, pages 150–157, Paris, France. IEEE.

Vidyasagar, T. R. (1999). A neuronal model of attentional spotlight: parietal guiding the temporal. *Brain Research Reviews*, 30(1):66–76.

Vogel, E. K., Woodman, G. F., and Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 27(1):92.

Vogel, E. K., Woodman, G. F., and Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 32(6):1436.

Von der Malsburg, C. (1999). The what and why of binding: the modeler's perspective. *Neuron*, 24(1):95–104.

Wachtler, T., Sejnowski, T. J., and Albright, T. D. (2003). Representation of color stimuli in awake macaque primary visual cortex. *Neuron*, 37(4):681–691.

Wang, X.-J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends in neurosciences*, 24(8):455–463.

Wheeler, M. E. and Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General*, 131(1):48.

White III, R. L. and Snyder, L. H. (2004). A neural network model of flexible spatial updating. *Journal of neurophysiology*, 91(4):1608–1619.

Wilimzig, C., Schneider, S., and Schöner, G. (2006). The time course of saccadic decision making: Dynamic field theory. *Neural Networks*, 19(8):1059–1074.

Wilson, H. R. and Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, 13(2):55–80.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, 9(4):625–636.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238.

Wurtz, R. H. (2008). Neuronal mechanisms of visual stability. *Vision research*, 48(20):2070–2089.

Xing, J. and Andersen, R. (2000). Models of the posterior parietal cortex which perform multimodal integration and represent space in several coordinate frames. *Cognitive Neuroscience, Journal of*, 12(4):601–614.

Zhang, W. and Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, 453(7192):233–235.

Zibner, S. K., Faubel, C., Iossifidis, I., and Schöner, G. (2010a). Scene representation for anthropomorphic robots: A dynamic neural field approach. In *Robotics (ISR), 2010 41st International Symposium on and 2010 6th German Conference on Robotics (ROBOTIK)*, pages 1–7. VDE.

Zibner, S. K., Faubel, C., Iossifidis, I., and Schöner, G. (2011a). Dynamic neural fields as building blocks of a cortex-inspired architecture for robotic scene representation. *Autonomous Mental Development, IEEE Transactions on*, 3(1):74–91.

Zibner, S. K., Faubel, C., Iossifidis, I., Schöner, G., and Spencer, J. P. (2010b). Scenes and tracking with dynamic neural fields: how to update a robotic scene representation. In *Development and Learning (ICDL), 2010 IEEE 9th International Conference on*, pages 244–250. IEEE.

Zibner, S. K., Faubel, C., and Schöner, G. (2011b). Making a robotic scene representation accessible to feature and label queries. In *Development and Learning (ICDL), 2011 IEEE International Conference on*, volume 2, pages 1–7. IEEE.

Zipser, D. and Andersen, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, 331(6158):679–684.

Lebenslauf

Persönliche Daten

Name:	Sebastian Schneegans
Geburtsdatum:	18.08.1980
Geburtsort:	Duderstadt
Anschrift:	Brückstraße 21 44787 Bochum
Telefon:	+49 151 11109948
E-Mail:	Sebastian@Schneegans.de
Familienstand:	ledig, keine Kinder

Berufliche Tätigkeit

seit 01/2007	Doktorand und wissenschaftlicher Mitarbeiter am Institut für Neuroinformatik, Ruhr-Universität Bochum
02/2002 – 04/2002	HiWi-Tätigkeit: Anwendungsentwicklung in C++
10/2001 – 02/2002	HiWi-Tätigkeit: Übungsgruppenleiter einer Informatik-Vorlesung

Studium

10/2000 – 09/2006	Studium der Bioinformatik an der Eberhard-Karls-Universität Tübingen mit Anwendungsschwerpunkt Neurobiologie Abschluss: Diplom in Informatik (Note: sehr gut)

Schullaufbahn

07/1993 – 06/1999	Eichsfeld-Gymnasium Duderstadt Abschluss: Abitur (Note: 1,2)
08/1991 – 06/1993	St.-Ursula-Schule Duderstadt
08/1987 – 07/1991	Grundschule Gerblingerode