

# Department of Computer Science, University of Otago

UNIVERSITY  
of  
OTAGO



*Te Whare Wānanga o Ōtāgo*

---

Technical Report OUCS-2017-01

## **Road map for an embodied model of language and cognition**

Author:

**Alistair Knott**

Department of Computer Science, University of Otago, New Zealand



---

Department of Computer Science,  
University of Otago, PO Box 56, Dunedin, Otago, New Zealand

<http://www.cs.otago.ac.nz/research/techreports.php>

# Road map for an embodied model of language and cognition

Ali

January 10, 2017

# Contents

<b>1</b>	<b>Introduction</b>	<b>7</b>
<b>I</b>	<b>Sensorimotor mechanisms</b>	<b>8</b>
<b>2</b>	<b>A model of low-level vision, attention, and object classification</b>	<b>9</b>
2.1	The architecture of the visual system . . . . .	9
2.2	A model of low-level vision . . . . .	9
2.2.1	Preliminaries: a self-organising map . . . . .	10
2.2.2	An array of SOMs for learning local visual patterns . . . . .	11
2.3	Spatial attention: the saliency map . . . . .	12
2.3.1	The saliency map . . . . .	12
2.3.2	An initial model of the saliency map . . . . .	12
2.3.3	Using the saliency map to select visual information . . . . .	14
2.4	A mechanism for selecting the next focus of attention: the gaze-orienting SOM . . . . .	14
2.4.1	Background: recurrent SOMs and reconstruction of inputs . . . . .	14
2.4.2	How the gaze-orienting SOM learns the orienting function, and a body-centred representation of gaze direction . . . . .	15
2.4.3	How the gaze-orienting SOM learns a saliency map that is stable over eye/head movements . . . . .	18
2.4.4	A circuit for covert attention . . . . .	19
2.5	Neural evidence for the media in the orienting circuit . . . . .	20
2.5.1	The gaze-orienting SOM is in frontal eye fields . . . . .	20
2.5.2	Modulation of low-level visual feature maps by spatial attention . . . . .	20
2.6	Object tracking . . . . .	20
2.6.1	Multiple object tracking . . . . .	20
2.6.2	A model of simple object tracking . . . . .	20
2.6.3	Multiple object tracking and object files . . . . .	22
2.6.4	A revision of the model: multiple tracking maps . . . . .	22
2.7	Visual object classification mechanisms . . . . .	23
2.7.1	Background: the brain's object classification system . . . . .	23
2.7.2	Recap: a layer of local SOMs detecting simple visual features . . . . .	23
2.7.3	A spatial pooling operation . . . . .	23
2.7.4	Iterating on the above scheme: hierarchical classification . . . . .	25
2.7.5	'Training' of the classification SOM . . . . .	25
2.7.6	Cardinality blindness in the classifier . . . . .	28
2.8	Multiple processing streams for visual property types . . . . .	28
2.8.1	Neural representations of shape and shape adjectives . . . . .	29
2.8.2	Neural representations of colour and colour adjectives . . . . .	29
2.8.3	Neural representations of texture . . . . .	29
2.8.4	Neural representations of affective properties . . . . .	30
2.8.5	Neural representations of category-relative geometric properties . . . . .	31

2.9	A circuit for representing object categories and properties . . . . .	31
2.9.1	A look-ahead to predicative propositions . . . . .	31
2.10	Representations of object categories, and associated attentional operations . . . . .	31
2.10.1	The rich property complex . . . . .	32
2.10.2	Unsupervised learning of object categories . . . . .	32
2.10.3	Property-level IOR: a mechanism for attending to properties of objects . . . . .	33
2.10.4	Using property-level IOR to learn a hierarchical type system . . . . .	34
2.11	Saliency at higher levels of the visual object classification network . . . . .	35
2.11.1	Local SOMs at different spatial granularities . . . . .	35
2.11.2	Saliency maps at different spatial granularities . . . . .	35
2.11.3	Popout . . . . .	36
2.11.4	Grouping by textural homogeneity . . . . .	38
2.11.5	Grouping by spatial proximity, and a representation of ‘parts’ of a visual stimulus . . . . .	39
2.12	Mechanisms for attending to components of a selected salient region at different spatial frequencies . . . . .	41
2.12.1	A mechanism for attending to wholes and parts of objects . . . . .	42
2.12.2	A mechanism for attending to the local and global form of objects . . . . .	44
<b>3</b>	<b>Spatial representations: environments and places</b>	<b>47</b>
3.1	A circuit for representations of places . . . . .	47
3.1.1	An actor-critic system for learning the next locomotion action . . . . .	47
3.2	A circuit for representations of environments . . . . .	48
3.2.1	Learning (and identifying) LTM environments . . . . .	48
3.3	A circuit for representations of navigational plans . . . . .	48
3.3.1	Learning of trajectories to places associated with external reward . . . . .	48
3.3.2	Working with multiple navigation goals: a mechanism for trajectory selection . . . . .	51
3.4	A circuit for representing the location of arbitrary individuals . . . . .	53
3.4.1	View cells . . . . .	53
3.4.2	Orientation cells . . . . .	53
3.4.3	A circuit for learning view cells: the orienting SOM . . . . .	53
3.4.4	Configuring the viewed-places SOM to generalise over ‘self’ and ‘other’ . . . . .	56
3.4.5	A scheme for learning to navigate to arbitrary places in the current environment: the goal places system . . . . .	57
3.4.6	Vision-based (‘egocentric’) navigation strategies . . . . .	58
3.5	Evidence for the media in the hippocampus and PFC . . . . .	59
3.5.1	Places MSOM . . . . .	59
3.5.2	Possible locomotion actions . . . . .	59
3.5.3	The current locomotion action . . . . .	59
3.5.4	LTM environments . . . . .	59
3.5.5	Navigational goals . . . . .	60
3.6	Consolidation within the place circuit . . . . .	60
3.7	Properties and types of LTM environments . . . . .	60
3.7.1	Properties of spatial environments, and spatial environment types . . . . .	60
3.7.2	Visual classification of spatial environments . . . . .	61
3.7.3	Spatial environment recognition . . . . .	62
3.8	The duality between individuals and environments . . . . .	62
3.9	Representation of groups of individuals as environments . . . . .	62
3.9.1	A model of the perception of group individuals . . . . .	62
3.9.2	When one object becomes a group . . . . .	62
3.10	Transitions between environments . . . . .	62
3.10.1	The major axis of an environment . . . . .	62
3.10.2	Transitions . . . . .	63

<b>4</b>	<b>Motor control of the arm: the reach visuomotor pathway</b>	<b>64</b>
4.1	Some interesting parallels between the navigation and motor control systems . . .	64
4.2	A model of the reach visuomotor pathway . . . . .	65
4.2.1	The motor system commands SOM . . . . .	65
4.2.2	The hand places SOM . . . . .	66
4.2.3	A system for learning trajectories to target objects . . . . .	69
4.3	Evidence for the reach model in parietal/premotor cortex . . . . .	74
<b>5</b>	<b>Representations of object geometry, and hierarchical spatial representations</b>	<b>75</b>
5.1	Affordance-based representations of physical objects . . . . .	75
5.2	Representations of object types in parietal cortex . . . . .	75
5.3	Representation of object dimensions in parietal cortex . . . . .	75
5.4	Mapping from early vision to parietal object representations . . . . .	76
5.4.1	Computation of 3D object type . . . . .	76
5.4.2	Computation of shape category transformations: shape and size properties	76
5.5	Mappings between dorsal and ventral object representations: types, size and shape	76
5.5.1	Influence of motor types on ventral types . . . . .	76
5.5.2	Representations of shape and size in the ventral visual pathway . . . . .	77
5.6	Representations of object parts and relations between objects . . . . .	78
5.6.1	Object parts . . . . .	78
5.6.2	Relations between objects . . . . .	78
5.7	Mappings between dorsal and ventral object representations: object parts and relations between objects . . . . .	78
<b>6</b>	<b>The grasp visuomotor pathway</b>	<b>79</b>
6.1	Evidence for the grasp model in parietal/premotor cortex . . . . .	79
6.2	Towards a definition of hand/arm motor programmes? . . . . .	79
6.3	A model of causative motor actions . . . . .	79
6.4	Transitive and intransitive body-centred actions . . . . .	79
<b>7</b>	<b>Action execution and action perception: the self-other distinction</b>	<b>81</b>
7.1	Early stages in the action perception pathway . . . . .	81
7.1.1	V1 units representing simple motion patterns . . . . .	81
7.1.2	MT and MST: encoding patterns of motion . . . . .	82
7.1.3	STS: joint attention and biological motion recognition . . . . .	82
7.2	A model of ‘the mirror system’ for reach/grasp actions . . . . .	82
<b>8</b>	<b>Summary: SM mechanisms involved in apprehending a simple transitive action</b>	<b>83</b>
 <b>II Working memory and long-term memory</b>		<b>84</b>
<b>9</b>	<b>Memory representations of individuals, episodes, situations and plans: an introductory model</b>	<b>85</b>
9.1	Motivating ideas . . . . .	85
9.1.1	Episodes are perceived in sequentially structured SM routines . . . . .	85
9.1.2	Individuals are perceived in sequentially structured SM routines . . . . .	86
9.1.3	Individuals and episodes are represented in WM as prepared SM routines .	86
9.1.4	Event representations make use of pointers . . . . .	86
9.1.5	Localist representations of events, and an associated probability model . . .	88
9.1.6	Localist representations of ‘situations’ . . . . .	88
9.2	A circuit for representing individuals, episodes, situations and plans . . . . .	89
9.3	Properties of units in each medium of the circuit . . . . .	90
9.3.1	The LTM/WM individuals system . . . . .	90

9.3.2	The current WM episode: agent, patient and action media . . . . .	90
9.3.3	The candidate episodes buffer . . . . .	91
9.3.4	The situations SOM . . . . .	91
9.3.5	The scenarios medium . . . . .	91
9.4	Evidence for the media in the hippocampus and medial temporal cortex . . . . .	91
9.4.1	Preliminaries: models of the structure of hippocampus . . . . .	91
9.4.2	The LTM/WM individuals system . . . . .	91
9.4.3	The current WM episode: agent, patient and action media . . . . .	92
9.4.4	The candidate episodes buffer . . . . .	92
9.4.5	The situations SOM . . . . .	92
9.4.6	Orthogonally from all this: a mixture of specific and general representations	93
9.5	Evidence for the media in PFC . . . . .	94
9.5.1	Preliminaries: models of the internal structure of PFC . . . . .	94
9.5.2	The LTM/WM individuals system . . . . .	94
9.5.3	The current WM episode: agent, patient and action media . . . . .	94
9.5.4	The candidate episodes buffer . . . . .	94
9.5.5	The situations SOM . . . . .	95
9.5.6	The scenarios medium . . . . .	95
9.5.7	Orthogonally from all this: a mixture of specific and general representations	95
<b>10</b>	<b>Memory representations of individuals, episodes, situations and plans: a detailed model</b>	<b>96</b>
10.1	Architecture of the complete model, and empirical motivation . . . . .	96
10.1.1	Architecture . . . . .	96
10.1.2	Empirical support for the architecture . . . . .	96
10.2	Application 1: Consolidation of hippocampal memories . . . . .	96
10.2.1	Evidence for the roles of hippocampus and PFC in consolidation . . . . .	97
10.2.2	How hippocampus encodes token episodes . . . . .	97
10.2.3	Principles for selecting hippocampal material to replay . . . . .	97
10.2.4	Implementing a buffer of recent episodes . . . . .	97
10.2.5	Learning token episodes/types, and generalisations, in PFC . . . . .	98
10.2.6	Remote and generic memories in the hippocampus? . . . . .	98
10.3	Application 2: Learning of optimal SM behaviour . . . . .	99
10.4	Application 3: Querying of episodic memories . . . . .	99
10.4.1	Post-retrieval processing . . . . .	100
10.5	Application 4: The interface with language . . . . .	100
10.5.1	A SM interpretation of syntactic structure . . . . .	100
10.5.2	A model of sentence generation . . . . .	100
10.5.3	Elements of a model of sentence semantics . . . . .	100
<b>11</b>	<b>WM representations of spatial and stative propositions</b>	<b>102</b>
11.1	WM representations of spatial propositions . . . . .	102
11.2	WM representations of predicative propositions . . . . .	102
<b>12</b>	<b>LTM for spatial and stative information</b>	<b>103</b>
12.1	LTM for object locations . . . . .	103
12.1.1	Object location memory and trajectory planning . . . . .	103
12.2	LTM for object parts and possessions . . . . .	105
12.3	LTM for object properties . . . . .	105
12.3.1	The dominant property assembly layer: representations of object types, and of properties . . . . .	105
12.3.2	The property memory circuit . . . . .	105
12.3.3	The location memory and property memory circuits combined, in an account of object recognition . . . . .	106

12.3.4	The continuity of LTM individuals . . . . .	106
12.4	Evidence for the media in the hippocampal region . . . . .	107
12.4.1	Object location memory . . . . .	107
12.4.2	How is the spatial representation system overlaid on the episodes/situations system? . . . . .	107
12.5	Evidence for the media in PFC? . . . . .	108
12.6	Interactions between the spatial representation system and the episodes/situations system . . . . .	108
12.6.1	Spatial components to situations . . . . .	108
12.6.2	Modelling the effects of episodes on their participant individuals . . . . .	108
<b>13</b>	<b>The reward system in the situation update network</b>	<b>109</b>
13.1	The reward system for action execution . . . . .	109
13.2	The reward system for action perception . . . . .	109
13.3	Situation updates and plan updates . . . . .	109
<b>14</b>	<b>Cross-modal representations of agents and patients</b>	<b>110</b>
<b>III</b>	<b>Language representation mechanisms</b>	<b>111</b>
<b>15</b>	<b>A model of clause syntax</b>	<b>112</b>
<b>16</b>	<b>A SM interpretation of clause syntax</b>	<b>113</b>
<b>17</b>	<b>A SM interpretation of stative sentences</b>	<b>114</b>
17.1	Locative sentences and LTM for object location . . . . .	114
17.2	Predicative sentences and LTM for object types . . . . .	114
17.2.1	A look ahead to DP-internal adjectives . . . . .	114
17.3	Possession sentences . . . . .	114
17.4	Stimulus-experiencer sentences . . . . .	114
<b>18</b>	<b>A SM interpretation of spatial PPs</b>	<b>115</b>
18.1	Syntax of spatial PPs . . . . .	115
18.2	SM interpretation of the syntax of a spatial PP . . . . .	115
18.2.1	WM spatial transitions . . . . .	115
<b>19</b>	<b>A model of sentences containing spatial PPs</b>	<b>116</b>
19.1	Stative spatial sentences . . . . .	116
19.2	Sentences denoting episodes involving movement . . . . .	116
<b>20</b>	<b>Emotional attitudes towards objects: perception, LTM and language</b>	<b>117</b>
20.1	Preliminaries: Damasio's model of emotions . . . . .	117
20.2	A model of emotion perception and representation . . . . .	117
20.3	LTM representations of emotional relations . . . . .	117
<b>21</b>	<b>Quantified propositions</b>	<b>118</b>
21.1	Quantified propositions and the semantic LTM system . . . . .	118
21.2	Reference to groups of objects perceived separately . . . . .	118
<b>22</b>	<b>The internal structure of DPs</b>	<b>119</b>
22.1	The syntax of DPs . . . . .	119
22.2	A SM interpretation of DP-internal adjectives . . . . .	119
22.3	Partitive DPs . . . . .	119
22.4	The <i>kind-of</i> construction . . . . .	119

<b>IV</b>	<b>Language processing mechanisms</b>	<b>120</b>
<b>23</b>	<b>Sentence generation</b>	<b>121</b>
<b>24</b>	<b>Sentence interpretation</b>	<b>122</b>

## Chapter 1

# Introduction

## Part I

# Sensorimotor mechanisms

## Chapter 2

# A model of low-level vision, attention, and object classification

### 2.1 The architecture of the visual system

Visual processing in the brain is hierarchical. At the early stages of visual processing, information in each area of the retina is processed separately, and simple, low-level representations of visual structure are extracted in parallel across the whole retina (see Schmolesky, 2007 for a review). Subsequent to these early processing stages, there are several distinct visual pathways: for instance, a pathway for classifying objects and identifying ‘semantic information’ about them (Logothetis, 1996); a pathway for computing the kind of geometric information about objects necessary to select and control simple physical actions on them, like reaching and grasping (Murata *et al.*, 2000; Fogassi *et al.*, 2005), and a pathway for classifying the actions of observed agents (Keyser and Perrett, 2004). In Section 2.2, I will introduce a model of low-level vision. In the remaining sections, I will discuss various attentional mechanisms that operate on low-level vision, and present models of these.

### 2.2 A model of low-level vision

A common idea in recent models of vision is that the early stages of the visual processing pathway learn primarily using unsupervised methods. This idea originated with Hinton (see e.g. Hinton, 2007); it was partly this idea that precipitated the explosion of interest in so-called ‘deep’ networks. In current deep networks, training in fact tends to use supervised methods throughout, so the earliest layers of vision are also trained with supervised methods (see e.g. Krizhevsky *et al.*, 2012 for a well-known example). There is good evidence that the training of low-level visual mechanisms is supervised by higher-level representations (see e.g. ), and I will provide a model of this supervision in Section 4.2.3 and Chapter 13. However, the representations learned at the lowest levels of a supervised deep network are not radically different from those learned by unsupervised methods (see e.g. Zeiler and Fergus, 2014 for an illustration). In the current chapter, I will focus on the unsupervised component of learning in the low-level visual system.

Hinton’s model of early vision hinges around mechanisms that identify *regularities* in the patterns that fall onto local regions of the retina. There are distinct mechanisms for identifying local regularities at each location on the retina, that operate in parallel, across the whole retina. They also operate at several different spatial frequencies, so that small patterns can be recognised at small places on the retina, and larger patterns can be recognised at larger places.

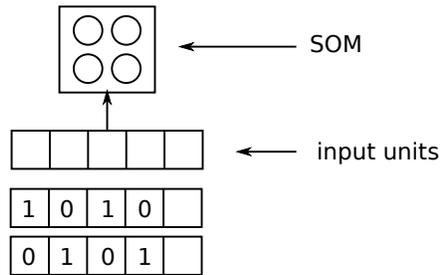


Figure 2.1: A simple self-organising map (SOM) containing 4 units. The map takes as input a vector of 5 units, which carry information about different ‘features’: 1 signals the presence of a feature. Each input unit is connected to each SOM unit. (These connections are indicated by a single line joining the SOM to its inputs.) Two sample training items ( $[1,0,1,0,-]$  and  $[0,1,0,1,-]$ ) are shown below.

### 2.2.1 Preliminaries: a self-organising map

I will use a learning device called a **self-organising map** or **SOM** (Kohonen, 1982) to model the early visual system. SOMs are the learning device we use throughout our model, so I will introduce them informally in this section.

A SOM is an unsupervised learning device, that identifies patterns in the input it receives. A simple example is given in Figure 2.1. The SOM receives input from an array of input units. Each unit represents a feature, which can be present in varying degrees: 1 signals a clearly expressed feature, 0 signals an absent feature, and values in between signal varying degrees of presence.

The units in a SOM are each connected to each of its input units, by links that can have variable weights. At the start of training, these weights are initialised to random values. It is trained on a set of training items presented on its input units. Each training item is a vector of numbers in the range  $[0-1]$ , signalling a particular set of features. For each training item, an activation level is computed for each SOM unit, based on the vector of numbers and their associated weights: the unit whose weights most closely correspond to the input values is selected as the ‘winner’, and the weights of this unit are incrementally modified, so they more closely resemble the input.<sup>1</sup> Say the example network in Figure 2.1 is presented many times with the two inputs shown in the figure. Notice there are correlations between features in these inputs: if the first feature is present (1), so is the third, and if the second feature is present (1), so is the fourth. During training, one of the SOM units will evolve weights to represent the first pattern, and another will evolve weights to represent the second pattern. These units can be seen as ‘localist’ representations of the distributed patterns they encode. Note that the SOM finds patterns by itself, without supervision. After it is trained, it performs an operation a little like principal components analysis, representing the strongest patterns that were present in the data, using different units for different patterns.

Since the units of a SOM hold localist representations of patterns, the activity across its units for a given input can be interpreted as a *probability distribution* over the patterns that the input might contain. If there is just one strong pattern in the input, one unit will be very active, and the others will all be inactive; if there are no identifiable patterns, there will be no strong differences between the units’ activity. It is easy to treat the SOM’s activity as a whole as a probability distribution, simply by scaling the activities of its units so they sum to 1. We systematically interpret the patterns in our SOMs as probability distributions. This permits a ‘Bayesian’ approach to representation and inference, which we use throughout the model. (A Bayesian model performs computations on whole probability distributions for variables, rather than trying to identify the ‘actual’ values of variables and then performing computations on those.) Interpreting SOM patterns as probability distributions also allows us to compute some

<sup>1</sup>There are various other factors involved in a SOM’s learning, but this description suffices for an informal understanding.

interesting properties of the SOM’s representation of its inputs, which we will briefly review here. These concepts will be used throughout the document.

**Entropy** The **entropy** of a probability distribution generated for a given input item provides a quantitative measure of how much information about the patterns represented by SOM units is present in this particular item. If the SOM can’t represent the item very well in terms of the patterns it has learned, its entropy will be high; but if it confidently identifies a single pattern, its entropy will be relatively low. If it clearly identifies two patterns, its entropy will still be low, but not as low as for a single pattern: a SOM is a competitive medium, that is trying to find ‘the single best pattern’, so the highest entropies are associated with the sparsest vectors of activity in the SOM. Entropy is a quantitative measure of ‘confidence’: low entropy means high confidence, and vice versa.

**Surprise** Another useful property of an observed SOM pattern is a measure of how **surprising** this pattern is, in relation to a population of patterns from which it is drawn as a sample. This can be defined using the Kullback-Leibler (KL) divergence between two probability distributions (Kullback and Leibler, 1951). Surprise can be defined as the KL divergence between the expected (or *prior*) distribution of SOM unit activity for the whole population, and the (*posterior*) distribution of this particular pattern (see e.g. Friston, 2010). A value close to zero denotes low surprise; high values denote high surprise.<sup>2</sup> This definition is consistent with the definition of confidence. If there is low confidence associated with the prior distribution, that means we have no strong expectations, and we won’t be very surprised by any particular observed distribution. Conversely, if there is low confidence associated with the observed distribution, we don’t know what patterns we are representing, so again we cannot be very surprised.

**SoftMax for selection** Sometimes we need to select candidate values from a SOM, based on a pattern of activity over its units. One way of doing this is simply to pick the most active unit, but a more nuanced way is to ‘sharpen’ the distribution, so as to emphasise the units with highest activation (which means de-emphasising the other units). A function called **softMax** is a good way of doing this (see e.g. Bishop, 2006).

### 2.2.2 An array of SOMs for learning local visual patterns

I propose to model the mechanisms that identify local visual patterns with an array of SOMs covering the whole retina. Each SOM learns frequently-occurring visual patterns that fall on its associated local patch of retina. The SOMs ‘stride’ over the retina in overlapping steps, an architecture which is standard in so-called ‘convolutional’ networks (see e.g. Le Cun and Bengio, 1995, Krizhevsky *et al.*, 2012). There are SOMs that identify patterns at several spatial scales, learning frequently-occurring combinations of simple visual features at the relevant scale. I will call the SOMs representing these local visual patterns **local pattern SOMs**.

The SOMs take their inputs from several distinct maps of simple visual features computed in parallel over the retina, encoding things like edges or bars at all possible orientations, blobs, the presence of different colours, and local movement in all possible directions. (These features must themselves be learned—and this learning could involve an earlier layer of ‘still more’ local pattern SOMs. But I won’t go into that detail.) The set of feature maps is shown in Figure 2.2.

The configuration of local pattern SOMs over these feature maps is illustrated in Figure 2.3. The figure just illustrates SOMs of a single spatial frequency: there will be several tiling patterns of this kind overlaid on the retina, taking input from visual features of different spatial frequencies, covering local areas commensurate with these frequencies.

If a variety of natural visual scenes is presented as input to this array of SOMs, they will learn patterns in the spatial structure of these scenes, at different scales. Note that the SOMs will all

---

<sup>2</sup>The KL divergence of two distributions ranges from 0 to arbitrarily large positive numbers.

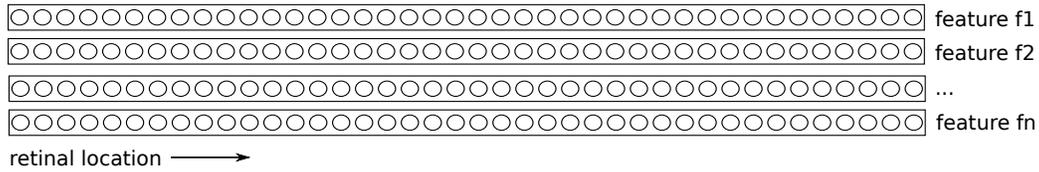


Figure 2.2: The set of feature maps. Each horizontal array of units represents a retinotopic map for a different type of visual feature.

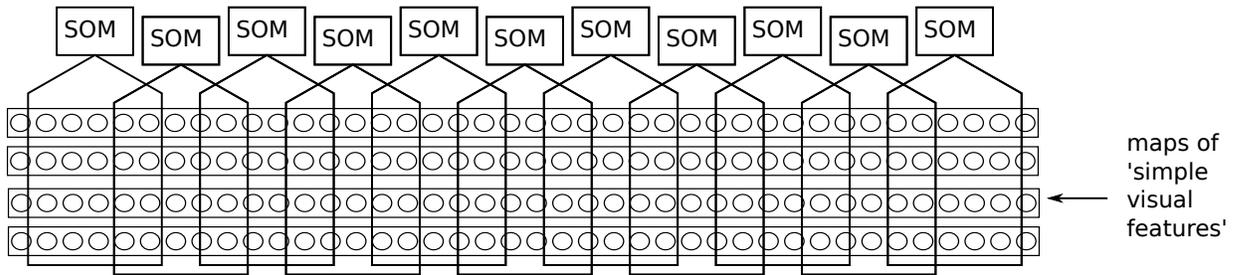


Figure 2.3: A set of local pattern SOMs tiling the retina

learn roughly the same thing, since we are not controlling for what types of pattern fall onto which locations on the retina.

## 2.3 Spatial attention: the saliency map

### 2.3.1 The saliency map

The visual field contains too much information to be processed in parallel beyond a certain early stage: the brain has a variety of mechanisms for selecting, or focussing on, certain spatial regions of the retina, and de-emphasising the others. This is often modelled in a structure called a ‘saliency map’, which computes a pattern of activation over the retina signalling the most important regions (Koch and Ullman, 1985; Wolfe, 1994; Itti and Koch, 2001). There is evidence that several regions of the brain compute something like a saliency map: in particular, there is something like a saliency map in lateral intraparietal cortex, or LIP (see e.g. Gottlieb *et al.*, 1998), and also in an area of frontal cortex called the frontal eye fields (FEF: see e.g. Bichot *et al.*, 2001). I will focus on the map in FEF in the current chapter; the parietal representations of space will be considered in Chapter 3.

### 2.3.2 An initial model of the saliency map

In this section, I will outline a simple model of the saliency map. This model will be elaborated in Section 2.11.

There have been several computational models of the saliency map. The model I introduce draws mainly on ideas from the models of Bruce and Tsotsos (2007) and Itti and Baldi (2009). These authors define the salience of a visual region using the information-theoretic concept of surprise introduced in Section 2.2.1. The basic idea is that visual attention focusses on regions of the retina that contain surprising information. It is straightforward to implement this principle in our model of local SOMs, if we allow them to maintain a running average of their distributions, so they can each compute a measure of how much the current distribution differs from the expected distribution.

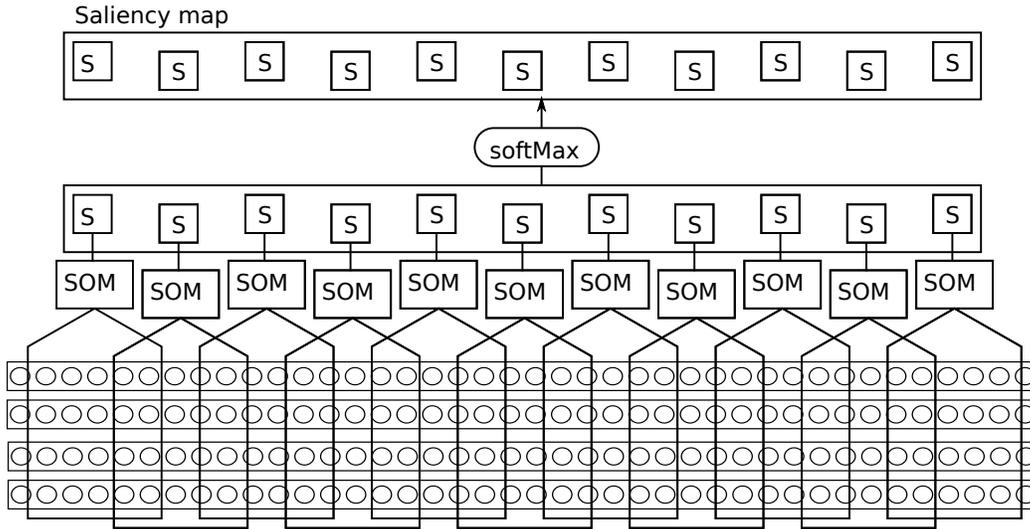


Figure 2.4: A saliency map based on local SOM representations

In the model we envisage, shown in Figure 2.4, each local SOM computes a distribution over its units at each moment in time; each SOM is also associated with a unit (denoted ‘S’), that computes a measure of surprise about this distribution. The saliency map is implemented as a medium in which these surprise units compete. We scale their activity to sum to 1, and then apply the softMax function described in Section 2.2.1, so the winning units have their activity strengthened, and the losing units have their activity weakened. This guarantees that the distribution of ‘most surprising’ locations on the retina is a sparse one.

I should say this is a very simple representation of saliency. Like many representations of saliency, it is not good at identifying *objects*: humans are much better at selecting objects as the focus of attention. I will mention some mechanisms later in this chapter that help allocate attention to object-like stimuli: in particular, in Section 2.6 I will consider how attention can be allocated to stimuli that move smoothly across the retina. However, our visual representations of objects are also trained when we interact physically with objects, by touching them and grasping them. I will discuss the visual representations learned through these mechanisms, and how they contribute to visual salience, in Section 4.2.3.2. In the meantime, the simple definition of salience just outlined will suffice to introduce a basic idea about how salience modulates visual processing.

### 2.3.2.1 Aside: an alternative definition of saliency

A very successful recent model of saliency implements a related idea, with some differences. In the model of Duan and Wang (2015), the notion of surprise is defined in relation to each *single unit* in the early visual system. Units represent simple visual features approximating oriented Gaussians.<sup>3</sup> A distribution is computed for the activity level of each unit, independently of other units: a measure of surprise can then be computed separately for each unit. However, there is also a requirement that units be active in a sparse pattern, so I’m not sure how different the SOM-based measure of surprise is from Duan and Wang’s measure.<sup>4</sup>

<sup>3</sup>These features are actually learned through an unsupervised mechanism slightly different from SOMs: details can be found in King *et al.* (2013).

<sup>4</sup>Duan and Wang’s measure has a few other very useful novel features. For one thing, their simple visual features are computed separately from several separate colour space representations of input images. (The importance of colour in saliency was emphasised in Achanta *et al.* (2008; 2009), and it is useful to pick up on this dimension of images.) Separate saliency sub-maps are created for different components of each colour space. For another thing, there are global computations that select the best saliency sub-maps based on their amount of ‘contrast’, measured using entropy; these again echo the global computations over whole images used in Achanta *et al.*’s

### 2.3.3 Using the saliency map to select visual information

Local SOM units provide the inputs to several visual pathways that will be discussed later. The processing in all of these pathways is modulated by spatial attention. The basic mechanism that implements this modulation is very simple; I will describe it briefly here. (Extensions to this mechanism will be given in Section 2.12, after a more detailed model of saliency has been introduced.)

Assume a particular visual pathway—say for object classification. Each local SOM provides input to a layer of units that compute ‘the next level of representation’ in that pathway. We simply need to assume that the inputs provided by local SOMs to this higher level of representation are gated by the saliency level of the SOM in question. If the saliency of the SOM is zero, the SOM will not contribute any inputs to the next level of representation. If it is 1 (the theoretical maximum for a saliency value), it will contribute its inputs in full. Naturally, a SOM unit that is not active will not contribute any inputs. So the saliency of a SOM is a genuinely modulatory influence: it can emphasise the visual features that the SOM has identified at the relevant location in the retinal image, but it cannot cause features to be ‘hallucinated’ at this location.

## 2.4 A mechanism for selecting the next focus of attention: the gaze-orienting SOM

The saliency map is a competitive medium: we can imagine a winning unit emerging as a result of this competition. This unit could be anywhere on the retina. One of the basic mechanisms in the attentional system is one where selection of a winning location in the saliency map triggers a ballistic movement of the eyes and head, that establishes the winning location on the retina. In this section we will talk about the circuit that does this.

The circuit responsible for moving the eyes and head has to perform two key functions. Firstly, it has to select a salient location from the many candidate locations on the retina-centred saliency map, and generate a movement of the eyes and/or head that brings the selected location onto the fovea. I will call this function its ‘orienting’ function. Secondly, it has to learn a representation of the saliency map that is *stable* over these movements, so it can keep track of salient locations over saccades. I will call this its ‘stabilising’ function.

I will use another SOM, which I’ll call the **gaze-orienting SOM**, to implement the required circuit. I will introduce the SOM in three stages. In Section 2.4.1, I introduce two more general preliminaries about SOMs, relating to recurrent architectures and reconstruction of inputs. In Section 2.4.2, I discuss how the gaze-orienting SOM learns its ‘orienting’ function—and how, in this process, it learns its own internal representation of ‘gaze direction’ that serves to link these representations. In Section 2.4.3, I discuss how it uses its internal representation of gaze direction to learn a representation of the saliency map that is stable over movements of the eyes and head.

### 2.4.1 Background: recurrent SOMs and reconstruction of inputs

#### 2.4.1.1 Recurrent SOMs

The gaze-orienting SOM is a variety of *recurrent* SOM. This is a SOM that takes a stream of inputs over time. (We will discretise time, so there is one input at each timestep.) A recurrent SOM takes a representation of its state at the previous timestep as an input at the current timestep. This simple device means it learns *sequential* patterns in its inputs as a whole, rather than just static patterns within individual inputs. In our model, we use an efficient computational implementation of a recurrent SOM called a **merge SOM** or **mSOM** (see Strickert and Hammer, 2005 for details).

The representations learned by an mSOM have the same localist character as those learned by a regular SOM—only in an mSOM, localist units come to represent particular stages in frequently-

---

methods. (Another relevant paper using similar methods is Liu *et al.*, 2011. See Borji and Itti, 2013; Borji *et al.*, 2014 for surveys.)

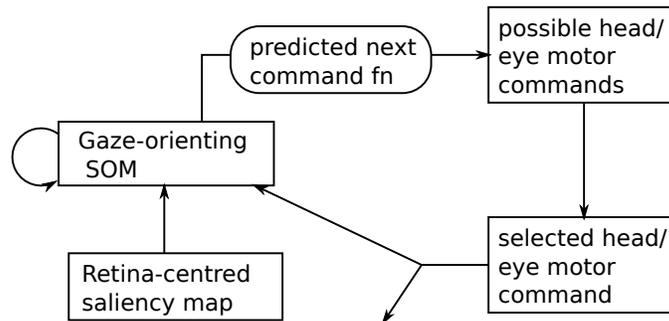


Figure 2.5: The gaze-orienting SOM

encountered sequences of inputs. A useful way of thinking of an mSOM is as a system that maintains some ‘internal state’, and updates this internal state as a function of its new inputs.

I will explain why the gaze-orienting SOM needs to have recurrent connections in Section 2.4.2.1.

#### 2.4.1.2 Using a trained SOM for pattern completion

Even though during training, a SOM simply learns common patterns between its input units, after training there can be directionality to its use. If a *partial* pattern is provided on the SOM’s input media, its connections allow it to make predictions, or inferences, about the likely values of the missing elements of the pattern. If the SOM takes inputs from several distinct media, as in the current case, we can present a pattern in some media, and leave other media unspecified, and it can *reconstruct* a likely pattern in the unspecified media. In fact, there is an attractive probabilistic interpretation of this reconstruction operation, that we can make under certain assumptions. We must first assume that each SOM unit keeps a record of the number of times it was the ‘winner’ during training. We must also assume that the units in the input media each represent a probability distribution—i.e. each represent all the possible values of some particular variable, exclusively and exhaustively. Under those assumptions, the reconstructed pattern in the ‘missing’ medium is the **Bayesian** estimate of the distribution over the missing variable, given distributions over the supplied variables.<sup>5</sup>

There are two attractive properties of this reconstruction operation. For one thing, queries can be posed for any input media. Any of them can be left blank: the reconstruction operation is symmetrical for all media. For another thing, if the distributions in the ‘input’ media are all sparse, the SOM can effectively compute a *many-to-many* mapping from values in the input media to values in the output media. The gaze-orienting SOM makes use of both these properties, as I will discuss in Section 2.4.3.

### 2.4.2 How the gaze-orienting SOM learns the orienting function, and a body-centred representation of gaze direction

The gaze-orienting SOM is shown in Figure 2.5. It takes three inputs. One is the retina-centred saliency map we have just introduced in Section 2.3.1. It also takes representations of the motor commands sent to the eyes and head. Finally, it takes a recurrent input: that is, an input from a representation of its own state at the previous time point.

I begin in Section 2.4.2.1 by discussing how motor commands to the eyes and head are represented. Then I will introduce the SOM’s learning algorithm bit by bit. In Section 2.4.2.2 I explain how the SOM learns its own internal representation of space ‘in the world’—in this case, a body-centred representation of ‘gaze direction’. (This method recurs many times throughout

<sup>5</sup>The frequencies associated with SOM units provide the all-important ‘prior distribution’ for this computation. More information is given in [Citation].

our wider model.) In Section 2.4.2.3 I will outline how the SOM learns to map this body-centred representation onto the retina-centred representations of salient locations it receives as input. In Section 2.4.2.4 I will explain how the SOM learns movements of the eyes and/or head that bring a selected location in the retina-centred saliency map onto the fovea. I conclude in Section 2.4.2.5 by describing how the trained SOM can map candidate retinal locations in parallel onto candidate head/eye movements.

#### 2.4.2.1 Representation of motor commands to the eyes and head

While many possible combinations of head and eye movements can move the agent’s gaze to a given orientation, empirical studies show that human eye and head movements are strongly coordinated (see Sağlam *et al.* (2011 for a review of evidence). One interesting aspect of human movements is that gaze is stabilised on the target *before the head movement finishes*: there is an initial phase of movement in which both head and eye move in the direction of the gaze target; and a second phase in which the head overshoots, and the eye ‘counter-rotates’ to compensate.

There is good evidence that this coordination results from optimisation processes, identifying the movements with best speed and accuracy. In the best-performing models, what is optimised is the variance of the endpoint of the action (Harris and Wolpert, 1998). Variance reflects the presence of noise in the motor system. A key insight is that variance increases with the strength of the motor signal (Harris and Wolpert, 1998); this discourages maximally strong motor impulses. At the same time, there is another ‘constant’ component of variance that increases linearly with movement duration (van Beers, 2007); this discourages overly long movements. There is a tradeoff between the strength of motor commands and the length of the resulting movements; an optimisation process can find movement parameters that hit the right point in this tradeoff.

For a system where the gaze can be shifted by both head and eye movements, this optimisation process finds a particular stereotypical pattern of head and eye movements for each gaze target (Sağlam *et al.*, 2011). The solution found by optimisation reproduces interesting properties of humans’ head/eye movements—in particular, the fact that gaze is stabilised on the target before the head movement finishes, as described above.

In the current discussion, I will assume that motor commands are specified as goal gaze directions, in a body-centred coordinate system.<sup>6</sup> From each goal gaze direction, a single optimal head/eye movement sequence can be selected.

Of course, this way of defining motor commands presupposes that units in the gaze-orienting SOM can be reliably *mapped onto* body-centred gaze directions. Given that the SOM also takes a retina-centred saliency map, this is far from obvious. In the next section I will discuss principles which (I hope) will allow the gaze-orienting SOM to learn a body-centred representation of gaze.

#### 2.4.2.2 How the gaze-orienting SOM can learn a body-centred representation of place

It is important to have some way of linking the SOM’s current pattern of activity to an actual position of the eyes or head. In the case of the head, there are proprioceptive cues that can provide an additional input. But there are no such cues for the eyes: somehow the eyes can maintain an accurate representation of their current angle in the head purely by ‘dead reckoning’. In our model, this happens naturally through the recurrent SOM’s learning. It is instructive to consider how this can happen, because this form of learning crops up in several other places in this document. I will illustrate with the case of eye position.

If our mSOM receives a sequence of head or eye commands, its internal state will represent the recent trajectory it has described. But there are strong *constraints* on what trajectories the eye can describe. These are imposed by the limits of the eye’s movement in each direction. For instance, consider a trajectory where the eye starts at a point at the extreme left of its range, and then

---

<sup>6</sup>This is the representation format used by Sağlam *et al.*, as far as I can tell, though their concern is not with coordinate system transformations are learned. They simply assume that gaze direction is eye direction + head direction, which implies all three directions are expressed in the same coordinate system.

the agent executes a motor command that brings it to a point at the extreme right of its range. A motor command with that magnitude can only be done if the eye is at one particular starting position (the extreme left), and it always results in the eye being at another unique position (the extreme right).<sup>7</sup> So this simple one-step trajectory uniquely specifies the eye’s current location. If the agent then executes another locomotion action, the new state the SOM gets into is also uniquely associated with a new eye location. After training, it turns out that even if there are no proprioceptive inputs identifying the current position of the eye in its orbit, *the states of the gaze-orienting SOM indicate this position*. This happens because its learning encourages it to represent frequently-occurring trajectories, and, for the reasons just discussed, these trajectories ultimately correlate with locations. Importantly, the locations learned by the SOM will be in a coordinate system associated with the body, because ultimately it is constraints due to the agent’s body (to which the head is connected) that limit the agent’s gaze.

We will reuse this trick throughout our model of the motor system; see also Chapter 3 on spatial navigation and Chapter 4 on motor control of the hand and arm.

### 2.4.2.3 How the gaze-orienting SOM deals with retinal input

A complication to the gaze-orienting SOM’s learning is that it also receives input from retina-centred representations. Can it still learn a body-centred frame of reference if its units must also represent retina-centred locations? I will argue here that it can.

Assume as a starting point that activity in the retinal saliency map is sharply focussed on the ‘most salient’ location. Assume further that the objects that generate the saliency map tend to retain their location across head/eye movements. And finally, assume that the most salient location in the saliency map remains tied to the same object over eye/head movements. (This is analogous to a situation where there is just a single object in the observer’s field of view, and no competition in the saliency map at all.)

In this situation, since SOM units are trained to represent body-centred gaze direction (through the SOM’s recurrent connections) but also inputs from the retinal saliency map, after training, they should come to encode particular *combinations* of retinal location and body-centred gaze direction. If the object generating the single salient location is stationary, there should in fact be a *constant mapping* between the salient retinal location and the body-centred gaze direction. For each body-centred gaze direction, there will be a unique retinal location, and vice versa.

### 2.4.2.4 The next command function

As shown in Figure 2.5, the next motor command is generated through a next-state prediction function that is trained separately from the SOM itself. While the recurrent SOM must learn something after each eye/head movement, in order to learn a body-centred representation of gaze direction, the next-command function is only trained in specific ‘lucky’ cases, where the action that is executed happens to foveate the selected peripheral retinal stimulus.

I assume the next-command function is trained in situations where the salient retinal location is some arbitrary peripheral location, and the agent has established some arbitrary body-centred gaze direction. In this situation, the winning SOM unit should represent precisely this combination of peripheral retinal location and body-centred gaze direction. We then generate a random head/eye movement. If it happens that the new salient retinal location *falls on the fovea*, we train the next-command function to map the active SOM unit onto this head/eye movement. If not, we simply continue making random head/eye commands (so that the dynamics of the SOM continue to encode body-centred gaze direction). Over time, the next-command function should learn a head/eye movement that foveates any arbitrary peripheral retinal location, regardless of what the agent’s current eye/head position is.

---

<sup>7</sup>We have to assume that motor commands that have no effect, because the eye or head is at its limit, are not passed into the SOM.

#### 2.4.2.5 Parallel processing in a trained gaze-orienting SOM

When all this learning is complete, it should be possible to use a less sharply focussed retinal saliency map, containing *several* salient locations, as input to the SOM. The learned connections into the SOM should now represent several different SOM units. We can imagine selection taking place within the SOM, rather than simply within the retinal saliency map.

This is helpful for several reasons. One is that SOM units are directly linked to motor commands, through the next commands function. If this function is simple enough, we can generate a distribution over possible motor commands too, so that the decision about which retinal location to foveate can be construed as a decision between alternative motor actions. This would allow us to implement a preference for small attentional shifts over large ones: if two potential targets are equally salient, we would rather saccade to the one closest to the current gaze direction.

Another benefit of envisaging competition between saccade targets happening in the SOM is that it provides a medium where bottom-up retinal salience measures can combine with top-down expectations about interesting (body-centred) locations. I will discuss this further in Section 2.4.3.

#### 2.4.3 How the gaze-orienting SOM learns a saliency map that is stable over eye/head movements

A saliency map is a device that supports systematic exploration of interesting locations in the visual field. If it is to do this, there must be a mechanism for keeping track of locations in the map that have already been attended to. There is a good model of the circuit that supports this. Each location in the saliency map is passed in parallel to a medium where locations compete, and a winner is selected, as described above. Then the winning location in the map is *inhibited*, and remains inhibited for some period of time, so that other locations can compete to be selected. Evidence for this **inhibition-of-return (IOR)** operation was provided by Posner *et al.* (1984); computational models of the mechanism are reviewed in Itti and Koch (2001).

This model of serial attentional processing via inhibition-of-return is complicated by the fact that selecting a winning item in the saliency map drives overt attentional movements of the eyes and head, as just described. If the saliency map is a retinotopic structure, inhibited regions will lose their correspondence with points in the visual field. Instead, we must envisage that the agent maintains a saliency map that is *stable* over movements of the eyes and head (see e.g. Rao and Ballard, 1996).

The gaze-orienting SOM can contribute directly to a circuit that creates a stable saliency map of this kind. As discussed in Section 2.4.2.2, units in the gaze-orienting SOM come to represent particular locations in the agent's field of view. However, since SOM units encode locations in a way that makes reference to head and eye position, there can be several SOM units that encode the *same* location. A saliency map that holds just a single representation of each point in a body-centred visual field must be a separate medium. We will call this new medium the **body-centred saliency map**. A circuit which connects this map to the gaze-orienting SOM is shown in Figure 2.6. In this circuit, the body-centred saliency map is simply another input to the gaze-orienting SOM.

For each point in the body-centred visual field, the gaze-orienting SOM has a whole set of distinct patterns that represent this point. It must learn to link each set of patterns to a single pattern in the body-centred saliency map. It does this in a special training mode, in which the principle that the body-centred saliency map must be stable over eye and head movements is simply enforced as an axiom. The agent begins by activating a pattern in the retina-centred saliency map with a static gaze. From this, a pattern of activity is reconstructed within the body-centred saliency map medium. Then the agent shifts his gaze in some arbitrary way, with a movement of the head and/or eyes, *but retains the pattern in the body-centred saliency map*, so it provides an input to the gaze-orienting SOM at the next point in time. Over time, with this training regime, the gaze-orienting SOM will come to map locations in the retina-centred saliency map *in parallel* onto locations in a saliency map stable over head and eye movements. This happens through a process called **cross-situational learning** (see e.g. Siskind, 1996). At

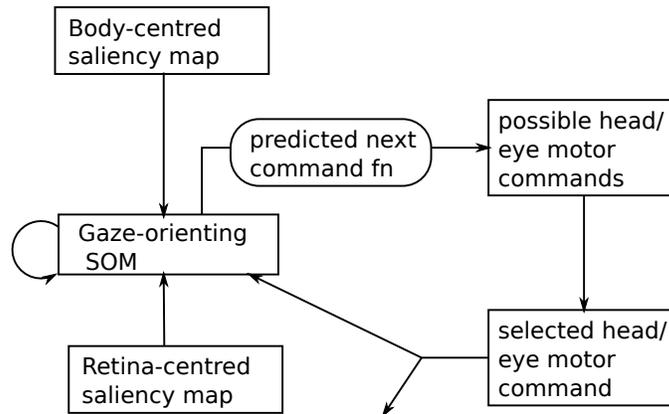


Figure 2.6: The gaze-orienting SOM’s interface with a saliency map that is stable over eye/head movements.

each iteration, *each* active unit in the gaze-orienting SOM is linked to *each* active unit in the body-centred saliency map—but only ‘the right’ units will be *consistently* associated. The others will just provide noise to the training. Note this requires the body-centred saliency map to be ‘reconstructed’ from the pattern in the gaze-orienting SOM, in the opposite direction to the arrow shown in Figure 2.6.<sup>8</sup>

Recall from Section 2.4.2.5 that the gaze-orienting SOM can be construed as the medium where a location to be attended to is *selected*, from amongst a set of competing alternatives. Now that the gaze-orienting SOM is connected to a body-centred saliency map, it can be thought of as mediating competition in a broader sense: it now functions as a medium where ‘bottom-up’ inputs from the retina can be compete against ‘top-down’ inputs from a more stable representation of salient locations that might be derived from something other than vision: for instance memory. In any case, after a location is selected, from amongst these competing bottom-up and top-down candidates, we can envisage that it is inhibited—in the *body-centred saliency map*, which is stable over eye movements. This inhibition operation is depicted by an arrow labelled ‘IOR’ in Figure 2.6.

#### 2.4.4 A circuit for covert attention

As already discussed in Section 2.3.3, the most salient location in the retina-centred saliency map is a location from which visual features are preferentially ‘selected’ and passed forward for further processing. This is the case even if the most salient location is not foveated. Attention biased towards a salient peripheral retinal location in this way is called ‘covert’ visual attention.

Note that the circuit for learning a body-centred saliency map just introduced in Section 2.4.3 also allows for a succession of salient points to be selected purely covertly via inhibition-of-return, without *any* overt movements of the eyes or head. This is certainly attested empirically—and it is known that the system supporting covert attentional shifts is connected to that controlling overt shifts (see e.g. Nobre *et al.*, 2000). In this case, note that IOR can also be implemented directly within the retinal saliency map. Indeed there is some evidence for a retina-centred IOR mechanism, as I will discuss in Section 2.5.1.

<sup>8</sup>To make learning efficient, it might be a good idea to impose the same amount of sparseness in the body-centred saliency map as in the gaze-orienting SOM.

## 2.5 Neural evidence for the media in the orienting circuit

### 2.5.1 The gaze-orienting SOM is in frontal eye fields

Our proposal is that the gaze-orienting SOM is implemented in FEF. (A distinct saliency circuit related to the reach system is described in Section 4.2.3.2; I will propose that this circuit provides a good model of intraparietal areas that encode location.)

There's debate as to whether FEF cells encode prepared eye movements, or a saliency map, or both (see your book for a summary). In our model, they can be interpreted both ways.

There is evidence that a midbrain structure called the superior colliculus is involved in sending reafferent copies of eye and head movements to the saliency map—at least the one in LIP. If the superior colliculus is damaged, LIP still represents salient locations, but only in a retinotopic frame of reference (see e.g. Sapir *et al.*, 2004).

### 2.5.2 Modulation of low-level visual feature maps by spatial attention

For this, refer to Moore and Armstrong's (2003) experiment showing that stimulation of FEF enhances representation of features in the corresponding area of early visual area V4.

## 2.6 Object tracking

The locations associated with objects in the agent's visual field do not just move about in discrete jumps, due to eye or head movements. They can also move about due to actual movements of the objects themselves, or due to movements of the agent in his environment. In these cases the retinal projection of an object in the world moves gradually: spatiotemporal continuity is an important part of the definition of 'an object'. In this section, we will consider how this principle is implemented as an axiom in the attentional system. (Later, in Section 12.3, we will see how it is implemented in the system that represents individuals in long-term memory.)

To support continued attention to objects over time, there has to be a second mechanism for shifting attention, often referred to as a **tracking** mechanism (citation). We introduce evidence for tracking in Section 2.6.1, and a model in Section 2.6.2.

### 2.6.1 Multiple object tracking

Pylyshyn and Storm (1988) showed that human observers can track a smallish number of objects as they move around the visual field.

### 2.6.2 A model of simple object tracking

We will model tracking mechanisms on the retina by assuming that the agent has already learned the gaze-orienting SOM described in Section 2.4, and the links that connect it to the body-centred saliency map (see Section 2.4.3).

In our model we envisage that there is a component of *recurrency* to the body-centred saliency map, so that it learns frequently-occurring sequences of salient locations. We model the body-centred saliency map as a recurrent SOM, that takes a representation of its previous state as input, as well as a pattern in the gaze-orienting SOM. In this connection, we will refer to it as a **tracking SOM** in the discussion below.

The tracking SOM needs to learn patterns of continuous movement in the visual field in a special training mode, in which the eye is stable: either fixating a single salient location, or possibly tracking the object at this location using pursuit eye movements. We also assume that a single pattern is selected in the gaze-orienting SOM, so only one object location is to be tracked. (We will relax this latter assumption later.)

When the tracking SOM is being trained, it is exposed to sequences of contiguous body-centred locations (as expressed in the gaze-orienting SOM), originating from continuous movements of the



### 2.6.2.2 Attentional focus on tracked items

Another interesting effect in the tracking map is that once the agent *starts* tracking certain objects, and updating based on its recurrent connections, as well as by the direct links to the gaze-orienting SOM, any *new* objects that appear are likely to be ignored. This is because the SOM is configured to represent only locations that are contiguous with locations that were present at the previous iteration—[because those are the only training data it receives consistently](#).

I’m not sure if this effect is found empirically. But it seems plausible that tracking a moving object makes it harder to divert attention to a newly appearing stimulus.

### 2.6.2.3 Momentum effects

A final interesting effect of the tracking SOM’s dynamics is that it should allow tracking across brief occlusions of the tracked item. This is because the state predicted by the next-state prediction function can be passed straight back into the tracking SOM even if there is no bottom-up evidence from the gaze-orienting SOM that confirms this prediction. (At least, this should be possible for a few timesteps.)

### 2.6.2.4 Tracking ‘exploding’ or ‘breaking’ items

An interesting perceptual scenario is one where a single object is monitored as it breaks into several pieces. The tracking map as described above should handle this scenario by beginning to track each of the pieces of the object. The key point is that there will be several different retinal motion cues at the currently tracked location, leading in different directions. In this scenario, the next-state prediction function will predict a set of several points adjacent to the current salient location, and these will be activated in the tracking map at the next moment.

I will say more about how breaking actions of this kind are represented in Section 3.9.2.

## 2.6.3 Multiple object tracking and object files

The above model is able to track a certain number of objects around on the retina. But it is not able to *individuate* objects while tracking them. There is evidence that human agents can represent a very small number of objects as individuals, while tracking them. The classic experiment demonstrating this is by Kahneman *et al.* (1992). In this experiment, subjects saw two boxes in a visual display: in each box, a letter briefly appeared. The boxes then moved smoothly to new locations, and a letter appeared in one of the boxes. It was found that subjects were faster to identify the letter in a given box if it was the same one that originally appeared in that box, and slower if it was different from the one that originally appeared there. This ‘object-specific preview advantage’ suggests that observers create representations of the boxes that survive movements of the boxes. These notional representations were dubbed ‘object files’ by Kahneman *et al.*

The neural basis of object files is still not known. In this section we will suggest one component of a neural model of object files.

## 2.6.4 A revision of the model: multiple tracking maps

The idea here combines the concept of discrete attentional operations selecting objects (Section 2.4) with the idea of object tracking. The main idea is that there are *multiple tracking SOMs*: let’s say two, for the time being. They are each basically a full copy of the tracking SOM shown in Figure 2.7, including the associated next-state prediction function. It is assumed that each of these specialises in tracking a single retinal location as it moves.

These tracking maps are activated one by one. First the agent attends focally to one object, and initiates a tracking map on this object. Then the agent attends focally to another object, and initiates a second tracking map on this second object. (Note that the tracking maps in our scheme are robust to head and eye movements, and so can continue to be tracked across saccades. There is

some evidence for this ability in human tracking; see e.g. Verfaillie *et al.*, 1994.) Now a particular token object—or a stand-in for such tokens, such as letters of the alphabet in Kahneman *et al.*'s experiment—can be associated with a *whole map*, and generate top-down expectations that obtain no matter where the tracked object is in the visual field.

It may be thought that this device of multiple tracking maps is somewhat profligate. However, the number of objects in experiments showing the object-specific preview advantage is severely limited; it is only two, or at the most three (see e.g. Liddle, 2010). In fact we will argue that two distinct tracking maps are built tightly into the architecture of the working memory system that represents events; for more on this, see Section 12.6.2.

## 2.7 Visual object classification mechanisms

In this section, I'll develop the idea of a 'convolutional network' or 'convnet' using SOMs. A standard convnet consists of alternating layers of neurons performing 'convolution' and spatial 'pooling'. These layers are normally trained using back-propagation, but they have also been designed using SOMs (see e.g. Vanetti *et al.*, 2013; Liu *et al.*, 2015). My proposal is loosely based on Liu *et al.*'s model, though my accounts of spatial pooling, training and cardinality blindness are novel.

### 2.7.1 Background: the brain's object classification system

There's a consensus that object categories are represented in the ventral visual stream, while their affordances are represented in the dorsal visual stream (see e.g. Bracci and Op de Beek, 2016 for a good recent survey). I will focus on the ventral classification process here; the dorsal process will be discussed in Chapter 5.

In the human ventral stream, the occipitotemporal cortex (OT) is a key location, analogous to monkey IT. In OT, the most visible distinction is between animate and inanimate objects, with a distinction in the animate category between faces and bodies (see e.g. Kriegeskorte *et al.*, 2008; Connolly *et al.*, 2012) and possibly also conspecific vs other faces and bodies (Caspari *et al.*, 2014). There is also good evidence that representations in the ventral visual pathway are progressively more complex, and abstract progressively over space, in the manner of a convolutional network (see Güçlü and van Gerven, 2015 for a recent study that makes this link explicitly).

The idea that the ventral visual pathway learns representations of object types using unsupervised methods, emphasising the most frequent or typical objects, is supported by a recent study by Jordan *et al.* (2016).

### 2.7.2 Recap: a layer of local SOMs detecting simple visual features

Recall from Section 2.2 that 'low-level vision' is implemented in a set of 'local SOMs', that tile the retina. Each local SOM learns to represent the commonly occurring combinations of simple visual features in its local area of retina. This layer of SOMs does something very analogous to the convolution later of a convnet. In fact, we could even envisage that these SOMs all 'share the same weights', as happens in a convnet (although this would not be biologically plausible, as discussed below).

### 2.7.3 A spatial pooling operation

In a standard convnet, the outputs of a convolution layer are passed to a 'spatial pooling' layer. The spatial pooling operation takes a set of local feature maps collectively covering a certain area of retina, and computes activity in a single isomorphic feature map, by activating the maximum value for each feature. This retains information about the visual features present in this area of retina, but loses information about exactly where they occur within this area. In biological terms, these correspond to a particular class of 'complex cells' in the primary visual cortex that respond

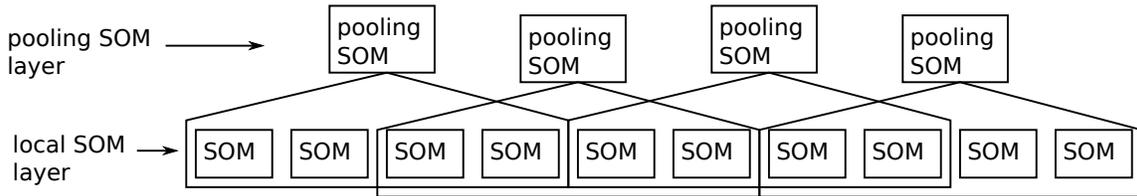


Figure 2.8: A layer of pooling SOMs taking input from the local SOMs. Each pooling SOM takes input from the local SOMs in a particular region of the retina, and learns to represent simple visual features wherever they appear within this region.

to a given simple visual feature anywhere within a small local retinal area (see classically Hubel and Wiesel, 1962).

There are many models of the process by which complex cells learn their tuning preferences; see Antolík and Bednar (2011) for a review. I will use a SOM-based scheme based on the model of Sullivan and de Sa (2012), which is in turn derived from the model of Földiák (1991).

In the scheme I propose, this operation is implemented by a **pooling SOM**, trained using the same clamping-while-tracking trick that’s used to train the body-centred saliency map (see Section 2.4.3). The retina is tiled with pooling SOMs, at a slightly coarser granularity than local SOMs. Each pooling SOM takes input from *all* the units in *all* the local SOMs within its area of retina, as shown in Figure 2.8. The activity of these input units is modulated by the activity of their associated saliency-map/SOM units, as already discussed in Section 2.3.3.

The procedure for training the pooling SOM layer takes advantage of the fact that objects in the world tend to project similar patterns onto the retina at successive time points. An important qualification is that these patterns might move from one point on the retina to an adjacent point, either because the eye is moving smoothly, or because the object is moving, or both. This means that if a pooling SOM is trained to use the same unit to represent simple visual features at successive time points, it will learn to respond invariantly to these features if they move smoothly over its associated region of retina.

Training proceeds as follows. First, the saliency map selects a local SOM. The units in this local SOM will provide a particularly strong input to the pooling SOM, because it is salient. A unit will activate in the pooling SOM, to represent the salient pattern. The key learning step is that this activated unit is *held active* at the next time point, so that it also becomes associated with whatever pattern is active at this next point.

Consider what happens when the tracked stimulus moves to an adjacent local SOM unit. That local SOM will represent the visual feature(s) best evoked by the stimulus. The visual features are likely to be pretty much the same as those evoked in the previous local SOM, since as a rule, objects don’t change much in the features they present from moment to moment. The principle that the same pooling SOM unit must represent the features of the tracked stimulus over consecutive moments, together with the contingent fact that the features of a stimulus tend not to change from moment to moment, means that pooling SOM units come to respond to some particular visual feature anywhere in a certain retinal area. In a convnet, the link between features in different local SOMs would be hard-wired into the architecture, through the principle of weight-sharing. In this case, the spatial pooling operation just involves finding the maximum activity over *a known unit* in all the local SOMs in a particular region. In the network I’m thinking of, the pooling SOM does pooling, but also identifies correspondences between local SOM units: these correspondences are not hardwired into the architecture.

There are also complex cells that respond to particular patterns of moving stimuli. I will discuss those in Section 7.1, in the context of a model of the visual pathway for action perception.

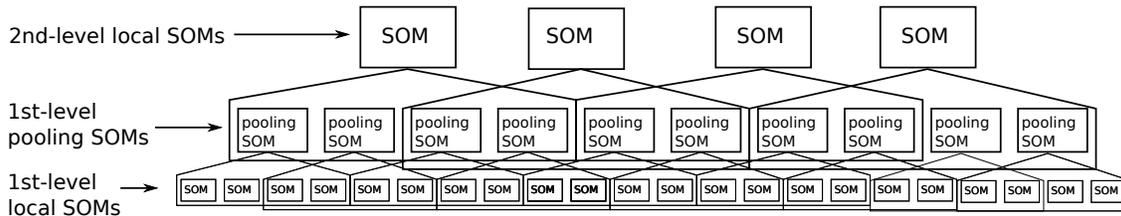


Figure 2.9: A layer of ‘second-level’ local SOMs taking input from ‘first-level’ pooling SOMs.

### 2.7.4 Iterating on the above scheme: hierarchical classification

The above sections described a layer of SOMs tiling the retina that compute visual features in their local area, and a layer of pooling SOMs at a slightly coarser spatial scale that do an analogue of spatial pooling. We can envisage the remainder of the ‘object classification network’ as being further iterations of this basic scheme. There is a layer of ‘first-level’ local SOMs, that pass input to a layer of ‘first-level’ pooling SOMs. These pooling SOMs provide input to a layer of ‘second-level’ local SOMs, with their own ‘second-level’ pooling SOMs. The principle of iteration is illustrated in Figure 2.9.

The second-level local SOMs are different from the first-level local SOMs in two respects. Firstly, they represent visual features over a wider receptive field. Secondly, their units encode visual features that are more complex: they will come to represent frequently-occurring combinations of the features encoded by the first-level SOMs. Note that because the first-level features are represented in pooling SOMs, there is some latitude about the exact retinal location of these features. The second-level SOM knows that a given first-level feature occurs within a certain *region* of its visual field, but does not know exactly where in this region it appears.

The alternating layers of feature-combining and spatial-pooling layers work on exactly the same principle as a regular convnet, trained using backpropagation, and I expect to see the same kinds of benefit. However, if the network’s weights are trained using SOM learning rather than backpropagation, there are two additional benefits we can expect. Firstly, note that after training, the units of the network at every level will hold *localist* representations of high-level visual features. This has several advantages in terms of how representations of visual features can be manipulated. I will discuss some of these advantages in the model of property representations I develop in Section 12.3.1. Secondly, if the components of the network are SOMs, it should be possible to propagate activity *downwards* in the network as well as upwards, through the process of ‘reconstructing the SOM’s inputs’ outlined in Section 2.4.1.2. Top-down propagation of activity happens through the same kind of mechanisms as in Hinton’s restricted Boltzmann machines.<sup>9</sup>

### 2.7.5 ‘Training’ of the classification SOM

In the scheme discussed so far, all learning in the classification SOM is unsupervised: it will learn certain high-level representations, based on the stimuli it sees most frequently. This is useful, but we also need a way to skew its learning towards representations that are useful for it. To begin with, we will assume these representations are simply given to the network as category labels, by an external ‘supervisor’. Later, in Section 4.2.3.2 and Chapter 13, we will see how the agent can generate training labels for himself.

A simple SOM-based classifier is sketched in Figure 2.10. In this network, I assume the second-layer pooling SOM holds representations that are sufficiently complex that they could identify categories of object: for instance, dogs and cats, or more plausibly, categories of simple geometric shape like squares and triangles. This SOM’s receptive field subtends the whole retina, so it could recognise instances of a given object type at any retinal location. The representations that emerge

<sup>9</sup>In fact a SOM-based classifier is a lot like a RBM classifier: this is something I should probably make more of.

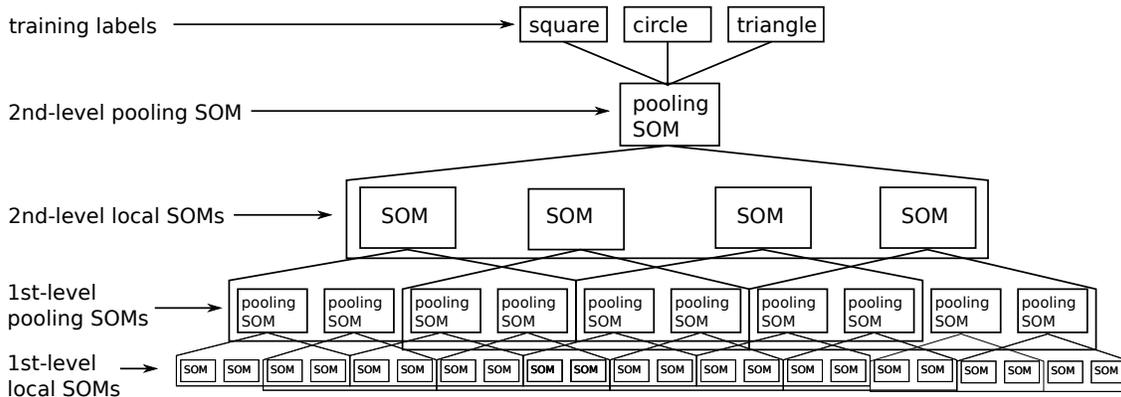


Figure 2.10: A simple SOM-based classifier, using two layers of alternating local SOMs and pooling SOMs. The training labels at the top are also provided as inputs to the SOM (during training).

in this SOM without supervision will identify high-level patterns in the stimuli presented on the retina. But we would like to be able to bias its learning towards certain categories: for instance, circles and squares. To this end, the top-level pooling SOM also takes input from a set of labels, provided by an external teacher, that identify useful categories: in this case, categories of geometric shape. Whenever a circle is presented, the label ‘circle’ is also provided as input to this top-level pooling SOM. The SOM’s representations will thereby be biased towards representing the visual patterns that reliably co-occur with the presented labels. After training, if we just provide a visual pattern, the activity in the top-level pooling SOM can be used to ‘reconstruct’ a label. Conversely, a label can be provided as input, and used to reconstruct an expectation of the associated visual stimulus.

### 2.7.5.1 Size-invariance in the classifier

Cells at the end of the inferotemporal visual pathway are frequently insensitive to the *size* of stimuli they represent, as well as to their retinal location (see e.g. Tanaka, 1997). To model this, it is useful to recognise that both the first and second level pooling SOMs can potentially hold complete object type representations. The second-level SOM holds representations of more complex retinal objects than the first-level SOM—and in virtue of this, the objects will also subtend larger areas of retina, since their atomic components are of the same size. However, we can also envisage a layer of first-level SOMs that take input from coarser-grained visual features, whose receptive fields are of the same size as those of the second-level SOMs. This additional layer is shown in Figure 2.11, along with a corresponding coarse-grained pooling SOM. This pooling SOM holds representations of a comparable *complexity* as the original layer of 1st order local SOMs, but of a larger retinal size.

Imagine there is an object of a type that can be represented by a 1st order local SOM. If the object is far away, it will be represented by a fine-grained 1st order SOM. But when it approaches, it will be represented by a coarser-grained 1st order SOM. We would like the same unit to fire in both cases. To train such a unit, we can envisage another kind of pooling SOM, that takes input from both fine-grained and coarse-grained local SOMs. In Figure 2.11, this is called the ‘size-invariant properties SOM’.<sup>10</sup> As with the other pooling SOMs, this SOM is constrained to hold the same unit active over a certain period of time. If objects regularly move in depth in relation to the observer, without rotating, this SOM should learn to represent object types in a way that is invariant to their size. I will assume that it is this SOM that receives supervised

<sup>10</sup>The size-invariant properties SOM also receives input from the 2nd level pooling SOM, so it can recognise large complex things as well as large simple things. The current section focusses on size-invariance, rather than the construction of hierarchically complex object representations.

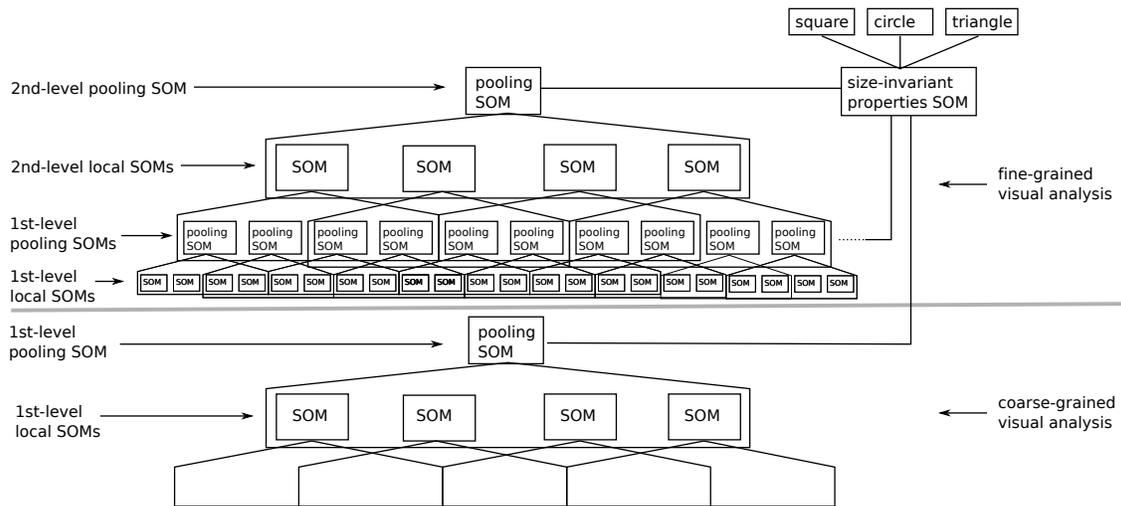


Figure 2.11: The classifier extended with a circuit to learn size-invariant object type representations. (The coarse-grained and fine-grained visual analyses should be understood as operating on the same area of retina.)

training.

### 2.7.5.2 Training at intermediate levels in the classifier

Consider again the case where an object category is associated with a complex enough visual pattern that it must be represented in a 2nd level SOM, which represents combinations of simpler 1st level SOM patterns. It might be useful if the 1st level SOM patterns could be biased towards representing the *constituent visual features* of more complex object types. There is a limit to how much information about a complex object type is provided by a low-level visual feature: for instance, a low-level textural feature like ‘furry’ is probably consistent with a number of object types. But nonetheless, furriness provides *some* information about object type.

I think maybe the size-invariant properties SOM provides a mechanism for biasing low-level SOMs towards features associated with object types. Say the 2nd level pooling SOM identifies a complex object type like a dog. The size-invariant properties SOM can perhaps be thought of as supplying another input to the 1st level SOMs, identifying this high-level object type. This may have the effect of skewing the representations learned by these SOMs towards those that are associated with high-level object types. Importantly, if there are *no* associations between low-level features and high-level object types, the 1st level SOMs will just learn to ignore the high-level information. So it doesn’t do as much damage as would be caused by trying to associate simple visual features with high-level object types using something like the delta rule. In this case, the error will always be high, and wildly variable, and weights will be changed by large amounts in arbitrary directions.

An existing model that implements some of the above ideas is that of Lefevbre and Garcia (2008). This is a SOM-based model of face recognition. The training data consists of images of several different individuals (many separate pictures of each individual). In a preprocessing stage, a set of ‘regions of interest’ (ROIs) is computed for each face in the training set. Each of these is a small square patch of the image. These ROI patches are then used to train a single SOM, whose units come to represent the patterns associated with these different ROIs. Of course, each ROI comes from a particular image of a particular individual—so in representing ROIs, the SOM’s units carry some information about individuals. Crucially, after training, each SOM unit is labelled with the number of times it was the winner *for each individual in the training set*. During a test phase, when presented with ROIs from an unseen image of one of the individuals, these numbers

can be plugged into the Bayesian formula to compute the probability distribution over faces, conditional on the ROI data. With a large enough SOM, this results in quite good identification of individuals: between 94 and 98%. (This method is quite close to the method suggested in the current section—except that the supervised component of training is implemented by labelling SOM units with category information, rather than simply by supplying the category as an input to the SOM. Apart from that, the Bayesian model of reconstruction is extremely similar.)

### 2.7.6 Cardinality blindness in the classifier

Consider again the classification network shown in Figure 2.11. Assume we train the network on small-sized squares, appearing all over the retina. After training, the network should be able to respond to a square appearing anywhere on the retina with the object category ‘square’. But interestingly, it will also respond this same way to *more than one* square. The classification network is (I believe) *blind to cardinality*: it responds the same way to a set of homogenous squares no matter how many items there are in this set.

Cardinality blindness also appears to be a property of the representations at the end of the classification pathway in IT (see e.g. Nieder and Miller, 2004). Skipping ahead temporarily to a discussion of language, the cardinality blindness of the IT classifier makes it ideally suited for delivering the denotations of noun stems. A common noun comprises a ‘stem’ denoting a category, and an ‘inflection’ denoting a number—in the simplest case, singular or plural. In semanticists’ parlance, an inflected noun denotes a homogeneous set of objects of a given type: its stem denotes the type, but says nothing about the size of the set, while its inflection provides information about the size of the set but says nothing about the type of its elements. These correspondences will be discussed in more detail in Chapter 22.

The network described so far provides some of the mechanisms that enable a homogeneous group of objects to be classified as such. But we have not yet considered how *attention can be allocated* to a retinal area containing a group of objects. This will be discussed in Section 2.11.

## 2.8 Multiple processing streams for visual property types

In the model I have in mind, there are several classifier mechanisms of the kind described in Section 2.7, that specialise in different *types* of visual information. The types of information are the kind of thing that can be separately reported as ‘properties’: the kinds of attribute that would be expressed in adjectives. I will assume there are four classifier streams, that represent what I’ll call ‘shape’, ‘colour’, ‘texture’ and ‘affective properties’. I will propose that the medium representing object categories sits above these four classifier streams, and learns frequently-occurring combinations of properties. (Evidence that the ventral visual cortex computes information on several independent dimensions comes from Haxby *et al.*, 2011.<sup>11</sup>)

I don’t want to suggest that the visual system doesn’t combine information about properties of these different types: it clearly does. My main point is that the shape, colour, texture and affect pathways compute properties that are kept somewhat distinct from the ragbag of things that define basic-level object categories, which allows them to be reported separately in adjectives.

I won’t dwell on the linguistic distinction between nouns and adjectives here: that will be discussed in more detail in Chapter 22. But I’ll mention one linguistic fact that is of relevance in motivating this model of separate streams. Macoir *et al.* (2015) find that patients with progressive semantic dementia have spared retention of property-denoting adjectives compared to category-denoting nouns: this suggests (to me!) that categories are *composed of* complexes of properties.

In this section I will review evidence that shape, colour, texture and affective properties are computed in somewhat distinct visual pathways. In Section 2.9 I will sketch circuits that compute these properties, and a circuit that performs object classification, that partly recruits these circuits.

---

<sup>11</sup>This analysis of the brain’s representations into independent components is extended to the whole brain in Guntupalli *et al.* (2016), so it’s not clear that the result says anything about object representations specifically.

### 2.8.1 Neural representations of shape and shape adjectives

Using fMRI adaptation, Kourtzi and Kanwisher (2001) found that shape is represented in the lateral occipital cortex. But many other properties are also represented here. More recently, using a multivoxel decoding technique, Coutanche and Thompson-Schill (2015) found that the shape of a visually presented stimulus could be decoded from right V4 (but not its colour). So there is evidence that in some regions, at least, shape may be computed separately from colour.

Another recent study (Bracci and Op de Beeck, 2016) found evidence that object shape is represented somewhat independently from object category in various areas of the ventral stream (occipitotemporal cortex, parahippocampal place area and transverse occipital sulcus) as well as in areas of the dorsal stream (superior parietal lobe). The experimental stimuli here varied in 2D shape, not necessarily in 3D shape. The shape-sensitive areas distinguished between ‘elongated’, ‘round’ and ‘triangular’ shapes (which in 3D could in many cases be termed ‘conical’). They might all plausibly relate to motor primitives, as I discuss in Section 5.4. While shape and category representations were dissociable, there were also strong correlations between them, suggesting that shape is a cue to category membership. However, the authors stress that object category representations in the ventral cortex often abstract away from shape. Also, interestingly, ‘elongated’ shape representations in ventral cortex frequently abstract away from orientation—a helpful revision of early models of IT object representation (e.g. Tanaka *et al.*, 1996).

A similar finding is reported by Freud *et al.* (2015): in a study of subjects with lesions to the ventral visual pathway, they show that representations of 3D shape are computed in parietal cortex independently from those computed in ventral cortex.

### 2.8.2 Neural representations of colour and colour adjectives

There are two brain regions that seem to be particularly activated by colour (though neither of them *only* represent colour): an area in the lingual gyrus labelled V4, or sometimes VO1, and an area in the medial fusiform gyrus, labelled V4 $\alpha$  (see Murphey *et al.*, 2008 for a review). Murphey *et al.* tested a human patient with electrodes implanted in the latter region: these responded more to chromatic than non-chromatic stimuli, and were selective for colour, particularly blue/purple; stimulation of these electrodes in the absence of visual stimuli elicited the percept of a blue/purple colour near the fovea, without any associated shape percept.

There’s evidence that colour percepts activate V4 $\alpha$ , even when induced by a colourless stimulus (Morita *et al.*, 2004). Coutanche and Thompson-Schill (2015) found that the colour of a visually presented stimulus was represented in anticipatory activity in right V4, even prior to a stimulus being presented. (Shape was also represented in V4, however.)

### 2.8.3 Neural representations of texture

Cant and Goodale (2007) showed subjects nonsense objects that varied in form and in surface properties (naturalistic textures like marble or wood grain, in different colours). While attention to form activated the lateral-occipital (LO) area, as expected, attention to texture (and to a lesser extent, colour) activated two distinct regions, the collateral sulcus and the inferior occipital gyrus. This was replicated in a fMRI adaptation study (Cant *et al.*, 2009). (Neither study found extrastriate regions sensitive only to colour.) Cavina-Pratesi *et al.* (2010), using a fMRI technique examining selective rebound from adaptation, found distinct areas sensitive to form (LO), colour (anterior contralateral sulcus and lingual gyrus), and texture (posterior contralateral sulcus). This texture region is different to those found by Cant *et al.*, perhaps because the textures in the experimental stimuli were tactile textures (rough, smooth, spiky etc). (All Cant *et al.*’s stimuli were smooth and required vision to be identified.) The areas identified by Cavina-Pratesi *et al.* were also consistent with data from patients with selective deficits in identifying colours, shapes and forms.

Cavina-Pratesi *et al.* (2010) also found regions in the fusiform gyrus that were selective to

specific combinations of shape, colour and texture.<sup>12</sup>

#### 2.8.4 Neural representations of affective properties

Emotive perceptual stimuli (including objects) activate a network of brain regions, including the amygdala for fear (Öhman, 2005), the anterior insula for disgust (Wicker *et al.*, 2003), the ventral occipital cortex for facial attractiveness (Chatterjee *et al.*, 2009) and the medial orbito-frontal cortex for aesthetic judgements (Ishizu and Zeki, 2011).

There is a certain amount of evidence for an autonomous system processing the emotional valence of perceived stimuli, though exactly what this processes must be carefully delineated (see Pessoa, 2005) for a good review). There's quite good evidence that at least some forms of perceptual processing of emotional content require focal attention; but there is also evidence that some aspects of emotional content are processed pre-attentionally. (This makes sense if you consider that emotional content contributes to a computation of salience, that might trigger an interruption of the agent's current task, à la Corbetta and Shulman.) There's also evidence that while some emotional processing is early, other more sophisticated processing is late: again see Pessoa (2005) for a discussion. With these caveats, here is a discussion of the autonomous system for processing emotional properties of objects.

**Fear and the amygdala** There are two routes to the amygdala during perceptual experience: a 'fast' route through subcortical regions (the superior colliculus and pulvinar), and a 'slow' route via temporal cortex (see Öhman, 2005). Fear-eliciting stimuli activate the amygdala prior to visual cortex (see again Öhman, 2005).

**Attractiveness and the ventral occipital cortex** The attractiveness of human faces is also recognised very fast. Olson and Marshuetz (2005) presented subjects with pictures of attractive and unattractive faces at very short exposures: short enough that they reported they could not see a face at all. (Stimuli were masked to eliminate persistence of vision.) Subjects' judgements of facial attractiveness was nonetheless significantly better than chance. There are also areas of the visual pathway that appear to be automatically activated by facial attractiveness, in particular the ventral occipital cortex, which includes the fusiform face area and the lateral occipital cortex (see Chatterjee *et al.*, 2009).

**Dangerousness and superior temporal cortex** There's also evidence that right superior temporal sulcus classifies the 'predacity' (dangerousness) of animals, relatively independently of their type; see Connolly *et al.* (2016).

**Aesthetic properties and orbitofrontal cortex** There is some evidence that a domain-independent representation of 'aesthetic beauty' is activated in medial orbito-frontal cortex. For instance, Ishizu and Zeki (2011) found activity in this area for both pictures judged beautiful and excerpts of music judged beautiful, and the activity in this area was proportional to the judged degree of beauty. (Orbitofrontal cortex is also differentially activated by stimuli judged beautiful and ugly; see Kawabata and Zeki, 2004.) Again there seem to be both fast and slow responses in this area: some judgements of beauty are considered, while others are fast (though there's no indication at all that this area generates evaluations that arrive faster than neutral class labels).

**Look-ahead to linguistic/syntactic issues** There's evidence that reading emotional adjectives activates the (left) amygdala more than reading neutral adjectives—and that positive adjectives elicit more activity than negative ones (Herbert *et al.*, 2009). Emotional adjectives also

---

<sup>12</sup>The fusiform gyrus runs right along temporal cortex, above the inferior temporal gyrus and below the parahippocampal gyrus, but it would be generally regarded as 'inferior temporal' (IT). Neurons in monkey IT cortex are well known to be sensitive to complex stimuli, including combinations of colour, form and surface pattern (Komatsu and Ideura, 1993).

activate the (left) inferior/middle occipital gyrus (BA18/19), and superior frontal gyrus (BA9). So some of the same areas that are activated by emotion-eliciting stimuli.

The emotive responses that are *post* focal attention, but *pre* classification, are particularly interesting, because that's where they sit in the syntactic structure, according to our general hypothesis.

### 2.8.5 Neural representations of category-relative geometric properties

Category-relative properties are things like 'big', 'small', 'thin', 'wide', 'narrow' and so on. They pick up on aspects of geometric form, but importantly, they do so in a way that highlights differences between the form of the described object and that of the generic object of its type. They are interesting linguistically, because they appear to occupy a specialised syntactic position (see Chapter 22 for discussion). I will say more about them in Chapter 5 (I think in Section 5.4).

I'll focus on size, since this has been most studied. There is some evidence that size differences are represented in the lateral occipital cortex and in the parahippocampal place area, with the former area specialising in small sizes and the latter in larger ones (see Cate *et al.*, 2011, Gabay *et al.*, 2016).<sup>13</sup> There is also evidence that absolute size of objects is represented in the organisation of occipitotemporal cortex, with big objects (e.g. a bathtub, a table) represented more medially and small objects (e.g. a paperclip, a cup) represented more laterally (see Konkle and Oliva, 2012). However, this distinction explicitly measured absolute size of objects. When subjects were asked to imagine large *examples* of the objects they were tested on, there was no alteration of the medial/lateral preference.

It's interesting that the ventral visual pathway is increasingly less sensitive to (retinal) size; the anterior area is fairly size-independent (see e.g. Watson *et al.*, 2016). The flip side of this is that earlier parts of the pathway are distinctly sensitive to size (see again Watson *et al.*).

## 2.9 A circuit for representing object categories and properties

### 2.9.1 A look-ahead to predicative propositions

In here, refer forward to Section 11.2 for an account of how propositions about object properties are expressed as WM episodes. (That's a preliminary to the syntactic account of predication given in Section 17.2.) The basic idea is that the *WM episode medium* can hold a trace of the processes through which an observer attends to and classifies an object and then attends to one of its properties.

## 2.10 Representations of object categories, and associated attentional operations

In this section I will discuss in more detail the size-invariant properties SOM shown in Figure 2.11, and how this can function as the basis for learning a set of object categories.

To recap from Section 2.7.5.1, the size-invariant properties SOM's units hold representations of the complexes of visual features that have been learned by the local SOMs in the visual system. (Its representations will be skewed towards the features encoded by the local SOMs that are currently selected as salient, as discussed in Section 2.3, and as will be further discussed in Section 2.11.) Importantly, the size-invariant properties SOM generalises over the features encoded by local SOMs, in two respects. Firstly, since it takes input from pooling SOMs, rather than directly from local SOMs, its units generalise over the location of features. Secondly, since it takes input from pooling SOMs operating at different spatial granularities, its units also generalise over the size of

---

<sup>13</sup>Amit *et al.*, 2012 is also somewhat relevant: though this is about distance rather than size.

features. This means that the representations of visual properties in its units have a high level of abstraction: they will represent the properties associated with an object in a way that is relatively invariant to its position in the visual field, and to its distance from the observer.

As already mentioned in Section 2.7.5, I assume the units in the size-invariant properties SOM are sufficiently complex and abstract that that *groups* of these units can be associated with representations of *object categories*. (The examples of object categories in Figure 2.11 are simple shapes: ‘square’, ‘circle’, ‘triangle’. In a more naturalistic context, I envisage groups of units in the size-invariant properties SOM could be associated with natural kinds like ‘dog’, ‘cat’, ‘car’, ‘cup’ and so on.) I begin in Section 2.10.1 by discussing the kinds of representation that will be activated within the size-invariant properties SOM. Then in Section 2.10.2 I will discuss how these representations can provide the training data for an unsupervised method for learning object categories. Finally in Section 2.10.3 I discuss how learned object category representations can serve in a mechanism for drawing attention to particular properties of token objects.

### 2.10.1 The rich property complex

Imagine an object is placed in the visual field, and the retinal region it occupies is associated high salience. The pattern of activity in the size-invariant properties SOM will represent the object’s visually observable properties.

These properties will be of different kinds. Some of them will reflect the object’s type: for instance, dogs will be likely to project many visual properties characteristic of dogs. Others will be more idiosyncratic: if we are looking at a particular dog, it may have certain properties that are unusual for dogs (for instance, an unusual head or body shape, or unusually long or short fur, or it may be unusually spotty or stripy or wet or muddy). All of these properties will be represented simultaneously in the size-invariant properties SOM.

I will call the collection of properties evoked by a token object in the size-invariant properties SOM the **rich property complex** or **RPC**. In future I will also refer to the size-invariant properties SOM as the ‘RPC SOM’.

#### 2.10.1.1 Aside: an alternative model of the RPC

An alternative way of thinking of the RPC is to model it as a *set* of SOMs that receive inputs from quasi-independent parts of the visual processing pathway. We could imagine one part deals with shape, while another separate part deals with colour analysis, and yet another part deals with visual texture, and another part with affective properties. There is some fairly good evidence for this: see in particular Cavina-Pratesi *et al.* (2010). If these visual features are computed separately, and represented in separate SOMs, we can envisage the RPC as the collection of these SOMs. Each SOM will then encode its own probability distribution: in this model, the RPC would be a collection of distributions, rather than a single distribution.

### 2.10.2 Unsupervised learning of object categories

Assume that the classification system is presented with a large number of token objects as training data. For each of these objects, the RPC SOM will activate a collection of units representing its collection of visual properties. If we assume that the objects presented are of different types, there will be some correlations among these properties: the properties associated with dogs will tend to co-occur, as will those associated with cats, cars, cups and so on.<sup>14</sup>

I will envisage that the medium representing distinct object categories that interfaces with the RPC SOM is a SOM itself: I will call it the **dominant property assembly SOM** (or **DPA SOM**). Through regular SOM learning mechanisms, units in the DPA SOM will come to hold

---

<sup>14</sup>Actually, if these properties co-occur then if the RPC is a single SOM, we expect it to learn them directly. It’s only if the RPC is a collection of separate SOMs that it’s worthwhile having a higher-level SOM sitting above the RPC. Thanks to Martin for that point!

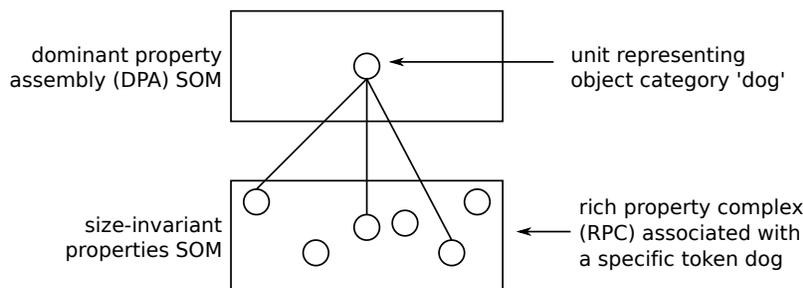


Figure 2.12: A rich property complex (RPC) encoding the properties of a token dog in the size-invariant properties SOM is presented as input to the trained dominant property assembly (DPA) SOM. The active unit in the DPA SOM represents the object category ‘dog’. The arcs linking this unit to properties in the RPC identify the ‘normal’ properties of dogs, on the basis of which it was selected as the winning category.

localist representations of object categories. As usual, activities in the DPA SOM are normalised to sum to 1, so a pattern of activity in this SOM represents a distribution over object categories.

A scenario where a token object’s RPC is presented to the trained DPA SOM is shown in Figure 2.12. The properties in the RPC include some properties that dogs commonly possess, and others that are ‘idiosyncratic’ properties, possessed by this particular dog but not by most dogs. Because units in the DPA SOM come to represent commonly occurring combinations of DPA properties, the winning DPA unit is activated by the prototypical ‘dog’ properties of the token dog: this unit can therefore be thought of as representing the object category ‘dog’. The same unit will be activated by other token dogs.

### 2.10.3 Property-level IOR: a mechanism for attending to properties of objects

Consider the situation illustrated in Figure 2.12, where a token object represented by a pattern of activity in the RPC SOM is identified with a particular object type in the DPA SOM. In this situation, it is interesting to consider how the observer’s attention can be drawn to the token object’s idiosyncratic properties—that is, those properties that are *not* typical for objects of the identified type. This is an important attentional process: we can imagine that it is the sort of process that is reported linguistically in sentences like *This dog is wet!*, *This dog is hairy!* and so on.<sup>15</sup>

In the current model, the mechanism for isolating idiosyncratic properties is quite simple. We can just *inhibit* activity in the RPC as a function of the weights of the selected DPA unit back into the RPC—because the strongest weights identify the most prototypical properties for this DPA unit. I will call this operation **property-level inhibition of return**, or **property-level IOR**. Its effect in our example scenario is illustrated in Figure 2.13.

I assume that the reduced sets of idiosyncratic properties isolated by property-level IOR provide additional training data for the DPA. This allows DPA units to identify regularities that obtain within these remaining properties. In this scheme, properties like ‘dirty’ and ‘hairy’ are read from the same medium as object categories like ‘dog’ and ‘cat’. What distinguishes object categories is not the medium they occupy, but the *temporal order* in which they occupy this position.

Note that the mechanism of property-level IOR makes strong assumptions about the *localist* nature of property representations in the RPC. If the pattern in the RPC is a distributed representation, of the kind that a traditional backpropagation algorithm will learn, the subtraction operation I am envisaging will not work at all. But given that the object classification net-

<sup>15</sup>The process reported by sentences like *This dog has a long tail!* is a little different, since it involves attention being drawn to a *part* of an object. I will suggest a model of this process in Section 2.12.1.2.

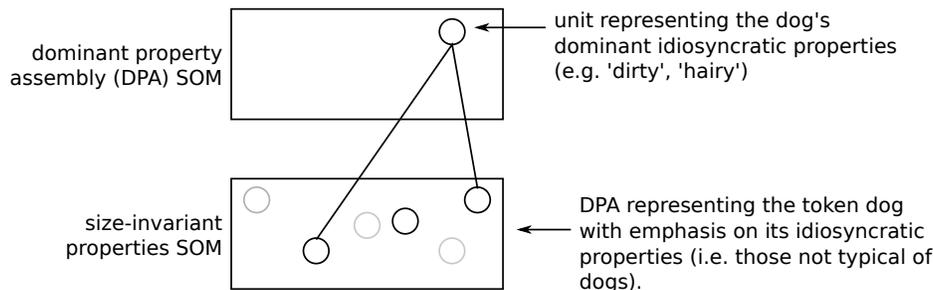


Figure 2.13: The pattern of activity in the RPC and DPA SOMs after property-level IOR takes place. The ‘typical’ dog properties which are inhibited are shown greyed out. The newly winning assembly of properties activates a new unit in the DPA SOM.

work is built entirely out of SOMs, we can rely on these units holding localist representations of commonly-occurring properties. In fact, a key reason why I am using SOMs to build the visual classification network is because they support this IOR operation, which is crucial for modelling how an observer’s attention can be drawn to particular properties of objects.

[Somewhere in here, I can refer to Arora *et al.*’s (2015) fMRI results about brain regions associated with predicative sentences. They find activity in the superior temporal gyrus (compared to identity statements, that also feature the copula *is*). I think I should look for studies with other comparisons, though.]

#### 2.10.4 Using property-level IOR to learn a hierarchical type system

Property-level IOR is useful for identifying the kind of properties that are reported in adjectives: for instance ‘hairy’, ‘dirty’. However, skipping briefly ahead to linguistic considerations, the syntactic contexts where adjectives can appear in a clause are also contexts where a certain type of noun phrase can appear: thus alongside *The dog is hairy*, we have *The dog is a spaniel*. In these positions, nouns are interpreted as predicates, that are semantically very much like adjectives. I will discuss these ‘predicate nominals’ in more detail in Chapter 22. (Incidentally, the existence of predicate nominals is a strong argument for supposing the denotations of nouns and adjectives are read from the same neural medium.) In the current section, I just want to discuss how it is that property-level IOR can result in an object type being activated in the DPA. This is work based on a model by Gorman and Knott (2016).

The key idea is that property-level IOR allows the observer to learn *finer-grained* representations of object categories than can be learned by the DPA SOM by itself. The DPA SOM is hard-wired to identify the strongest correlations in the inputs it receives, and allocate units to representing these. In the model of Gorman and Knott, these correlations identify a ‘basic level’ of object categories. However, there are also categories that descend below the basic level: for instance, there are sub-types of dogs and cats, characterised by particular correlations of properties *in addition to* those that identify basic level categories. Property-level IOR gives the DPA SOM a chance to learn these more subtle correlations. In Gorman and Knott’s model, after property-level IOR, a new DPA unit is selected based on the remaining properties, and then this unit is trained on *all* the token object’s features, including the inhibited ones. This training mode allows the DPA SOM to allocate units to represent finer-grained object categories, while also allowing these units to represent the prototypical properties of their parent category.

During training, the system of categories learned in the DPA SOM starts off very broad, and due to IOR, becomes more specific. This process mirrors the developmental process in infants, who begin (for instance) by classifying all domestic pets / vehicles / into the same category, and as they gain experience with these categories, develop finer-grained categories. In summary, the DPA SOM, in conjunction with the property-level IOR operation, provides scope for learning a

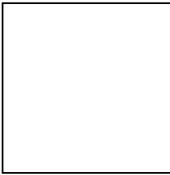
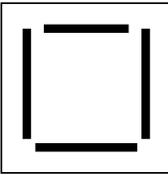
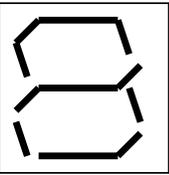
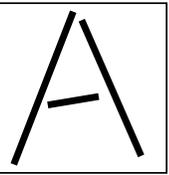
	Receptive field size	Example patterns using default frequency spatial features				
large						
small						

Figure 2.14: An illustration of ‘default frequency’ spatial features

rich hierarchical system of object categories.

## 2.11 Saliency at higher levels of the visual object classification network

In this section I will introduce some extensions to the basic object classification pathway introduced in Section 2.7, that incorporate measures of saliency associated with higher-level representations in the pathway. All of these extensions reflect the fact that saliency is not just allocated to the smallest-sized retinal regions, but larger regions too. These extensions are mostly reworkings of ideas in Walles *et al.* (2008; 2014).

### 2.11.1 Local SOMs at different spatial granularities

As already mentioned in Section 2.2.2, ‘local SOMs’ tile the retina at a range of spatial frequencies. The lower-frequency local SOMs have larger receptive fields, and take input from correspondingly lower-frequency atomic visual features, as illustrated in Figure 2.11. I will first make this notion of ‘correspondingly lower-frequency features’ more precise.

I will assume there is a range of **default frequencies** for a local SOM’s input features, which is a function of its receptive field. The default frequency features for a given SOM are those that naturally function as ‘building blocks’ of shapes within the SOM’s receptive field. A toy example is shown in Figure 2.14. For the large receptive field shown on the top row, the oriented visual features needed to form shapes filling the field have a certain range of spatial frequencies. For the smaller receptive field shown on the bottom row, the visual features must be correspondingly smaller. It would not be useful, or economical, to represent the larger shapes in the top row with combinations of the smaller visual features on the bottom row.

### 2.11.2 Saliency maps at different spatial granularities

Recall from Section 2.3 that a set of local SOMs tiling the retina functions as a saliency map. If the retina is tiled by local SOMs at several different spatial granularities, then we can envisage *several distinct saliency maps*, operating at different granularities.

In most circumstances, SOMs in these distinct saliency maps compete against one another. After competition, the most salient regions on the retina might include regions of several different sizes. This is a common circumstance in natural scenes: there is often something large competing for attention, but also something small.

However, there are some circumstances where small features on the retina contribute to the saliency of larger regions. These circumstances involve the phenomenon of ‘popout’, in which a stimulus stands out from its surroundings, and the phenomenon of ‘perceptual grouping’, in

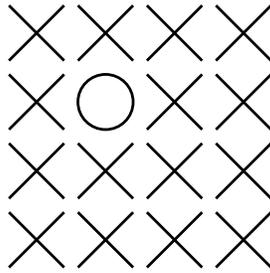


Figure 2.15: An illustration of ‘popout’. The O emerges clearly from a field of Xs.

which several small features form a single perceptual ‘figure’, distinguished from the surrounding ‘ground’. I will consider popout in Section 2.11.3. I will consider two of the classic grouping effects grouping by textural homogeneity and grouping by spatial proximity, in Sections 2.11.4 and 2.11.5.

In our model, both popout and grouping are implemented in the saliency map. If a region pops out from its surroundings, it is represented as having higher saliency than neighbouring regions. If stimuli occupying several retinal regions are grouped, this means a single larger region encompassing these smaller regions is represented as having high saliency, so these two stimuli will be passed to the object classifier together, rather than individually.

### 2.11.3 Popout

All other things being equal, a stimulus is salient if it is different from nearby stimuli. This effect is most readily seen in displays like that shown in Figure 2.15: the item that is different ‘pops out’, and is encoded as a figure, while the nearby stimuli are treated as ground. While popout was originally modelled as an all-or-nothing phenomenon, experiments by Duncan and Humphreys (1989) showed that it admits of degrees. The key variables determining how easily a target stimulus emerges from a field of distractors are the homogeneity of the distractors and the similarity of the target to the distractors. Figure 2.16a illustrates how the target O emerges less readily when the distractors are heterogeneous; Figure 2.16b illustrates how a target that is similar to the distractors emerges even less readily.

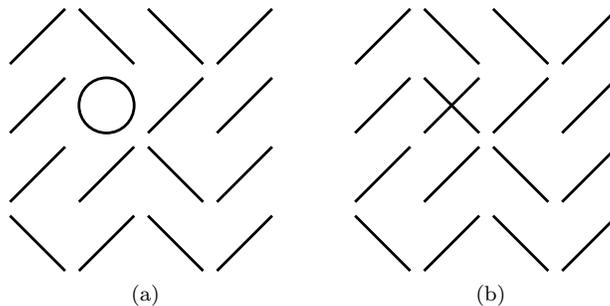


Figure 2.16: Graded popout effects found by Duncan and Humphreys (1989). (a) A target (O) emerges less clearly in a heterogeneous field of distractors. (b) A target (X) emerges even less clearly when it is similar to the distractors.

The basic popout phenomenon is modelled by a circuit that increases the salience of a local SOM as a function of the *difference* between the pattern it encodes and the patterns encoded by its neighbouring SOMs. Since local SOMs represent distributions over visual features, this difference is readily expressed using KL divergence. A simple computation of popout saliency for a given local SOM would be the average KL divergence between its pattern and the pattern in

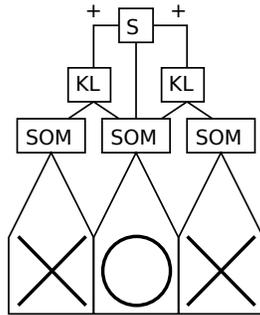


Figure 2.17: A circuit for enhancing the salience of a local SOM by its contrast with adjacent SOMs. The local SOM whose salience is computed is the one representing the circle stimulus in the centre. The adjacent SOMs represent Xs flanking this circle. (The box labelled ‘KL’ denotes the KL divergence between the local SOM and its adjacent SOMs.)

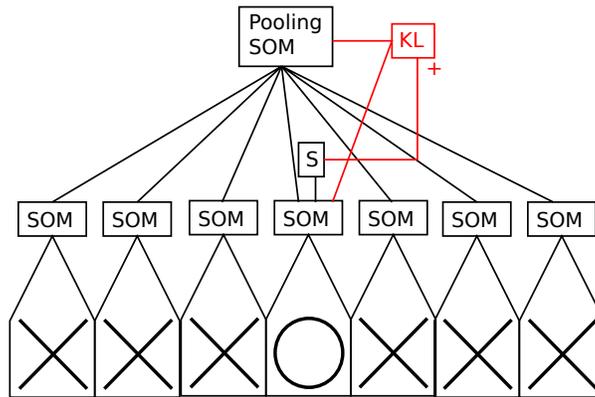


Figure 2.18: A circuit for enhancing the salience of a local SOM by its collective contrast with objects in its local region. The KL divergence computation is shown (in red) for the SOM representing circle; a similar computation is applied for each local SOM.

each of its neighbouring SOMs. This metric identifies a measure of ‘local contrast’: the principle is illustrated in Figure 2.17.

However, this metric does not capture the effects on popout due to the homogeneity of the larger area in which a stimulus appears, that were discovered by Duncan and Humphreys. To model this larger effect, we can make use of the pooling SOM introduced in Section 2.7.3. In that section, I described how pooling SOMs are trained using a *single instance* of a visual feature, as it moves around a local area of retina. But now consider what happens when a group of identical stimuli are present within a given area of retina. If there is a unit in the pooling SOM that responds to a single instance of this stimulus anywhere within this area of retina, this unit will respond very vigorously if many instances of the feature are present in this area.<sup>16</sup> Such a unit effectively represents a homogeneous field of stimuli. If we compute the KL divergence between the pattern expressed in a local SOM *and the pattern expressed in the pooling SOM to which it contributes*, as illustrated in Figure 2.18, this provides a measure of salience which will increase as the homogeneity of distractors increases. This measure will be low for all local SOMs representing the element repeated in the homogeneous field, but high for a local SOM representing some unusual

<sup>16</sup>This assumes spatial pooling uses a ‘sum’ operation, rather than a ‘max’ operation. As far as I understand, both sum and max can be used in the spatial pooling layer of a convnet; I think to model textural homogeneity grouping, we have to use sum.



Figure 2.19: An illustration of grouping by textural homogeneity

element, that only occurs rarely in this field.

#### 2.11.4 Grouping by textural homogeneity

Textural homogeneity also features positively in the salience computation. All other things being equal, two retinal stimuli are more likely to be grouped, and encoded as a ‘figure’, if they are similar to one another. For instance, the two crosses in Figure 2.19 are more likely to be grouped than a cross and a square, even though they are roughly equal distances apart.

To model grouping by textural homogeneity, we can again make use of the pooling SOM introduced in Section 2.7.3. But here, high homogeneity within a given region should increase the salience of *this whole region*, compared to neighbouring regions. Specifically, if there is a strongly active unit in the pooling SOM, indicating the presence of multiple instances of a certain pattern in a certain retinal area, we want to increase the saliency of this whole retinal area.

We first need a measure that indicates that there is a strong pattern of this kind in the pooling SOM. Note that the units in the pooling SOM can also be treated as encoding a probability distribution over pooled features. If a pooling SOM’s receptive field contains a heterogeneous set of visual features, it will have a relatively uniform distribution (that is, a relatively high entropy, or low confidence). If its receptive field contains a homogenous set of features, it will have a relatively sharp distribution (that is, a relatively low entropy, or high confidence). So high confidence in the pooling SOM should indicate the presence of textural homogeneity within its receptive field.

We now need to consider what measure of saliency is increased by this textural homogeneity. I suggest it should increase the saliency of a regular local SOM that represents the combinations of *simple* visual features within the same region of retina represented by the pooling SOM. The patterns this local SOM represents will be coarser grained than those represented by the pooling SOM. (The pooling SOM represents patterns that can occur in any *sub-region* within its receptive field, while the large-scale local SOM represents patterns that occupy the *whole* of its receptive field.) We can now envisage two measures of the saliency of this receptive field. The pooling SOM’s confidence measures saliency deriving from textural homogeneity within this receptive field. The large-scale local SOM’s ‘surprise’ as to its current pattern of activity measures saliency due to the ‘global’ figure present within this receptive field (see Section 2.3.1). Given that we are using local SOMs to define saliency maps, we can simply stipulate that the confidence of a pooling SOM with a given receptive field is added to the surprise of the local SOM with the same receptive field. This will boost the saliency of regions with textural homogeneity.

The circuit I have in mind is illustrated in Figure 2.20. It is related to the circuit for learning size-invariance shown in Figure 2.11. There are circuits that analyse a given retinal region at two spatial scales. The lower circuit comprises a single coarse-grained local SOM analysing the figure at a given region. The upper circuit comprises a pooling SOM covering the same region, taking input from finer-grained local SOMs covering the region. The coarse-grained SOM will be salient if the coarse-grained figure it represents is ‘surprising’, as discussed in Section 2.3.2. The fine-grained pooling SOM covering the same region will be salient if there is a fine-grained figure that is *frequently repeated* within this same region—especially if this figure is also ‘surprising’.

Note that a region with textural homogeneity can represent a figure in two different senses. These are illustrated in the stimulus shown in Figure 2.21. This stimulus has both a **global form** (‘A’), and a **local form** (‘X’). The global form will be identified by the coarse-grained local SOM that represents the region. The local form will be identified by the pooling SOM that represents the region. I will talk more about how the classifier identifies the global and local form of visual stimuli in Section 2.12.2. For the moment, I am simply modelling how the textural homogeneity

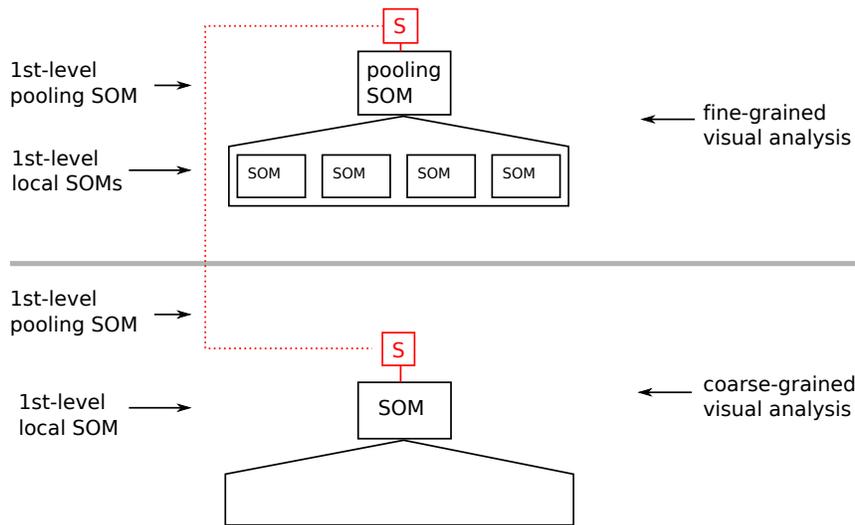


Figure 2.20: A circuit that increases salience for regions with high textural homogeneity. The salience of the upper pooling SOM contributes to the salience of a local SOM covering the same region at a coarser spatial frequency. The local SOM detects the ‘global form’ of a stimulus in this region; the pooling SOM detects its ‘local form’.

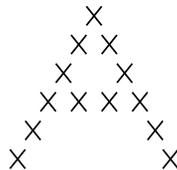


Figure 2.21: An illustration of local and global visual form. The global form of this stimulus is ‘A’; the local form (i.e. the form of its homogeneous texture elements) is ‘X’.

of a region contributes to its saliency.<sup>17</sup>

### 2.11.5 Grouping by spatial proximity, and a representation of ‘parts’ of a visual stimulus

The spatial proximity of stimuli is another cue to perceptual grouping. For instance, consider Figure 2.22: here, the crosses are more likely to be grouped with their nearby square, rather than with each other. The stimuli grouped within a single salient region here have the flavour of



Figure 2.22: An illustration of grouping by spatial proximity

<sup>17</sup>It is also important to consider what happens when there is an area of textural homogeneity that extends beyond the scope of a single pooling SOM, and covers several pooling SOMs. In this case, several large-scale local SOMs with adjacent receptive fields will have their saliency increased. In the normal scheme, these saliencies would compete—but in Section 2.11.5 I offer a scheme whereby adjacent salient regions collectively activate the saliency of a still-larger retinal region. I’ll assume large areas of homogeneity are recognised as salient by this adjacency-based mechanism, working on top of the regular homogeneity mechanism.

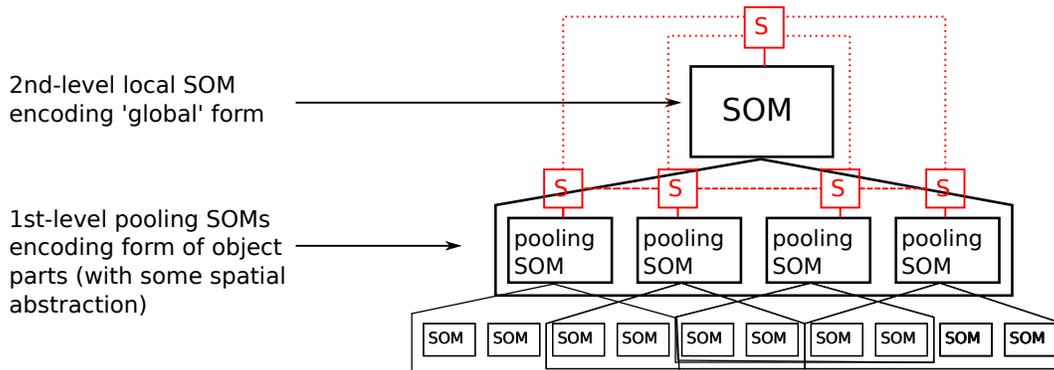


Figure 2.23: A circuit showing how saliency of figures occupying several close/adjacent small regions (representing ‘parts’ of a larger form) can contribute to the saliency of the larger region they occupying (representing the ‘whole’ form). Horizontal dashed lines represent modulatory effects of the saliency of pooling SOMs on the saliency of their neighbouring pooling SOMs. Dotted lines represent small contributions of the saliency of ‘parts’ to a ‘whole’. (These will only significantly influence the saliency of the whole if there are several salient ‘parts’, making up some sizeable proportion of the region occupied by the whole.)

‘compound objects’, each made up of several *parts*. The parts are also in some sense representable in isolation; however, their proximity induces the perceiver to represent them as parts of some larger whole.

To model grouping by proximity, there should be a way whereby two adjacent SOMs that both have a strong ‘surprising’ signal to be selected together, and passed to the classifier together. The classifier would then be able to process coarse-grained aspects of the ‘global form’ of the combined object—for instance, the characteristic features that identify the overall shape of a car. It would also be useful for it to be able to identify the car by its component parts—for instance, wheels, windscreen, bumpers.

To implement this kind of grouping, it is important to consider the fact that fine-grained local SOMs representing object parts interface with a coarser-grained local SOM representing a whole object *through a layer of pooling SOMs*. The coarser-grained SOM’s representation of object parts is delivered through these pooling SOMs, and only indirectly through the finer-grained local SOMs. The pooling SOMs allow some flexibility as to the exact spatial relationship between the object parts. Notwithstanding this flexibility, I assume that the relevant notion of spatial proximity applies *between pooling SOMs*, rather than directly between the finer-grained local SOMs. To this end, we need a metric of saliency that operates on pooling SOMs. I assume we can use exactly the same measure that applies to a local SOM: the KL divergence between the expected distribution over activities in the pooling SOM and the distribution that is currently evoked within it.

With these preliminaries, I propose grouping by spatial proximity is implemented through a combination of two mechanisms. The first mechanism is lateral connections between the saliency measures of adjacent pooling SOMs. I assume the saliency of a given pooling SOM (computed as just described) is *modulated* by the saliency of its neighbouring pooling SOMs (at the same spatial frequency), so that two adjacent pooling SOMs which are both individually salient have their saliencies enhanced. (Importantly, this modulation is more than just lateral excitation: a pooling SOM’s saliency is only boosted by the saliency of an adjacent pooling SOM if it *already* has some measure of saliency.)

The second mechanism is one whereby the saliencies of all the pooling SOMs in a given retinal region all *contribute* to the saliency of the larger-scale local SOM whose receptive field covers this larger region. These new saliency metrics are illustrated in Figure 2.23.

This hierarchical pooling of saliency has the effect that a given region can become salient *either* because it contains a single surprising ‘figure’ composed of features at its default frequency, *or*

because it contains a sufficient number of smaller surprising figures identified by finer-grained local SOMs—especially if these finer-grained SOMs are adjacent. A ‘sufficient number’ should be some fairly large proportion of the total number of finer-grained local SOMs in the region. (Naturally, the region can also become salient if it has a mixture of a surprising global form and surprising adjacent salient parts.)

Note this hierarchical pooling of saliencies only applies between SOMs with a particular ratio of spatial frequencies. This ratio is different from those discussed in the model of visual texture discussed in Section 2.11.4: the spatial scale of the ‘texture elements’ making up a visual stimulus can be much smaller than the spatial scale of the ‘adjacent parts’ of a visual stimulus. This difference in spatial scale is apparent in Figures 2.21 and 2.22: the crosses that form the texture elements of the ‘A’ in Figure 2.21 are much smaller in relation to the whole figure than the crosses that form the ‘parts’ of the figures in Figure 2.22.

I also assume that the transfer of salience activity between adjacent spatial frequencies works top-down as well as bottom-up. If the larger region is salient because of a large-scale figure in this region, some measure of this salience is passed top-down to all the SOMs that represent its constituent sub-regions (at the relevant ratio). This is helpful in highlighting patterns in these local SOMs that represent the ‘parts’ of the stimulus at the larger region. This top-down mechanism should ensure a transitive transmission of salience down to the very lowest-level SOMs within a selected larger region. (Though there will be a focus on those sub-regions which are also salient for other reasons, naturally.)

As should be clear from this section, the grouping-by-proximity mechanism provides a meaningful way to define the ‘parts’ of a complex retinal stimulus. I will conclude with two more general suggestions about the role and nature of these ‘part’ representations.

#### **2.11.5.1 A mechanism for representing salient regions with different forms**

Firstly, I suggest that the grouping mechanism discussed in this section is also useful in defining salient regions that have a component of *form* to them. In many models of salience, salient regions are uniformly circular. But objects come in different shapes; ideally we want to select a region shaped like the object for processing by the classifier, so that arbitrary regions of background close to the object can be withheld. The mechanism which increases the salience of a pooling SOM as a function of the salience of its neighbours is well suited to selecting a particular *group* of *connected* pooling SOMs as the salient ones within a given larger region, so that information is passed from *these* pooling SOMs to a higher-level SOM.

#### **2.11.5.2 3D representations of object parts**

Secondly, I should emphasise that the current definition of ‘object parts’ is very vision-centred; ‘parts’ are defined purely as retinal regions containing two-dimensional visual stimuli. We also need to represent objects as three-dimensional geometric shapes; the notion of ‘part’ that will be required in this representation will be very different. I will discuss representations of three-dimensional object geometry in Chapter 5. In Section 5.7 I will discuss how a representation of 3D object parts can be *mapped* to a 2D retinotopic representation of parts.

## **2.12 Mechanisms for attending to components of a selected salient region at different spatial frequencies**

In Section 2.11 I described two perceptual grouping mechanisms, that encourage objects with internal spatial structure to be represented as salient regions: grouping by textural homogeneity (Section 2.11.4) and grouping by spatial proximity (Section 2.11.5). In each case, having selected a large region, it should be possible to focus attention on the smaller figures within the region. In this section, we will consider these processes. I suggest two separate mechanisms, one for focussing attention on the ‘parts’ of a compound stimulus made up of multiple adjacent (and

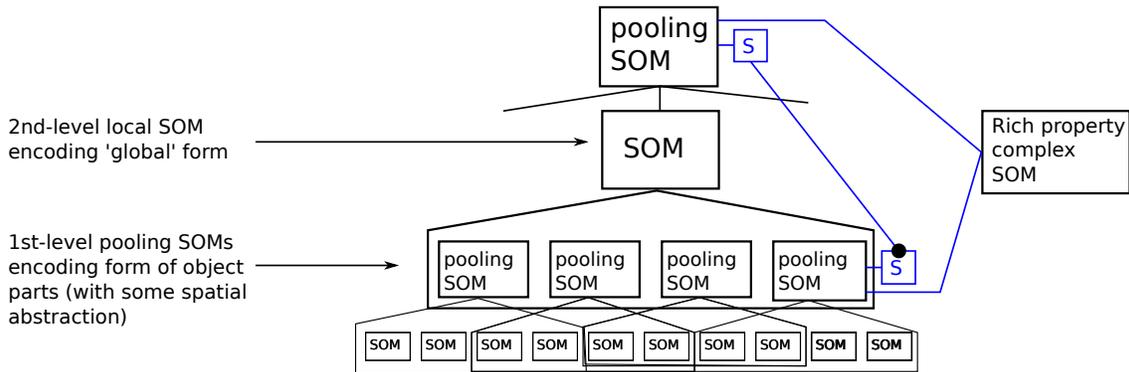


Figure 2.24: A system for classifying a large form and for classifying the form of its parts, supplemented with an ‘outward’ saliency mechanism (shown in blue) for determining which SOM reports the ‘result’ of classification.

potentially heterogeneous) local figures; the other for focussing attention on the ‘local form’ of a texturally homogeneous stimulus. These mechanisms will be discussed in Sections 2.12.1 and 2.12.2 respectively.

### 2.12.1 A mechanism for attending to wholes and parts of objects

Assume the scenario discussed in Section 2.11.5: a high-level local SOM subtending a large region represents a whole car, and a set of lower-level local SOMs representing sub-regions within this region (at the relevant ratio) represent parts of the car: wheels, windscreen, and so on. The higher-level SOM is salient in part because the lower-level SOMs are salient, and pass their salience up to the larger region, as shown in Figure 2.23. In addition, the salience of the lower-level SOMs should enable them to pass their outputs up to the higher-level SOM, so that information about the parts of the object can help identify the whole. However, it’s important to bear in mind that these local SOMs can also communicate with the *top level* of the classifier, the size-invariant properties SOM (i.e. the rich property complex SOM). Both the higher-level SOM and the lower-level SOMs can deliver output to the rich property complex SOM, as shown in Figure 2.24. We don’t want the representations of object parts evoked in the lower SOMs to *compete* with the representation of the whole object in the higher SOM. What we see in the large selected region is a car, not a wheel or a bumper. A finesse is required, whereby information about parts from the lower salient SOMs is passed ‘up’ to the higher SOM, but is *not* passed ‘out’ to the rich property complex SOM. (At least, not to begin with.)

To achieve this, I assume there are *two separate saliency mechanisms* threaded through the visual pathway: one for communicating information ‘up’ to higher levels, and one for passing information ‘out’ as a result of classification. The ‘upwards’ saliency mechanism picks a single spatial region, of a certain specific size. (The saliency of sub-regions within this region can contribute to the saliency of the larger region, as discussed above.) The ‘outwards’ saliency mechanism implements competition between SOMs representing regions of different sizes. The competitive rule here is simple: the SOM representing the largest region wins, to ensure we read out an object category from a SOM representing the whole object, rather than one of its parts.

Outward saliency mechanisms are illustrated in Figure 2.25 by the lateral (blue) saliency measures on SOMs. (Note these are applied to pooling SOMs rather than directly to local SOMs, since ‘read-out’ is done from pooling SOMs.) The blue arcs leading to the rich property complex SOM provide channels allowing the result of classification to be read out from different pooling SOMs. The outward saliency mechanism implements competition between SOMs representing regions of different sizes, with the SOM representing a larger region inhibiting those representing smaller regions. Note that this competition does not stop the SOMs representing parts from passing their

information to the SOM representing the whole.

Having identified the whole object as a car, however, there should be a mechanism to attend to its parts. At this point, I assume the salience of the large SOM is inhibited, *and the pooling SOMs at the level below compete with one another*. This involves disabling the lateral modulatory links between them.

We now need a principle for selecting the most salient object part. We could simply rely on the normal salience metrics for the local SOMs to achieve this, based on ‘surprise’ (see Section 2.3.2). But there is a more interesting possibility. What we would really like is a mechanism that draws an observer’s attention to the parts of an object that are surprising *as parts of the currently-identified object*.

For the sake of concreteness, assume that the activated pattern in the high-level SOM represents an object category (e.g. a car), and that the pooling SOMs that provide the input to the high-level SOM represent parts of an object (e.g. the wheels, windscreen, bumper of a car). Also assume that the high-level SOM representation of ‘car’ is an orientation-specific one, so that different units represent a car at different orientations/poses. Now imagine that we use the top-level pattern to *reconstruct* the patterns in the lower-level SOMs which are most likely to have activated it. This will generate a representation of the patterns we *expect* in these lower-level SOMs, *given that* the higher-level SOM represents a car with a particular orientation. (The orientation-specificity of the high-level SOM pattern is important, because cars with different orientations have different parts at different retinal locations.) We can do something quite useful with these patterns: we can generate a new measure of the ‘surprise’ of the actual patterns found in all of these lower-level SOMs. This measure assesses our surprise about these actual patterns not in absolute terms, but *conditionally* on the fact that an object of a certain category, at a certain orientation, has been identified at the level above. Recall from Section 2.3 that the saliency associated with a SOM is given by a measure of how surprising the activated pattern is. But the measure of surprise in that section was a *prior* measure, assuming no additional information. The new surprise measure I am introducing here is a *posterior* measure, taking into account what occupies the larger region of retina. What’s more, since each lower-level SOM represents *a different sub-part* of this larger region, there can be different expectations about what occupies each lower-level SOM, which make *different low-level features surprising* in different places within the larger region. If the larger SOM hypothesises a car, this will induce expectations about wheel-shaped features in certain places, about ‘window’ features in other places, ‘bumper’ features in other places, and so on.

There are two interesting uses for this measure of the visual features expected in different sub-parts of a retinal region, *given* a hypothesised object category. I will conclude by outlining these.

### 2.12.1.1 A mechanism for testing a hypothesis about the object category

Used by itself, a convolutional network sometimes makes catastrophic mistakes about object categories, even if it is normally very reliable (Nguyen *et al.*, 2015). For instance, in one case, an abstract pattern of black and yellow stripes was recognised with high confidence as a ‘school bus’. Maybe the reason why humans do not make such errors is that after hypothesising an object category, we attend to sub-parts of the hypothesised object to see if we find what we expect.

The expectations associated with sub-regions which I introduced above could be directly used in a verification system of this kind. The idea would be to treat the sub-regions within an object as a saliency map—that is, as a set of competing regions—in its own right. Bottom-up saliency already highlights those regions containing positive figures—especially those, that are connected or adjacent. But on top of this, we could activate a measure of saliency based on expectation: sub-regions where there are stronger expectations could be made more salient, so our attention is drawn first to these (in a special ‘hypothesis-checking’ phase of object classification). It is fine for an object to have a few unexpected parts—indeed, I will discuss this case below. But I assume we need some minimum number of the strongly expected parts of the the hypothesised object to be present, to confirm our hypothesis.

This mode of ‘attention to expected parts’ might also be useful in the representation of generic

facts about object parts. If we activate a representation of the category ‘car’ top-down, along with a selected retinal region, we should be able to generate similar expectations about the parts we will find. Assuming we are working in some ‘imagination mode’, where top-down expectations suffice by themselves to activate representations in the visual system, we can readily imagine stepping through sub-regions of the selected region, activating representations of the expected sub-parts of a (generic) car. I will talk more about generic propositions in Section ??.

Martin notes that if the local SOM that represents properties characteristic of a car *takes inputs* from finer-grained local SOMs that hold representations of parts of a car, it should already be immune to the kind of catastrophic errors I’m discussing in its first guess, and no hypothesis-testing should be needed. I need to think about this. . . but one possibility is just that the immunity arrives because the SOM takes input from finer-grained local SOMs that can under different attentional circumstances represent whole objects. Maybe that’s not the case with regular convnets—and it’s this that makes them susceptible to catastrophic errors.

### 2.12.1.2 A mechanism for attending to surprising sub-parts of an object

Another use for the new conditional measure of saliency for sub-parts of an object is in drawing attention specifically to parts of an object that are *unusual*, given its type. (Assuming we have already confirmed that it has enough ‘usual parts’ to *qualify* as an instance of its type.)

Again, to do this, we can treat the whole region occupied by an object with a hypothesised type (e.g. a car) as a saliency map. Now assume that the salience of a given sub-region is measured not by strength of expectation of the features in this region, but by discrepancy between the expected features in this region and the features that are actually found. In this saliency map, the most salient (i.e. surprising) regions will be those that contain visual features that are unusual *for a car*. For instance, if we were expecting a black wheel in a given sub-portion of the region, the presence of a green wheel should register as surprising: our attention would be drawn to the wheel, as the most salient region within this new saliency map. To begin with, of course, we would classify the object in the selected salient region (as a wheel). But there is also a very natural property-level IOR operation that can be executed, that subtracts the expected features of the wheel from the actual features. Note this is the same IOR operation as was outlined in Section 2.10.3: however, the expectations now come not from the category derived from the object classifier, but from the *expectations* as to the object at the current salient location, given the position of this location in relation to the larger area identified as a car. That is, we subtract the expected ‘black wheel’ features from the encountered ‘green wheel’ features, and we are left with ‘green’.

Note that this salience mechanism can also draw attention to expected object parts that are *entirely absent*. For instance, if we are expecting a wheel in a given location, the complete absence of a wheel will also register as surprising—presumably as *more* surprising than a wheel with unusual properties. In this case, when we initially pass the selected location to the object classifier, we will likely get a null result (that is, a distribution over object categories with high entropy). In this situation, I suggest there is a mechanism that registers a classification failure and then attempts to classify the set of expected features *by themselves*. If this is successful, we have identified a negative polarity fact: there is *no wheel* at the expected location.

I will later argue that this saliency map provides the basis for an account of the semantics of concrete predicative sentences featuring the verb *have*, such as *The car has a red bumper*, or *This car has no wheels*. This is discussed in Section 12.2 (and also draws on ideas about attention to properties developed in Section 12.3.1).

### 2.12.2 A mechanism for attending to the local and global form of objects

Now assume the scenario discussed in Section 2.11.4 and illustrated in Figure 2.21: a homogeneous region of Xs that forms the shape of an ‘A’ is allocated high saliency, partly because of its local form (‘A’) and partly because of its high textural homogeneity. Again, we need a mechanism for *deciding* whether the classifier should represent the stimulus’ global form (‘A’) or its local form

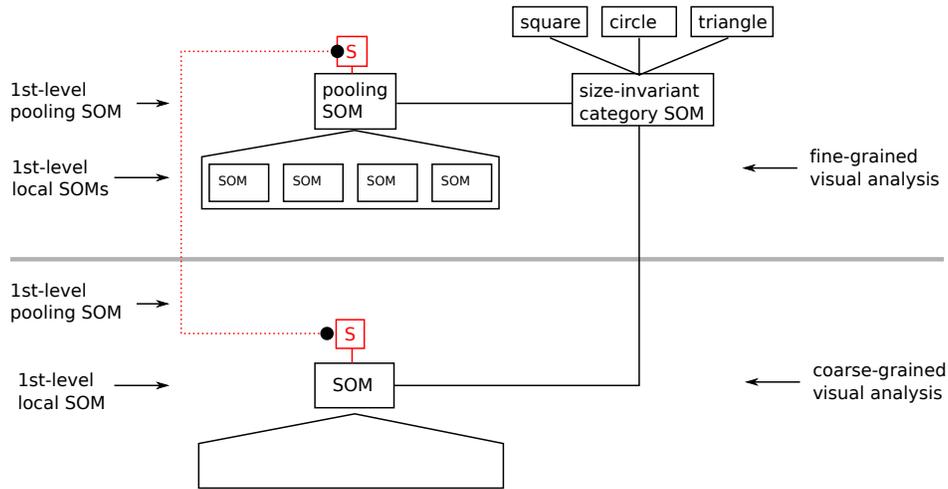


Figure 2.25: A circuit for choosing between the local and global form of a selected salient region. When the region is competing for selection, the salience of the upper pooling SOM contributes to that of the local SOM covering the same region at a coarser spatial frequency (refer back to Figure 2.20). Once the region is selected, however, these two saliences *compete*, so that only the representation in the winning SOM is passed to the size-invariant category SOM. (Inhibition is denoted by the black dots on the ends of the connection between the SOM saliences.)

(‘X’). (Note that if we classify its local form, we are using the ‘cardinality blindness’ property of the classifier that was discussed in Section 2.7.6.)

It is important a decision is made between these alternatives, because there is an important semantic distinction between them. If we classify the stimulus’ global form, we are identifying a *single* object. But if we classify its local form, we are identifying a homogeneous *group* of objects. (The stimulus represents a single A, but a group of Xs.) In Walles *et al.* (2014) we proposed that attention to the local versus global form of a stimulus is a perceptual analogue of selecting the syntactic NUMBER feature for a noun phrase: SINGULAR or PLURAL. I will talk more about this perceptual conception of the singular-plural distinction in Chapter 22.

For the moment, the question is how the observer who has selected a stimulus keeps separate the options of classifying its global form and its local form, and how the observer decides which to do. I propose a mechanism similar to that used to distinguish between attention to wholes and parts of objects: having selected a salient region, there is competition *between* the spatial scales associated with the global form of the region (i.e. the ‘default frequencies’ discussed in Section 2.11.1) and the spatial scales of the texture elements in the region (i.e. the scales of elements that are *pooled* across that region, as discussed in Section 2.11.4). This mechanism is illustrated in Figure 2.25.

Note that the process of attending to the local form of a stimulus is quite different from the process of attending to the parts of a stimulus. For one thing, attention to local form retains the whole of the currently selected salient region, while attention to parts selects sub-regions within this region. For another thing, as already mentioned in Section 2.11.5, the spatial scales of ‘local form’ elements in relation to the selected region’s default frequency are different from that of ‘parts’. Parts can be larger than texture elements.<sup>18</sup>

I assume there is a mechanism that records whether the winning scale is that associated with the region’s global form or local form. This will indicate whether the category that is read out is a single object or a homogeneous group of objects. This mechanism has to operate across all retinal

<sup>18</sup>There may be some overlap between the relevant spatial scales: for instance, a pair of adjacent Xs might stand out as salient both because they can be analysed as adjacent ‘parts’ of a larger whole, but also because they can be analysed as a homogeneous group. My main point is that the relative spatial scales do not *fully* overlap.

locations and spatial frequencies, and deliver a single bit of information, from which the labels ‘singular’ and ‘plural’ can be read. I’m still thinking about how this mechanism might work.

Whether the local or global form of a stimulus is selected depends very much on the stimulus. If a selected region has high textural homogeneity and little global form, local form will likely be selected; if it has a clear global form and little textural homogeneity, global form will likely be selected. There are some stimuli, like the A made of Xs in Figure 2.21, where both local and global form compete strongly. In these cases, we suggest there is a general tendency for the local form to win first (as Navon, 1977 found in his classic study). However, I also envisage an IOR operation, which allows attention to be focussed on the local form *after* the global form has been identified (and before attention passes to a new location). This allows the stimulus to be reparsed for its local form. I suggest this ‘IOR-of-classification-scales’ operation is one of the denotations of the particle *of*, as it features in expressions like ‘a line of soldiers’, or ‘a basket of apples’. This is discussed more in Chapter 22. Likewise, if local form is selected first, I suggest there is an operation that can reparse for global form. I suggest this surfaces in language in prepositional phrases that report ‘configuration’ of objects—for instance *the soldiers were in a line*, or *the apples were in a basket*, or (potentially) *the paper was curled up ‘in’ a ball*). This operation will be discussed in more detail in Chapter 12 (I think).

## Chapter 3

# Spatial representations: environments and places

There's another circuit involving PFC and hippocampus that specialises in representing spatial environments and locations. It can represent the location of the agent, or of some observed object, or a goal location. I like the idea that the circuit is isomorphic to the circuit for LTM/WM of episodes, and that representations in these parallel circuits can communicate with each other.

### 3.1 A circuit for representations of places

The places SOM: the basic idea goes here. (MSOMs represent commonly occurring sequences of inputs; in this case, the SOM receives locomotion commands, and so learns trajectories; an environment can be represented by the trajectories that are possible.) Refer to a paper for the details.

The architecture for the places SOM is shown in Figure 3.1.

#### 3.1.1 An actor-critic system for learning the next locomotion action

In here, you should mention that the rewards in this case are *internally generated* within the agent: they are tailored to learning the spatial structure of the environment. There's a reward associated with travelling straight, and a small punishment associated with turning: this scheme encourages locomotion actions that discover the full extent of the environment. There's also a (larger?) punishment associated with encountering an obstacle to forwards navigation: bumping into a boundary, or other obstacle in the environment will incur this punishment.

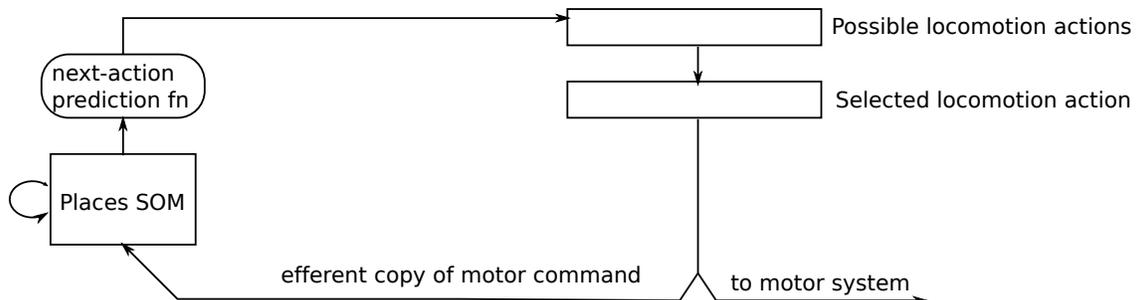


Figure 3.1: Architecture of the places SOM

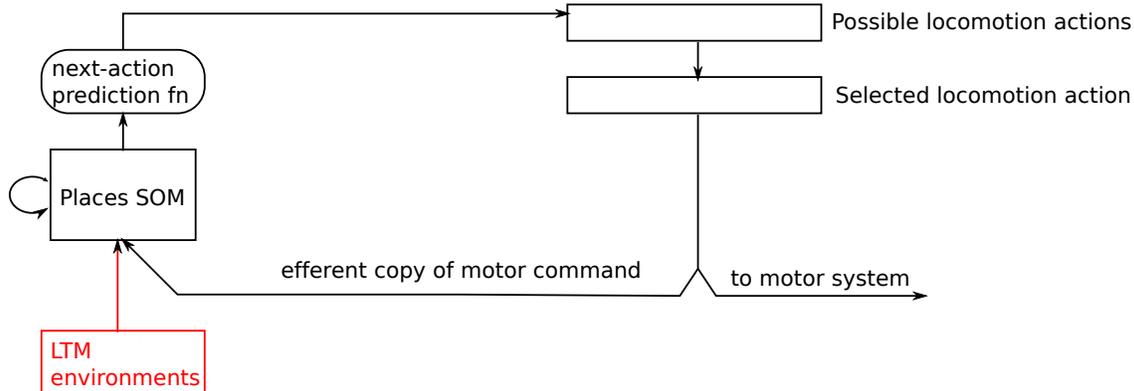


Figure 3.2: LTM environments as a tonic bias on the places SOM

## 3.2 A circuit for representations of environments

This is where you introduce **LTM environments**, and explain the circuit that learns them. The circuit is shown in Figure 3.2: the LTM environments medium that exerts a tonic bias on the dynamics of the places SOM is shown in red.

LTM environments provide static input to the places SOM, to bias its dynamics: they reflect the shape of the environment. Refer to a paper for the details. I will later suggest that there is a type system for environments, analogous to the type system for objects; see Section 3.7 for more on this idea.

### 3.2.1 Learning (and identifying) LTM environments

Marco's stuff here.

These similarities mean that sequences of navigational actions can be *simulated* in our model, just as sequences of episodes can be simulated in the situations SOM. There's plenty of evidence that animals can simulate trajectories offline, as I will discuss in Section 3.5.1.

## 3.3 A circuit for representations of navigational plans

Here, I introduce two new media: one holding **navigational goals** and one holding **goal trajectories**. As I will discuss in Section 3.5.5, navigational goals model representations in medial PFC, while goal trajectories model a particular type of hippocampal place cell (in CA1); for now I will just introduce the computational architecture.

In the following sections, I'll be talking about a particular style of navigational learning, that makes use of allocentric representations of place in the hippocampus. I'm not talking about navigational methods that involve visual cues (sometimes called 'piloting') though these are also very important, at least for sighted agents.

### 3.3.1 Learning of trajectories to places associated with external reward

In this section, I will discuss how a navigational goal enables the agent to learn trajectories to locations associated with reward. I will begin by considering a simple reward schedule where the place associated with reward is directly and reliably indicated by an externally provided cue, that the agent perceives somewhere distant from the goal. For instance, let's say cue  $C_1$  indicates that a reward will be found at place  $P_1$ , while cue  $C_2$  indicates that a reward will be found at place  $P_2$ . If  $C_1$  is activated, we want the goal trajectories medium to represent a trajectory from the agent's current place to  $P_1$ , and if  $C_2$  is activated, we want it to represent a trajectory to  $P_2$ .

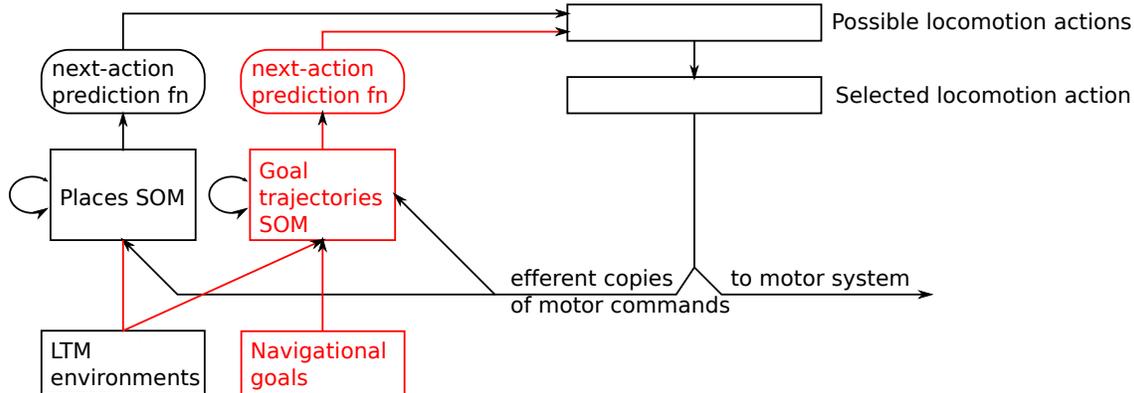


Figure 3.3: Architecture of the model of trajectory learning

There are several key ideas in the learning algorithm. One group of ideas derive from a model of plan learning by Braver and Cohen (2000). Another group of ideas involve the introduction of a variant on the places SOM that is sensitive to navigation goals. I will introduce these ideas in turn.

### 3.3.1.1 The navigational goals medium

The first key idea in the model of trajectory learning is that a navigational goal unit in the PFC is activated *as a function* of the perceptual cue that indicates the reward schedule. To be concrete, imagine a 1:1 mapping between perceptual stimuli and units in the navigation goals medium, so when the agent perceives  $C_1$ , an associated navigational goal  $X_1$  is activated. (I will talk more about how this function from perception to goals is learned in Section 4.2, in particular in Section 4.2.3.) The second key idea is that after a navigational goal unit is activated in PFC, *it stays active*, even after the perceptual cue that triggered it is no longer present. In this sense,  $X_1$  represents the special task-state in which the agent has to get to place  $P_1$ . This is a classical property of PFC units representing plans: once turned on, they are somewhat resistant to perceptual disruption (for discussion and evidence see e.g. Miller and Cohen, 2001; Freedman *et al.*, 2003). (This idea implies that perceptual stimuli are only mapped to PFC plans at particular moments; I will talk more about the moments when such updates happen in Section 13.3.) The third key idea is that at the start of learning, the PFC unit that represents the navigational goal is turned on *speculatively*. When the agent first activates the navigational goal unit  $X_1$  in response to cue  $C_1$ , he still hasn't learned anything at all about its *content*—that is, about how it biases navigation. All he knows is that a plan is required for the navigation reward schedule signalled by  $C_1$ : and  $X_1$  is the assembly that will eventually hold it. This distinction between the mechanism that *selects* a PFC assembly that will hold a prefrontal plan and the mechanism that learns how the plan should bias behaviour is a fundamental part of Braver and Cohen's model. I now turn to the latter mechanism.

### 3.3.1.2 The goal trajectories SOM

I envisage that the medium holding navigation goals provides input to a hippocampal circuit representing locations and trajectories, separate from the places SOM, as shown in Figure 3.3. The new parts of the model are shown in red.

The goal trajectories SOM is like the places SOM, in that it's updated by locomotion actions and its own recurrent inputs and a tonically active representation of the current spatial environment. (In this case, call the current environment  $E_1$ .) But it also takes one more input, from the navigation goals medium. While the navigational goal unit  $X_1$  is on, it exerts a constant bias on the dynamics of the goal trajectories SOM.

Like the places SOM, the goal trajectories SOM works in conjunction with a next-action prediction function. The prediction function is trained using the same actor-critic framework as the one for the places SOM; see Section 3.1.1. However the places SOM prediction function is trained on internally generated rewards, tailored to learning the spatial structure of the environment. The prediction function paired with the goal trajectories SOM is trained on these internally generated rewards as well—but it is also trained on actual, externally generated rewards. (At least, it is trained on external rewards in its simplest mode of operation. A more complex internally-generated reward scheme for training it will be discussed in Section 3.4.5 below.)

To explain how the system learns the ‘content’ of a navigation goal, I will step through an example of training. Assume navigational goal unit  $X_1$  has just become active. Now the agent goes wandering off through the environment, driven by the navigational affordances represented in the regular places SOM, and perhaps also by some component of exploration. As he walks, the places SOM updates as normal. The goal trajectories SOM also updates as a function of the agent’s sequence of actions. But crucially, the states it gets into are also *specific to the active navigational goal*,  $X_1$ . Remember, the system as a whole doesn’t yet know anything about what this goal is! All the unit  $X_1$  encodes is an assumption about the reward schedule currently in place: namely that the place currently associated with reward is some function of the cue stimulus  $C_1$ .

Now consider what happens if the agent gets lucky, and arrives at place  $P_1$ , and gets a reward. Now the critic function will learn to associate the current state of the goal trajectories SOM with a big reward (since none was predicted)—and the next-action-prediction function can learn to map this state onto the action that led to the reward. Of course the agent has not finished learning yet: the temporal difference learning scheme requires him to restart the learning routine many times. After each run it becomes easier for the agent to reach a place associated with a reward, as (discounted) rewards become associated with places progressively further from the actual goal place. Remember that reward is associated with states of the trajectory planning SOM, rather than with actual places. In all the relevant runs, the states in this SOM are updated under the influence of the same navigation goal unit  $X_1$ .

In summary, during temporal-difference learning under the influence of  $X_1$ , the critic function learns to associate a gradient of reward values to states in the goal trajectories SOM that represent places leading up to the goal state. There are a few interesting effects here. Firstly, the agent can start from different locations. Since the SOM encodes the agent’s current locations (being a type of place-cells SOM), and its dynamics encode the agent’s movement sequences, it learns *distinct trajectories* for different starting places. But there is a measure of generalisation over these trajectories: since the active navigation goal unit is just one input to the SOM, and with its other inputs it is simply representing places, if it is in a place it has never been, but the next-action function learned to map a \*neighbouring place onto action  $A$ , it is likely to pick  $A$ —which is not a bad decision.

The result of all this learning is that, provided the agent starts at a wide enough range of different places during training, the dynamics of the goal trajectories SOM, and the trained next-action function, will take him from *any* initial place to the specific goal place  $P_1$ . Say after training, the agent is placed at  $P_0$ . In this state,<sup>1</sup> the next-action function has learned to do the action that climbs the gradient of discounted rewards: say  $A_1$ . So the agent does  $A_1$ . Now the trajectories SOM updates (under the influence of  $X_1$ ). In its new state, the next-action function again knows what to do. And so on.

I conclude this section by making a few interesting observations to make about this training regime.

**Separable, declarative representations of many trajectories** As mentioned above, the temporal difference learning function, when trained on  $X_1$ , will learn learn a distinct trajectory *from several possible starting places* to the goal place  $P_1$ . If it is trained often enough, and has

---

<sup>1</sup>I think the state of the trajectory planning SOM at the start of an action might have to be copied directly from the places SOM—unless this effect happens automatically somehow.

a large enough capacity, it could learn trajectories from *all* places in the environment to the goal place. And if it was trained on several different cues, that predicted reward at several different locations, it could learn separate trajectories from many possible starting places to many goal places. The learning mechanism is set up to represent a large space of distinct trajectories.

Note that the system’s representations of trajectories are *implicit*. As Cisek (2005) notes, the brain does not ‘precompute’ the points in a planned trajectory: these points are activated within a dynamical system as the action takes place. (Although in our model, the execution of a trajectory can always be *simulated*, as discussed in Section ??.)

**A navigational goal represents task as well as place** Note that a unit in the navigational goals medium trained using the scheme described above can be thought of as representing a location—but it can equally well be thought of as representing a *task*. Recall that the unit in our example is associated with a perceptual cue,  $C_1$ : it becomes active after  $C_1$  is presented, and thereafter exerts a tonic influence on the goal trajectories SOM. It would be useful to decouple representations of goal locations from representations of plan-triggering cues—because potentially, there could be several different cues that trigger establishment of a given goal location, and it would be inefficient to learn a trajectory for each of these. In Section 3.4.5 I describe a scheme that learns a purer concept of goal locations.

### 3.3.2 Working with multiple navigation goals: a mechanism for trajectory selection

In the previous section, I assumed there was just a single unit active in the navigational goals medium. However, it is interesting to consider what happens if there are several such units, associated with several different goal locations, and each biasing the goal trajectories SOM towards a particular trajectory. Under these circumstances, we would like the goal trajectories SOM to function as a mechanism for *selecting* a goal trajectory, and subsequently for *executing* a sequence of actions that drive the agent along the selected trajectory.<sup>2</sup>

There are a few desiderata for the mechanism that selects and executes trajectories. Firstly, we would like the mechanism to choose between discrete alternatives, rather than to blend trajectories. (Blending trajectories would likely drive the agent towards a point between the evoked goals.<sup>3</sup>) Secondly, we would like the decision to take into account two separate factors: firstly, the length of the trajectory to the goal (with closer goals being preferred), and secondly, the size of the reward at its endpoint (with larger rewards being preferred). Thirdly, we would like to be able to impose a *prior* bias on goal trajectories, by activating them at different levels. (The decision about the best trajectory should therefore weigh prior bias for trajectories, as well as path length and size of reward.)

In fact, I think the goal trajectories medium fulfils these desiderata quite well. I will explain by giving another worked example.

Say there are two goal locations, cued by different cues,  $C_1$  and  $C_2$ —and therefore represented by different navigation goal units ( $X_1$  and  $X_2$ ). Under the influence of  $X_1$  and  $X_2$ , the SOM will represent two overlaid states. Each component of the state represents the agent’s current location, naturally: to that extent, the states overlap. But the states *also* represent the starting points of two very different learned trajectories.

Now consider what the next-action function will predict. It has been trained to map the SOM pattern onto the action that generates the largest (discounted) reward. There are two overlaid SOM patterns. If these do not interfere with one another, then the distribution of actions predicted by the function should approximate the summed distributions it would predict from the two individual patterns. There will be a winner in each separate distribution: in many cases, the

<sup>2</sup>Cisek, 2005 argues persuasively that the same neural circuit is responsible for these two tasks. He was discussing hand/arm actions rather than navigational actions. In fact we will argue that the trajectory learning mechanisms described here are also used in the system that learns hand/arm actions; details are in Section 4.

<sup>3</sup>Again there are huge parallels between this system and the system for selecting motor actions: see e.g. Tipper *et al.* (1992). No coincidence, I suggest.

winner of the summed distribution will be the action most strongly activated in the distributions considered separately: that is, the action that generates the largest discounted reward. Note this might be because the associated goal is close, or because it's far away, but correspondingly larger. If the above reasoning is correct, the goal trajectories SOM and next-action network, trained using temporal difference learning, should make a pretty good job of trading off size of reward and distance to reward, as discussed in the desiderata set out above.<sup>4</sup>

Now consider the other factor we wanted to take into account: a prior bias on navigation goals, expressed in the strength of navigation goal unit activations. Note that if our two navigation goal units are activated at different levels, we can expect this to have an effect on the strength with which the SOM patterns associated with the two different trajectories will be expressed. And since the next-action function needs to *recognise* an input in the SOM in order to generate the appropriate learned output, we can expect that the more strongly a navigation goal is activated, the easier it is to recognise a training input, and the more strongly its response to this input will be expressed. So there is some reason to think the SOM's decisions can be influenced by the activation levels of goal trajectory units.

Now consider the other desideratum: having selected a trajectory, the SOM must execute it, without being distracted by alternative competitors. For instance, say the agent has selected a trajectory to place  $P_1$ , but this happens to pass through, or close to, a goal place  $P_2$ , which was also competing for selection. Assume the agent does the action which sets him off on a trajectory towards  $P_1$ , which is the best action according to the critic. I think there is a natural mechanism for reinforcing his decision, because the new state of the SOM is one which *reflects* this decision—and moves away from the states in which trajectories towards the other goal were learned. Again, this mechanism is probably not infallible, and it probably makes sense to limit the number of competing alternative trajectories.

Another mechanism that should help is one whereby the navigation goal units *not* selected are inhibited. This requires a mechanism that recognises which goal unit *was* selected: not a trivial thing if the decision process is made in the SOM and next-state function, rather than simply by picking the most active goal unit. However, we should be able to implement the kind of mechanism that recognises 'the current environment' by treating the current SOM pattern as a query and retrieving the environment that best matches it (see Section 3.2.1). If the set of active navigation goal units is progressively adjusted by this mechanism, we should quite quickly lock into a single selected goal unit.

The above model of trajectory planning will be progressively refined, in two stages. Firstly, in Section 3.4.5 I will introduce an idea about how to learn a purer concept of 'goal places', decoupled from cues that signal reward schedules. Secondly, in Section 12.1.1 I will introduce a model of how an agent's memory for the locations of objects, together with learned associations between objects and rewards, can induce a distribution over goal locations, biasing the agent towards trajectories to places where rewarding objects are likely to be, and away from places where punishing objects are likely to be. However, I will conclude the current section with some thoughts about how the trajectory selection mechanism just described may function to *optimise* trajectories.

### 3.3.2.1 A mechanism for trajectory optimisation?

Note the above selection mechanism means that if there are two alternative trajectories that reach the same reward, the shorter of those will systematically be preferred. This is starting to feel a lot like a mechanism for *optimising* trajectories, which is quite cool.

---

<sup>4</sup>There are a few paradoxical cases where we can expect problematic interference between two navigational goal units. These centre on the output of the next-action network. This network essentially sums two separate distributions over possible actions. It is quite possible to imagine that the winner of the summed distribution is not the winner in either distribution individually, but (for example) the runner-up in both. This is probably a reason to suggest that the number of alternative candidate navigation goals active in parallel should be reasonably small. (Some form of softmax should probably be envisaged, to limit how many there can be.) At the same time, note that the SOM also provides a way for the agent to *serially* select candidate navigation goal units one at a time, and play forward a trajectory, to see what happens. There is very good evidence this actually happens in the hippocampal place system, as I will discuss in Section 3.5.2.

Note that in the above case, the longer (sub-optimal) trajectory will probably remain represented, even if it is not preferred. This is very useful, in case the shorter, optimal trajectory is unavailable for some reason.)

## 3.4 A circuit for representing the location of arbitrary individuals

Place cells just represent the agent’s own location. It’s important to be able to represent the location of external individuals in the environment too. This is an essential preliminary to the model of LTM for object locations, which is presented in the next chapter.

### 3.4.1 View cells

In monkeys and humans, the hippocampus also contains ‘view cells’, that respond to a given location in the local environment not when the agent is *at* that location, but when the agent is *attending to* that location: see Rolls *et al.* (2005) for monkeys and Ekstrom *et al.* (2003) for humans. Rolls *et al.* found view cells in the perirhinal cortex as well as the hippocampus.

### 3.4.2 Orientation cells

Here, introduce head direction cells.

(Actually in our model, units in the places SOM already encode orientation in an allocentric frame of reference. But orientation and place end up being encoded separately too.)

### 3.4.3 A circuit for learning view cells: the orienting SOM

In this section, I will discuss the function that computes an environment-centred representation of the location of an attended external object. I will call this function the **orienting** function. The representations of viewed places that it produces occupy a medium called the **viewed-places** medium.

The orienting function needs several inputs. One is the map of place cells that represent the agent’s own environment-centred location (that is, the places SOM). It also needs an allocentric representation of agent orientation, of the kind just described in Section 3.4.2. Then it needs information about the retinal location of the external object. I will assume this is provided in the form of a set of retinotopic feature maps, of the kind computed in early visual cortex. The function also needs information about the distance of the attended object. I assume this is contained implicitly within the retinotopic feature maps, since these include features representing different degrees of retinal disparity across the whole retina. Finally, it needs information about the angle of the agent’s eye in relation to his head, and about the angle of the agent’s head in relation to his body. I assume each of these angles is represented in a coarse-coded scheme, as a ‘bump’ of activity in a 1-dimensional array of units.

My proposal is that the orienting function simply takes the form of a SOM, that receives input from all these media, and also from the viewed-places medium, as shown in Figure 3.4. I’ll refer to it as the **orienting SOM** from now on. I will first describe how the orienting SOM is trained. The complete training regime for orientation also involves training within the viewed-places medium, which I will describe afterwards.

#### 3.4.3.1 Training the orienting SOM: stationary external object, moving observer

The first training regime involves adopting a special mode, where the agent is moving, and the viewed external object is stationary.<sup>5</sup> (A scheme like this is also used in a model by Wiskott and

---

<sup>5</sup>I’m not sure how to guarantee the external object is stationary. Maybe it’s simply the case that most external objects are stationary.

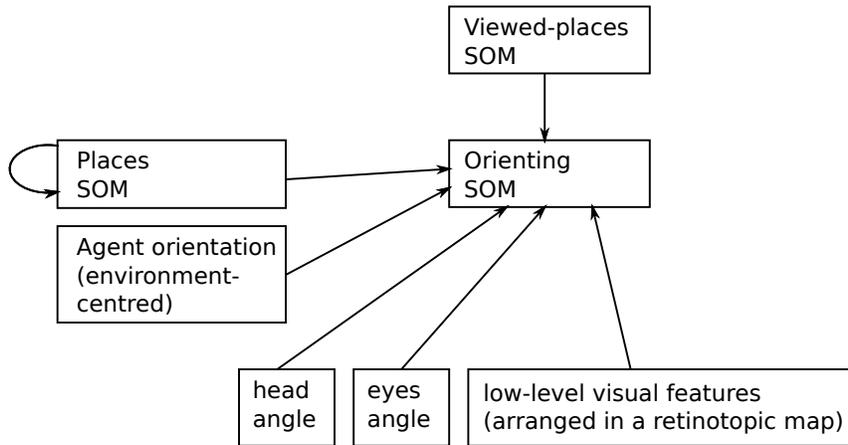


Figure 3.4: The orienting SOM. This SOM maps object locations in the agent’s visual field onto environment-centred places in the viewed-places medium.

Sejnowski, 2002.) In this mode, I have to assume the existence of a mechanism for *tracking* the projection of a selected external object on the retina.<sup>6</sup> The tracking mechanism I have in mind is similar to that described in Section 2.6, except that in that section, the tracking mechanism was body-centred and here it is environment-centred. I assume a ‘tracking map’, in a retina-centred coordinate system, that maintains a ‘bump’ of activity at the point on the retina projected by the object. During tracking, this bump on the tracking map must move, as a function of local movement signals at the selected location: to the right if movement is to the right, and so on.

Learning proceeds as follows. First, an external object is selected (at some arbitrary place on the retina), and tracking is initiated for this location in the tracking map. Next, a winning unit is computed in the orienting SOM, without input from the viewed-places medium, and this unit is used to reconstruct a unit in the viewed-places medium. Crucially, once a unit is selected in the viewed-places medium, it is *held active* throughout learning. During learning, the agent moves around at random, maintaining the object in his field of view, and tracking it in the tracking map. (He can choose to foveate the tracked object if desired, but he must not always do so, because the orienting SOM must train on every point on the retina.)

During training, the orienting SOM as usual learns frequently-occurring associations between all its inputs: retinal positions, environment-centred representations of the agent’s place and orientation, and the agent’s head and eye angles. Crucially, the inputs always include the same unit in the viewed-places medium. This enforces the axiom that the viewed object is stationary.

The orienting SOM trained like this has some interesting properties, which I will briefly mention before describing the other aspects of training.

**The orienting SOM operates in parallel over the retina** Firstly, the orienting SOM learns to map retinotopic locations of objects to environment-centred locations *in parallel*. This is useful; and what is more, there’s good evidence that the brain performs coordinate transformations of this kind in parallel (see e.g. Pouget and Sejnowski, 1997). We will revisit this parallel operation in Section 4.2.3.2.

**The orienting SOM is bidirectional** Secondly, the orienting SOM can work in both directions: after learning, it can be configured either to map retinal locations onto environment-centred places, or environment-centred places onto retinal locations. The former mode is bottom-up, and functions to register visual information in a more stable format, in particular in LTM. We will

<sup>6</sup>For a discussion of how the visual system learns what counts as ‘projections of external objects’ on the retina, see the end of this section.

discuss LTM representations of object location in Section 12.1. The latter mode is top-down, and functions to bias attention towards certain places on the retina, in virtue of objects which are remembered or expected to be at particular environmental locations. We will discuss this kind of top-down scenario in Section 12.1.1.

**The vertical component of retinal place representations** To a first approximation, the environments the agent moves around in are horizontal. However, the retina has a vertical component to it as well as a horizontal one. Something meaningful is often stored in this component: if the agent’s line of sight is horizontal, and he is standing on an uncluttered flat plane, places further away from him project to higher points on his retina. However, this generalisation does not hold if there are objects in the agent’s environment. Objects resting on the ground can have arbitrary heights, and other objects can rest on top of these high objects; some objects can even fly through the air. Objects only feature in the navigation system as boundaries or obstacles. However, they can feature in their own right in other motor systems—in particular in the system that controls reaches to target objects in the agent’s perispace, which is described in Section 4.2.3. This system has its own representation of space, and places in this space are also transformed in parallel into places on the retina—so retina-centred locations also communicate with the reach motor system.

**Bootstrapping the retinal representation of objects** The tracking mechanism, as introduced above, tracks ‘the projection of a selected external object on the retina’. For this to work, the visual system has to be able to compute representations of ‘object-like things’, in parallel, at every retinal location. If it happens in parallel, it must happen at early visual processing stages, in the set of retinotopically-organised feature maps. As just mentioned, the retina also participates in the ‘reach’ motor system. I suggest that it is through its participation in the reach motor system that early visual mechanisms across the retina learn the visual signatures of objects that are used by the tracking mechanism. I discuss this in more detail in Section 4.2.3.2.

### 3.4.3.2 Training the viewed-places medium: moving external object

In another training mode for the orienting system, the external object is moving (see again Wiskott and Sejnowski, 2002). There are several cues that can signal a moving object. (They all tend to be most informative when the agent is stationary, so I will assume a stationary agent in this mode.) One cue is when the agent’s head and eyes are both fixed, and the active location in the tracking map is moving over time. Another is when the active point in the tracking map is stationary over time, but the agent’s head and/or eyes are moving. (The eye movements must be of the ‘smooth’ variety, not saccades.)

Learning in this mode involves the axiom that the external object’s movements must be between *adjacent locations* in the environment. To enforce this assumption, the viewed-places medium is trained to have some dynamics of its own. To do this, we will model it as a recurrent SOM—an mSOM, as usual. From now on, I will call it the **viewed-places SOM**. This SOM takes as its input a representation of its previous state. Assume the agent has already learned to activate units in the viewed-places medium (now the viewed-places SOM!) through the first learning method—the one that assumes a moving agent and stationary object. Learning with the second method, with a *moving object*, causes the SOM to learn *sequential patterns* linking these SOM units together. These define a measure of spatial adjacency within the viewed-places SOM.

### 3.4.3.3 Method 2 continued: learning a measure of orientation in the viewed-places SOM

Note that a viewed object can move in several different directions. So the dynamics learned in the SOM is not deterministic: a SOM state with a particular unit active can update to several alternative states, each featuring its own active SOM unit. These define the two-dimensional ‘locality’ of the original unit, in some sense. It would be useful to be able to learn a measure of

orientation, or direction, in the SOM, so that it can distinguish between these alternative states. Objects in the world that can move often have a natural ‘forward’ direction, and their geometrical shape often indicates this direction: for instance, animals (including humans) tend to be roughly symmetrical in a vertical plane pointing in their forward direction (the ‘median plane’) but not in the vertical plane perpendicular to this (the ‘coronal plane’). As a consequence, the projection of movement-capable objects on the retina typically carry information about their orientation. A final assumption about our viewed-places SOM is that it receives input from vision.

When provided with visual information, the SOM should learn to associate particular orientations with particular dynamic trajectories. Note that this is a big ask: navigating objects come in many different shapes. However, there is a visual pathway that learns to represent the shape of arbitrary objects elsewhere, namely the ‘object classification’ pathway in inferior temporal cortex. (This pathway was briefly introduced in Section 9.1.1, and will be discussed in more detail in Section ??.) I assume that the visual inputs provided to the viewed-places SOM come from high-level representations of shape computed in this system.

The extra circuitry for training the viewed-places SOM’s dynamics, including inputs from a network computing viewpoint-specific representations of object shape, is shown (in red) in Figure 3.5.

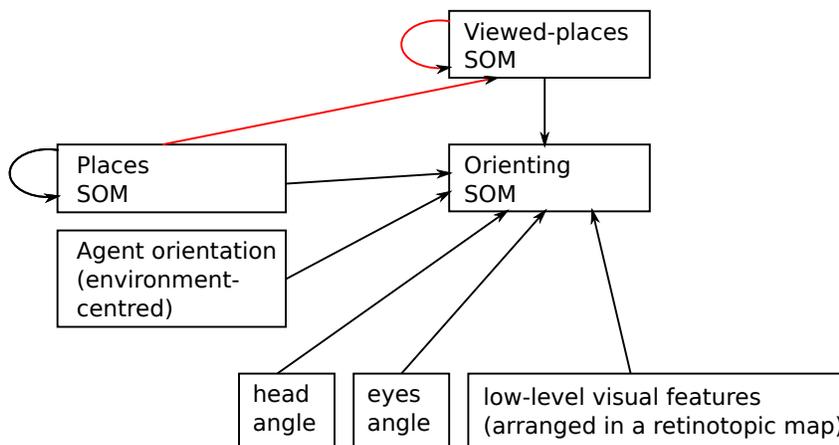


Figure 3.5: Circuitry for training the viewed-places SOM’s dynamics (shown in red). The SOM’s dynamics is trained when the observer is stationary, and tracking a moving external object.

Note that the system that combines representations of the shape of a viewed object with its trajectory looks suspiciously like the superior temporal sulcus: see e.g. Perrett *et al.* (1989 and much other work).

I’m not sure if this all works, but I like the look of it!

### 3.4.4 Configuring the viewed-places SOM to generalise over ‘self’ and ‘other’

Here introduce the idea that the viewed-places SOM can be configured to represent the agent’s own location, or that of an external individual. To allow it to represent his own location, the agent must be able to activate the viewed-places SOM via two completely separate circuits. When he is attending to an external object, he should activate the SOM using the circuitry described above in Section 3.4.3. When he is attending *to himself*, he should activate the SOM using a circuit that maps units in the places SOM (that always represents his own location) onto units in the viewed-places SOM. Of course, this mapping also has to be learned. The key axiom here is that the mapping should preserve the separately-learned dynamics of the two SOMs. I think we can assume the viewed-places SOM can take one further input, that comes from the places SOM. This



this.

A key idea is to posit a 1:1 mapping between units in the second-order places SOM and units in the navigational goals SOM.<sup>7</sup> I assume this happens in a mode where the agent himself is navigating—thus, the second-order places SOM represents his own location.

#### 3.4.5.1 A training regime for learning goal place units

I envisage a special training mode that can be activated while the agent is navigating an environment. In this mode, an arbitrary unit is picked in the second-order places SOM, and its associated navigational goal unit is activated. As described in Section 3.3.1, once such a unit is activated, its activation persists into the future, so it provides a tonic bias on the agent’s goal trajectories SOM. I will call navigational goals associated with places in this way **goal places**: I’ll assume they are a special variety of navigational goal, somewhat more general and symbolic than the units activated by arbitrary cues described in Section 3.3.1. In this special mode, a reward is ‘internally’ delivered when the agent’s current place (as represented in the regular places SOM) is the one associated with the active navigational goal unit. With this special reward schedule in place, temporal-difference learning proceeds as normal, as described in Section 3.3.1. If the capacity of the network is large enough, it should learn a very large set of well-delineated general-purpose trajectories: one from every possible current place to every possible goal place. (In every possible environment.)

#### 3.4.5.2 Selecting a goal place from a set of candidate goal places

Recall from Section 3.3.2 that a set of several alternative navigation goal units can be activated simultaneously, to different degrees: in this situation, the goal trajectories SOM functions to select, and then execute, a trajectory associated with one of these units, based on the activation levels of the goal units, the length of their associated trajectories, and the size of the reward likely to be obtained. This kind of competition can also be envisaged between goal place units. (Importantly, in this case, if we assume that the offline training of goal place trajectories uses a constant level of internally generated reward, the discounted reward associated with a SOM state only reflects the length of the associated trajectory.) This competitive environment provides a framework whereby trajectories can be selected based on the objects that are likely to be *at* locations in the environment. This type of competition will be discussed in Section 12.1.1, after we have introduced a model of memory for object locations.

#### 3.4.5.3 A network for orienting to a given place in the environment

The circuit described in Section 3.4.5 allows the agent to choose between arbitrary goal places in his local environment. There is a related circuit that is also worth describing, that allows the agent simply to choose *direct his attention* to arbitrary places in his environment. I will call this circuit the **orienting network**. This network exploits the circuit that maps retinal locations of objects onto environment-centred locations in the 2nd-order places SOM, as described in Section 3.4.3. (...)

### 3.4.6 Vision-based (‘egocentric’) navigation strategies

All the navigation strategies I have introduced so far have used the allocentric representations of places learned in the places SOM. However, the orienting network just described allows another kind of strategy: if the observer can visually attend to the object at a goal place, he can simply align his body with his direction of attention and ‘walk forward’. This kind of strategy, called **piloting** in the animal literature, makes direct use of egocentric, visual information. At the same time, however, since the actions that are taken can be passed as input to the places SOM in

---

<sup>7</sup>These are exactly analogous to the 1:1 mapping between the media representing the ‘current motor state’ and the ‘goal motor state’ in Lee-Hand and Knott (2015).

the usual way, these strategies can easily be *represented* in allocentric formats, even if allocentric representations are not ‘in charge’ during piloting.

A vision-based navigation strategy won’t always work if there are obstacles, or if there is no line of sight to the goal object; in this case, we have to fall back on allocentric strategies. But it is frequently a valid option, all the same. And when it is a valid option, it can provide a very efficient way of *training* the allocentric navigation strategies, so that the allocentric trajectory learning system is biased towards finding straight-line trajectories if they are possible.

Note that alongside egocentric (vision-based) strategies for navigating *to* a goal object, there are also egocentric strategies for navigating ‘along’ objects, or ‘past’ objects, or ‘away from’ objects. (These latter ones are not vision-based, but there are other modalities, in particular audition, that can be used to locate objects in an egocentric frame of reference. For instance, to run away from something that is making a noise, the agent just has to orient in a direction opposite to that which the noise is coming from, and the move forward.)

I will say a lot more about trajectories taken ‘in relation to’ landmark objects in Section ??, when I am discussing the semantics of prepositional phrases.

## 3.5 Evidence for the media in the hippocampus and PFC

### 3.5.1 Places MSOM

Representations of ‘the agent’s current location’: CA3 place cells. (This is also a medium where we can express probability distributions over *possible* locations.)

In humans, there are single-cell results confirming place cells (Ekstrom *et al.*, 2003) and MVPA results confirming place within an environment is represented in the hippocampus (Hassabis *et al.*, 2009).

The place SOM units in our model encode trajectories, rather than places as such. In animals, place cells often encode trajectories too; see e.g. Ferbinteanu and Shapiro, 2003. This happens particularly ‘when specific trajectories are taken repeatedly in constrained environments’ (see e.g. Chersi and Burgess, 2015).

#### 3.5.1.1 Offline simulation of place cell sequences

Say something about Matthew Wilson’s experiments here. Our network can certainly reproduce phenomena like replayed place cell sequences.

### 3.5.2 Possible locomotion actions

Evidence that a rat can sequentially evaluate alternative locomotion commands is provided by Johnson and Redish (2007); see Redish (2016) for discussion.

### 3.5.3 The current locomotion action

Representation of ‘the current action’: ‘speed cells’ in medial entorhinal cortex (2015); cells indicating the degree by which the agent is turning left or right (Jacobs *et al.*, 2010).

### 3.5.4 LTM environments

Representation of ‘the current LTM environment’: parahippocampal cortex and medial entorhinal cortex (Preston and Eichenbaum, 2013).

Hassabis *et al.* (2009) studied human subjects navigating in a VR environment with multiple different rooms. MVPA could decode which room the subject was in (regardless of location within the room) from parahippocampal cortex.

### 3.5.5 Navigational goals

What properties should a navigational goal cell in PFC have? Note we don't expect it to have a 'place field' of the sort seen in hippocampal cells. It would be a misconception to think that a cell encoding a navigation goal will have a response only when the animal reaches the goal location. If the cell encodes a navigation goal, we expect it to be active *throughout the agent's journey* to the goal location. (And actually, when the agent reaches the goal, it should turn off.)

Here refer to Ito *et al.* (2015). They have evidence that medial PFC maintains tonic information about the current navigational plan, and relays it to hippocampus (specifically, to area CA1) via an area in the thalamus called the nucleus reuniens.

You should also probably say something about Schultz *et al.*'s (1997) findings about dopamine firing during operant learning here.

## 3.6 Consolidation within the place circuit

You should refer to Ito *et al.* (2015) here.

## 3.7 Properties and types of LTM environments

In this section I'll argue that LTM environments can have properties too, just like LTM individuals. I'll introduce some interesting analogies between LTM individuals and LTM environments.

### 3.7.1 Properties of spatial environments, and spatial environment types

There are also types of spatial environment. There's my office, and then there are offices in general; there's my street, and then there are streets in general. I expect that there's a medium representing the 'properties' of a 'currently attended' spatial environment, just like the RPC holds properties of the currently attended object. And I expect there's a second medium, representing frequently-occurring assemblies of these environment properties, which define a system of environment *types*.

#### 3.7.1.1 An idea about environment properties

What would the properties of an environment look like? I'm mainly interested in the spatial, or geometrical properties. I'm thinking these might simply be, *the collection of place SOM units associated with a given environment*. Environments of particular shapes might call on similar *sets* of place cells, and types could be learned on that basis.

On that model, there are two representations of an environment. One is the LTM environment unit, that's used as a constant bias on the dynamics of the place cells SOM. The other is the set of units in the place cells SOM that are active (sequentially) while this environment is being traversed. If the model is correct, environments of different types will involve different sets of place SOM units.

#### 3.7.1.2 The mechanism that learns associations between LTM environments and sets of place SOM units

In order to learn types of environments, there needs to be a mechanism by which activating an LTM environment can activate its full set of place SOM units *in parallel*. I propose that at every update of the places SOM, the newly active pattern is directly associated with the currently active LTM environment, in a special bank of associative connections linking LTM environments directly to place SOM units. After thoroughly learning an environment, these links will activate an interesting *compositional* representation of the spatial structure of the whole environment. It is compositional in that it represents a collection of all the spatial elements of the environment simultaneously. (The LTM environment, on the other hand, is a 'holistic' representation of the spatial structure of the environment.) I'll call this representation the **environment property**

**complex**, to echo the ‘rich property complex’ (RPC) associated with object types. I assume it has a lot in common with the RPC.<sup>8</sup> In Section 3.7.2 I’ll discuss how visual stimuli can be associated with environment property complexes.

### 3.7.1.3 Is the environment property complex engaged during actual navigation?

Maybe the environment property complex is not only *learned* during actual navigation through the environment, but *active in parallel* during actual navigation: that is, responsible for some component of the activity of place SOM units, so a given set of units are constantly preferred. I’m not sure about that: my current idea is that an LTM environment imposes a bias on place SOM units purely by influencing the *dynamics* of the places SOM. However, it’s an idea to keep in mind.

### 3.7.1.4 Learning environment types

I assume that environment types are learned in a layer analogous to the dominant property complex layer, that identifies correlations amongst place SOM units that commonly go together (that is, feature in the same environment property complex).

To learn such correlations, there has to be some point when all the place SOM units in an environment property complex are activated simultaneously. This could happen during navigation, as just discussed. But I propose that the main time it happens is when the agent *leaves* a given environment, and re-represents it as an object. At this point, I assume the learned connections that define environment property complexes are activated all at once, and we activate the property complex for this particular environment. I suggest this provides an opportunity to learn a visual representation of the spatial structure of the whole environment (see Section 3.7.2 below)—but also an opportunity to learn environment types.

### 3.7.1.5 Environment types, and their relationship with token environments

If the analogy with objects persists, then we expect to see a token environment of a given type having a set of ‘environment properties’ (i.e. place SOM units) characteristic of that type (and thus also possessed by several other environments)—but we also expect to see some ‘idiosyncratic’ environment properties, that are somewhat more unique to this token environment. What might those idiosyncratic environment properties look like? I have two ideas.

Firstly, we could imagine that the shape of this token environment differs in some idiosyncratic way from the shape represented by its type—and that these idiosyncracies are represented in a little collection of idiosyncratic place SOM units. These units would have to play nicely with the general units that make up the property complex associated with the type. There are various ways that could happen. For instance, the general units could define a coarse-grained spatial structure, and the idiosyncratic units could define

## 3.7.2 Visual classification of spatial environments

A spatial environment can be processed visually in two ways, which I’ll consider separately.

If the observer is in the environment to be classified, there could be a function like Chang-Joo’s, that takes the complete visual field as input, and generates a spatial representation as output. It differs crucially from Chang-Joo’s function, in that it is trained to return *the same environment at every point*, namely the currently active LTM environment.

If the observer can see the whole environment (that is, if it’s projected onto a retinal region), he can classify its spatial structure as a whole. The function here would take as input the retinal region the environment projects onto, and return as output the LTM environment, and associated

---

<sup>8</sup>Ultimately, I want to think some component of the RPC for objects comes are spatial representations derived from haptic exploration, as in Hayim’s project. Those are probably the representations of object shape that are computed in the ventrodorsal visual pathway. These representations are discussed more in Section [spatial representations in Part 2].

set of place SOM units—that is, the spatial representations that support actual navigation within the environment. It would have to be trained by a process that operates when the observer has *left* the environment, I think. Call the environment to be used in training  $E_1$ . The observer has to leave  $E_1$  and get into the embedding environment  $E_2$ . At that point,  $E_1$  will be represented spatially for the observer as an *object*, at a *place* in  $E_2$ . Crucially, this object would be associated with the LTM environment  $E_1$ : so the observer can *activate*  $E_1$  for training purposes even when he’s not in it. In this mode, I assume the active LTM environment  $E_1$  *also activates its full collection of associated place SOM units*, and these also function as training inputs for the function. This should allow the agent to learn *generalisations* about how *aspects* of the spatial structure of an environment is manifested in visual patterns. Hopefully, some of these generalisations will extend to unseen environments.

### 3.7.3 Spatial environment recognition

Obviously, spatial environments (e.g. rooms, streets) are also represented as token individuals: ‘my street’, as opposed to ‘a street’, and so on. I’m going to assume that *token environments are represented by LTM environments*. I’m already committed to this idea, since I’m envisaging a 1:1 mapping between LTM environments and LTM individuals (which are definitely tokens). This deepens the analogy with properties. (...)

## 3.8 The duality between individuals and environments

Here’s where I introduce the idea that each object can be re-interpreted as an environment. In the LTM system, this is implemented in a 1:1 mapping between LTM individuals.

## 3.9 Representation of groups of individuals as environments

### 3.9.1 A model of the perception of group individuals

Here, introduce the Walles *et al.* (2008, 2014).

### 3.9.2 When one object becomes a group

We use spatial language in an interesting way when referring to objects that break into pieces. The interesting phrase is *into*. When referring to the configuration of a static group of homogeneous objects, we can use the spatial expression *in*, as in *The soldiers were in a line*: this suggests that the configuration of the group is in some sense an environment that its component elements occupy. We can also use a trajectory expression like *into* to refer to an object’s trajectory into a regular spatial environment: for instance *The dog went into the kennel* describes the action of entering the kennel environment. These two devices can apparently be combined, to talk about objects breaking. When we say *The cup broke into pieces*, we are describing a transition of an atomic individual into an individual with parts.

(I think this goes beyond parts, and applies also to shapes: thus we can say *Mary bent the wire into a circle*, or *John curled into a ball*.)

## 3.10 Transitions between environments

### 3.10.1 The major axis of an environment

There are several reasons to suggest that the coordinate system we build for a given environment has labelled axes. Firstly, there is evidence that language picks up on such axes: for instance, if we talk about going *along* a corridor (or equivalently going *up* or *down* it), that indicates travel in the direction of its long axis; if we talk about going *across* a corridor, that indicates travel in

the direction parallel to the long axis. I will call the long axis the **major axis** henceforth, and the perpendicular one the **second axis**.<sup>9</sup> Secondly, in principle we need information about labelled axes to be able to represent transitions between environments. As well as relocating the agent's location in the new environment, we need to specify the angle between the major axes of the two environments.

Here, introduce the idea that the major axis is the direction your hand spends most time travelling in during haptic exploration. (I need to work this out.)

### 3.10.2 Transitions

[I think I can take some material from the new book draft...]

---

<sup>9</sup>(There is also an obvious 'up'-'down' axis that supplies the third dimension, but I won't focus on that here.)

## Chapter 4

# Motor control of the arm: the reach visuomotor pathway

The model of environment-centred spatial representations and spatial navigation introduced in Chapter 3 provides a very interesting basis for a more general model of motor control. In this chapter I'll focus on a model of control of the hand/arm system, specifically for reaching. (I will consider grasping in Chapter 6.)

### 4.1 Some interesting parallels between the navigation and motor control systems

I'll start by introducing a few points of contact between the model of environment-centred places and navigation and the model that needs to be built of the hand/arm motor controller implemented in parietal and premotor cortex.

In Section 3.1 I described how the agent can learn an allocentric representation of his location in the environment, using nothing but reafferent copies of navigation commands. The parietal cortex learns an allocentric representation of the agent's peripersonal space: in particular, of the location of the agent's hand in this space. It would be nice to think it could learn this representation using nothing but reafferent copies of motor commands to move the arm.

In Section 3.2 I described how the agent can learn representations of different environments, that permit navigation actions that avoid the boundaries/obstacles in these environments. The parietal cortex learns to manoeuvre the hand in cluttered spaces, so as to avoid obstacles and barriers.

In Section 3.3 I described how the agent can learn navigation actions, for some arbitrary given environment, that allow him to reach some arbitrary goal point in the environment, starting off at some other arbitrary point. The learning mechanism found optimal trajectories, but also stored sub-optimal ones in case the optimal one is unavailable. What's more, there was a specified method for selecting between trajectories—and having selected a trajectory, there was a method for implementing it without distraction from the other trajectories. (Inhibition was involved.)

In Section 3.4 I described a network that learns goal locations, independently of other plans, each associated with a trajectory. In the parietal/premotor pathway, I think the analogy of goal locations representations are *places associated with target objects*. There's a special function for computing these places, e.g. from vision.<sup>1</sup> And there are special routines for reaching each place, with a learned trajectory.

Finally, in Section 12.1.1 I described how object location memory and memory for rewards associated with objects can define a distribution of goal locations, that modulates the decision

---

<sup>1</sup>Analogous to view cells, maybe? I think I might be missing something.

about trajectory selection using nicely separated knowledge sources, that can be productively combined. It would be nice if trajectories in the hand/arm system could be similarly modulated.

In this chapter I'll introduce a model of reaching and grasping. The reach model is presented in Section 4.2 and motivated empirically in Section 4.3; the grasp model is presented in Section ?? and motivated empirically in Section 6.1.

## 4.2 A model of the reach visuomotor pathway

In this section, I will present a model of the pathway that represents and controls the state of the agent's arm, and thereby the location of the agent's hand. The model uses similar mechanisms to those used in the system that represents the agent's location in his local environment, and controls this location through navigation actions. The basic idea is to model the hand as a locomoting entity, within the peripersonal space of the body.

There are two components to the model. One is a circuit that learns an allocentric spatial representation of the agent's peripersonal space, using a recurrent SOM driven by efferent copies of the agent's arm actions. This SOM is similar to the places SOM in the navigation model (see Section 3.1), but its units learn allocentric representations of the location (and state) of the agent's hand, rather than representations of the location of the whole agent. This circuit is described in Sections 4.2.1 and 4.2.2. The other component is a circuit that learns representations of trajectories that take the agent's hand from arbitrary initial positions to arbitrary goal positions, using a recurrent SOM fed by an additional layer of goal units. This SOM is similar to the goal trajectories SOM in the navigation model (see Section 3.3.1.2), but its units learn trajectories of the hand in relation to the body, rather than of the agent in his environment. This circuit is described in Section 4.2.3.

### 4.2.1 The motor system commands SOM

To begin with, we need a representation of the various different motor commands that can be given to the arm, to make it move. Assuming a stationary shoulder, the arm has four degrees of freedom: two at the shoulder, and two at the elbow (allowing for rotation of the forearm round its own axis). Each degree of freedom is associated with a pair of opponent muscle groups.

I will assume a motor command to each of these muscle groups is represented using a coarse-coding scheme in an array of 10 neurons, designated  $n_0 \dots n_9$ . The coarse-coding scheme activates a 'bump' at some point in this array: a bump close to  $n_0$  represents no motor impulse, and a bump close to 1 represents the strongest possible impulse that can be generated by that muscle group. The impulses for the opposing muscle groups are represented separately, so as to model variable levels of stiffness of the joint. (A given arm position can often be maintained with several different possible degrees of stiffness; where larger forces are applied the opponent muscle groups, the arm will be stiffer, and less susceptible to deflection by an external force to deflect. As smaller forces are applied, the arm's movements will increasingly be a function of its passive dynamics.) In summary, to represent the motor commands that can be applied to the arm, we have 8 arrays of 10 units, each representing a motor impulse encoded as a bump.

We need a system that can represent a distribution of alternative possible motor commands to the arm at each point. Clearly we cannot envisage separate distributions within these 8 arrays, as this would create insoluble binding problems. Instead, we envisage a SOM, called the **motor system commands SOM**, that learns localist encodings of motor commands that frequently occur together. This SOM is not recurrent; its role is like the candidate episodes SOM in the WM/LTM system (see Section 9.2). Each unit in this SOM can potentially represent a high-level motor command for the whole motor system of the hand/arm, that activates a specific pattern of motor commands to the 8 individual muscle groups that control this motor system. We will call each such high-level command a **motor system command**. There is very good evidence that commands to different muscles are coordinated in this way; see e.g. ???. (Although I don't know anyone who has modelled high-level motor commands with a SOM like this.) Note that the

place-coded representations of motor impulse strength that provide input to the SOM allow it to associate impulses of different strengths with different SOM units.

Obviously, the motor system commands SOM will only learn a useful high-level representation of commands if there is some structure in the individual commands it receives; that is, if there are identifiable correlations, or principal components of variation, within the component fields of the input command. I will discuss why we can expect correlations to arise in Section 4.2.3.3; for now, I will just assume that they do.

## 4.2.2 The hand places SOM

I now envisage a recurrent SOM (an mSOM), analogous to the places SOM, that takes as input at each time point the current pattern of activity in the motor system commands SOM, as computed from efferent copies of actual commands to the 8 arm muscle groups. The SOM also takes a representation of its previous state: so it is set up to learn frequently-occurring *sequences* of motor system commands. I will call this SOM the **hand places SOM**;<sup>2</sup> its architecture is shown in Figure 4.1. (The motor system commands SOM is also shown.) The SOM is coupled with a

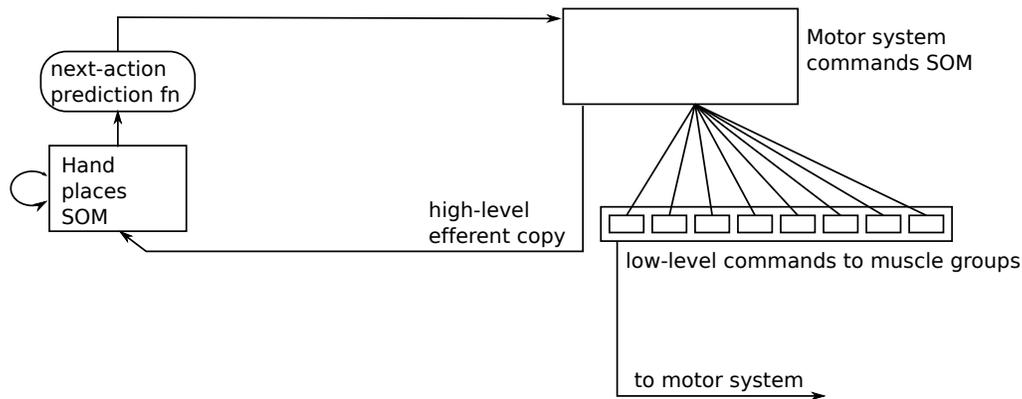


Figure 4.1: Architecture of the hand places SOM, and associated circuitry.

next-action prediction function, that is trained to map the current SOM state onto the action that actually occurs next. I envisage a simple actor-critic scheme is used for training, that delivers a few internal rewards that encourage the network to learn about the bounds of the space the arm can navigate, similarly to the scheme that trains the places SOM (see Section 3.1.1). Firstly, there is a reward associated with doing the same command you just did. Secondly, there is a punishment associated with commands that do not have any effect. (For instance, if a joint is at its limit, a command that attempts to move the joint past the limit is punished.) Thirdly, there is a punishment associated with commands that evoke pain.

### 4.2.2.1 How the hand places SOM learns a body-centred representation of locations in peripersonal space

I'll argue that after learning, the states of the hand places SOM will encode locations of the hand, in a coordinate system centred on the agent's body: in other words, that the SOM learns to transform a joint-centred representation of the location of the hand into a new frame of reference, centred on the agent's body, that makes no reference to joint angles of the arm. The argument is similar to the argument for the places SOM.

Firstly, being a recurrent SOM, the hand places SOM learns to represent commonly-occurring *sequences* of motor commands—which I will call *motor trajectories*. Each common motor trajectory will be encoded in a given SOM unit.

<sup>2</sup>Note, the motor system that controls the *arm* determines the *place* of the *hand*; hence the name for this SOM.

Secondly, there are constraints on the bounds of the hand's movements in joint space. These come from several sources. They are partly given by the limits of the four joints: each one has a maximal extent and a minimal extent. They are partly given by the fact that the elbow is attached to the agent's body, which is a barrier to the hand's movement. They are partly given by biomechanical constraints on the relative positions of different joints in the arm, arising from tendon stretching.

Thirdly, these constraints give rise to certain rough correspondences between motor trajectories and hand positions: certain motor trajectories always result in the hand reaching the same rough position. To take an extreme example, imagine the agent's arm joints are each at one of their limits, and the agent executes a sequence of joint impulses that move each joint progressively towards its other limit. Say it takes ten iterations to reach a boundary state where the hand can no longer move. The motor trajectory defined over those ten iterations can only occur in one way. So it always leaves the agent's hand in the same rough location.

Fourthly, the SOM states associated with unique locations can provide reference points for other locations, defined in relation to them. If a SOM state denotes a unique location, any movement made *from* this location will also be denoted by a SOM state, because the SOM is updated with movement commands. An important part of the learning process involves the establishment of cyclical patterns of movement. In a cyclical pattern, the SOM steps through a cycle of states, leading to a cyclic sequence of motor actions. In this case, hand positions *have* to correspond to SOM states. (I'm sure this is provable.)

Note I am not claiming that there is a single unit or region in the hand places SOM that is active whenever the agent's hand is at a certain body-centred location. Units in the SOM will be more readily associated with *motor states* of the hand, which involve a particular body-centred location, but also a particular body-centred speed. The hand places SOM would perhaps be better called the hand *states* SOM. The hand's speed is a crucial input to the motor controller, and there is lots of evidence that it is computed by integrating motor commands over time; see e.g. ???. A recurrent SOM is a good device for representing speed.

The representations of hand motor state learned by the SOM are certainly not given in joint-centred coordinates. But why should we think of them as *body-centred* representations? The answer is simple: the SOM learns the dynamics of a system of joints *attached* (at the shoulder) *to the agent's body*. The attachment of the arm to the body constitutes the fundamental constraint on how it can move: within this constraint, the SOM learns a set of ways the hand can move, just as the places SOM in the navigation system learns a set of ways the agent as a whole can move, within the constraint of a specified environment. All states represented by the SOM, therefore, are implicitly referred to the position and orientation of the agent's body.

#### 4.2.2.2 Parameterising the hand places SOM: body dynamics modes

In fact, the hand's movement in relation to the body is subject to different sets of constraints at different times. Some of these are due to the body's orientation: for instance, the hand's dynamics are completely different when the agent is lying on his back, compared to when he is standing up. Some of them are to do with objects near the body. For instance, if the agent is lying on a flat solid surface, there are many positions his hand can no longer reach. Other constraints are to do with the tiredness or coldness of the muscles; still other constraints are to do with loads that the hand might be bearing (for instance if the hand is holding something). All these factors mean that there must be *many* models of motor control for the hand/arm, rather than just a single model. This point was forcefully made by Wolpert and Kawato (1998). These authors suggest an approach in which each model is paired with a predictive model: the model (or models) that are best able to predict how motor commands update the hand's motor state are selected as the controllers to use at the current moment.

Note that our current SOM-based model of arm dynamics can readily be extended with a model of alternative dynamical regimes. We can envisage a medium analogous to the LTM environments medium (see Section 3.2), whose pattern of activity delivers a *tonic bias* on the hand places SOM. I will call this medium the **body dynamics modes** medium, and patterns of activity within it

**body dynamics modes.** Body dynamics modes can be learned by the same methods that govern learning of LTM environment representations (see Section 3.2.1). Note these same principles also account for how the agent *recognises* the current body dynamics mode: if the current SOM state is used as a *query* into the dynamics modes medium, it will indicate which modes are most consistent with the dynamics currently in place.

Note finally that some components of a dynamical regime change continuously rather than discretely. Some notion of weighted blends of body dynamics modes might be needed. This is something allowed for in Wolpert and Kawato’s model.

#### 4.2.2.3 Proprioceptive inputs to the SOM

To help the SOM learn, I assume it also receives proprioceptive input about the joint angles in the arm motor system. There are four joints; again I assume that angles are represented in a coarse-coding scheme, using an array of 10 units for each joint. Proprioceptive inputs therefore supply a further 40 input units for the SOM. (These input units do not need to be processed using a SOM, because there is no requirement to represent a set of alternatives; confidence about each joint angle can be expressed in its own array of units, by the width of the bump that encodes it.)

Of course, these proprioceptive inputs are expressed directly in terms of joint angles. But I want to argue that this is incidental: for the SOM, they function a little like ‘smells’ in the navigation circuit: perceptual stimuli that happen to be associated with particular allocentric places. As such, they provide very valuable orienting *cues* about the hand’s current location. But that’s all they are.

Note that while proprioception provides useful information about hand location, it does provide any information about hand speed: so it is all the more important that speed be computed by an internal model updated by motor commands. However, information about hand speed is directly delivered through vision; it is to vision that we now turn.

#### 4.2.2.4 Learning a visual representation of hand location/speed using the hand places SOM

The agent can learn to identify the location and speed of his hand using vision. In this section I will suggest a circuit that does this.

On the visual side, I assume the input to the function is the set of low-level feature maps computed in early visual areas (V1-V4), including feature maps responding to visual motion (e.g. in MT and MST) and maps responding to retinal disparity (e.g. in V3A, Tsao *et al.* (2003). (Retinal disparity is important for stereopsis, to compute the depth of the hand.) We also have to assume input from a medium encoding the angle of the head in relation to the body, and a medium encoding the angle of the eye in relation to the head. (One of the challenges for this function is that it must map retinal signals of the agent’s hand location into the body-centred coordinate system used by the hand places SOM.) The function must be trained to map these inputs to a suitable value of the hand places SOM.

I propose that the function is implemented in another SOM, that simply learns correlations between patterns in the hand places SOM, patterns in the head-angle and eye-angle media, and patterns in the set of retinal feature maps. The SOM is shown in red in Figure 4.2. The units in this SOM have a response similar to many parietal representations of visual object location: they represent a specific retinal location, but their response is ‘modulated by eye position’ (see e.g. Colby and Goldberg, 1999).

Since the visual function depends on prior learning in the hand places SOM, a natural developmental trajectory would be one where the visual function matures later. After it has matured, however, it provides a valuable new input to the hand places SOM, that can help it refine its representation of hand location. For instance, there are often several different joint configurations that place the hand at a given location in body-centred space. The hand places SOM cannot learn these invariances, even with proprioceptive information about joint angles. But the signals

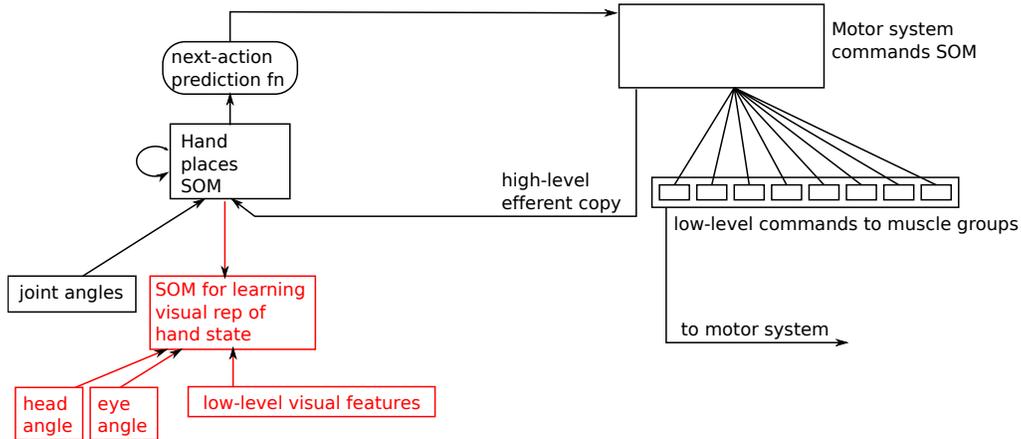


Figure 4.2: Architecture of a SOM (shown in red) that learns to identify hand states visually, using training data from the hand places SOM. During training, all the media linked to the SOM provide inputs. After training, the SOM can be used in another mode where input from vision (head and eye angles and low-level visual features) activates a SOM pattern, and this pattern is used to reconstruct a pattern in the hand places SOM.

arriving from a trained visual hand state recognising function will represent the hand at a given body-centred location in the same way, whatever the associated joint configuration is.

### 4.2.3 A system for learning trajectories to target objects

The hand places SOM introduced in Section 4.2.2 represents the arm’s actual movements, and in the process, the hand’s current place. We need to supplement this SOM with a system that represents goal states of the arm—that is, goal locations for the hand—and learns how to achieve these goals. In this section, I will introduce a SOM analogous to the goal trajectories SOM in the navigation model (see Section 3.3.1.2), which takes input from a medium representing ‘goal hand places’, and learns to represent motor trajectories that achieve different goals. This SOM is called the **goal hand trajectories SOM**.

I begin in Section 4.2.3.1 with a quick recap of how learning happens in the goal trajectories SOM in the navigation system.

#### 4.2.3.1 Recap: navigation goals and trajectory learning in the navigation system

In the navigation system, learning of goal states, and of the action sequences through which they can be achieved, involves two media: the navigational goals medium, and the goal trajectories SOM (see Section 3.3.1). As discussed in that section, the foundation for learning is an externally imposed reward schedule: in the example we used, if the agent perceives cue stimulus  $C_1$ , he will be rewarded if he reaches place  $P_1$  in his environment, and if he perceives stimulus  $C_2$ , he will be rewarded if he reaches place  $P_2$ . We begin by assuming a mapping between perceptual stimuli and patterns of activity in the navigational goals medium. In our example, cue  $C_1$  triggers activation of unit  $X_1$  in the navigational goals medium, and  $C_2$  triggers activation of unit  $X_2$ . States in the navigational goals medium *endure*, and therefore exert a constant influence on the goal trajectories SOM. The goal trajectories SOM and its associated action-prediction function are trained using temporal difference learning. At the start of learning, the tonic influence of the navigational goal unit that signals the reward scheme currently in place is of no practical use in determining how the agent acts. But over time, temporal difference learning ensures that this influence drives the SOM through a sequence of states that take the agent to the rewarded place. In temporal difference

learning, rewards are discounted in proportion to their distance from the current state, so the system is encouraged to find the shortest sequence of actions that leads to the rewarded place.

#### 4.2.3.2 Navigation goals and goal trajectories in the hand places medium

What are the analogue of navigation goal units like  $X_1$  and  $X_2$  in the system controlling reach movements to target objects? In some way, they should be representations of target objects themselves: specifically, the locations of target objects in the agent’s perispace. However, a newborn baby does not have a pre-existing concept of an external physical object. This concept has to be learned, through sensorimotor interactions with objects. In our model, the hand/arm motor controller plays an important role in this learning.

In our model of motor learning, the navigation goal units  $X_1$  and  $X_2$  were introduced simply as enduring representations of perceptual cues that identify the reward schedule currently in place for the navigation system. If we think about navigation goals in this way, we must assume there is a reward associated with reaching a target object. Several researchers have suggested that during training, the reach system treats *touch sensations* as intrinsically rewarding (see e.g. Arbib *et al.*, 2009 and many others). If this is the case, then putting an object into the agent’s perispace at a particular location imposes a particular reward schedule: the agent will be rewarded when his hand reaches that particular location.

Note that for a sighted agent, there is a visual cue that identifies the reward schedule created by the presence of an object at a given location: namely the retinal pattern produced by the object at that location. In some sense, this pattern is analogous to the cue stimulus (e.g.  $C_1$ ,  $C_2$ ) that signals to the agent a reward schedule that is currently in place. But in our discussion of navigation goals, we assumed a simple 1:1 mapping from cue stimuli ( $C_1$ ,  $C_2$ ) to navigation goals ( $X_1$ ,  $X_2$ ). In the case of visual stimuli, the mapping is more complex. We still have to assume a function that maps distinct patterns in the retinal array to distinct navigation goals that can guide reaching: it is important we can activate navigation goals *purely as a function of perceptual inputs*, because at the start of training, a navigation goal is *nothing more than* a function of perceptual inputs. However, we want this function to be *learnable*, so that over time, it *tunes* its mappings from perceptual inputs to navigation goals, and becomes better at identifying the opportunities for getting touch sensations by making arm movements. For this purpose, we will use a SOM, configured similarly to the SOM that learns to map between visual features and the agent’s hand states (see Section 4.2.2.4). We call this SOM the **visual reach affordances SOM**—or just the ‘reach affordances SOM’ for short. As we discuss below, a SOM is a perfect learning device for current purposes. Even before training begins, its winning units in some sense ‘represent’ the visual stimulus provided as input—so from the outset, sparse patterns in the SOM can stand in for visual patterns that identify the current reward schedule (i.e. the locations of target objects). At the same time, as training progresses, there is a mechanism that *tunes* SOM units, so they become *better* encoders of these visual patterns, and identify the specific visual stimuli that signal rewarded trajectories.

Importantly, this tuning process happens in parallel with the learning that takes place in the goal hand trajectories SOM. Learning in this SOM occurs just as it does in the navigation system (see Section 3.3.1). The reach affordances SOM is the analogue in the reach system of the ‘navigation goals’ medium in the locomotion system. Accordingly, the selected pattern in the reach affordances SOM *stays tonically active*, and exerts a tonic influence on the agent’s sequence of motor movements. The sequence of states activated in the goal hand trajectories SOM reflects these movements, but also, crucially, the tonic pattern in the reach affordances SOM. If we think of the system as a whole, the trajectory learned by the goal hand trajectories SOM is ultimately a function of the visual input at the time it was initiated. The SOM pattern is the intermediate point in this mapping from visual input to hand trajectory. There is a mechanism for optimising the trajectory that the SOM pattern generates, to get the highest discounted reward. And, separately from this, there is a mechanism for tuning *this same SOM pattern*, so that it becomes an increasingly good diagnostic of the trajectory that leads to this reward, by bringing the hand into contact with an object at a specific place.

The architecture of the circuit involving the goal hand trajectories SOM and the reach affordances SOM is shown (in red) in Figure 4.3. Note that the reach affordances SOM takes input

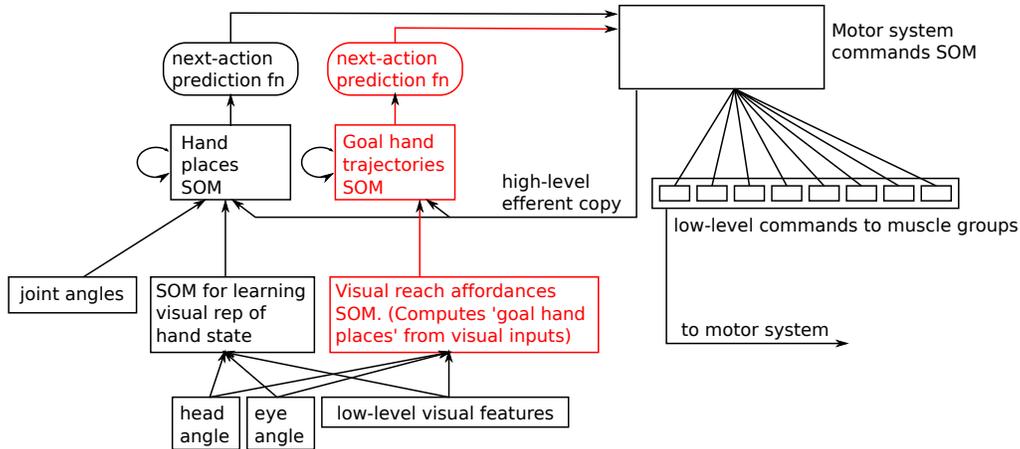


Figure 4.3: The goal hand trajectories SOM (and associated next-action prediction function), and the visual reach affordances SOM. (The visual reach affordances SOM also computes )

from the exactly the same visuomotor media used by the visual hand state SOM: a retinotopic map of low-level visual features, plus head and eye angle.

We conclude this section by discussing various interesting features of the circuit for controlling reach movements to target objects. (A discussion about neural regions that may correspond to components of the circuit is deferred until Section 4.3.)

**Parallel representation of reach goals** Firstly, note also that the visual reach affordances SOM can compute reach goals to *multiple objects in the visual field in parallel*. Each of its units receives input from all visual features at all locations on the retina. Units quickly come to specialise in specific retinal locations—and specific combinations of features at these locations—but importantly, there is nothing to stop units encoding objects at different retinal locations from firing simultaneously.

The situation where multiple reach goals are activated simultaneously in the reach affordances SOM is directly analogous to the situation in the navigation circuit, where multiple navigational goals are active simultaneously. As discussed in Section 3.3.2, the recurrent SOM that receives input from a collection of motor goals simultaneously can do a pretty good job of selecting between them, taking into account the costs of their associated trajectories. (Incidentally, the selection mechanism also takes into account the degree of activation of the different motor goals: so in the case of the reach affordances SOM, the selection mechanism will also show a preference visual stimuli which more strongly or clearly indicate reach goals.) Having selected a goal, the recurrent SOM also has a natural mechanism for ‘locking in’ this goal, and inhibiting its competitors; this is also discussed in Section 3.3.2.

Note that the mechanism for trajectory selection is likely to be confused if too many competing motor goals are given to it. (Again this is discussed in Section 3.3.2.) For this reason, we can expect *some* degree of selection to take place in the reach affordances SOM. As this medium is a SOM, we can impose an arbitrary sharpening on the pattern of activity expressed within it, so there are quite elegant mechanisms for finding the optimal degree of sharpness.

**The reach affordances SOM and object location memory** Note that patterns of activity in the reach affordances SOM can also be thought of straightforwardly as representing *locations*—for instance, locations where objects can be found. Of course, these locations are indexed to the agent’s body, not the environment: but if we know the environment-centred location and orientation

of the agent, we should be able to learn a function that momentarily maps between locations in the environment and body-centred locations in the reach affordances SOM. In combination with the environment-centred object location memory system, this function should provide a way of associating body-indexed locations in the reach affordances SOM with LTM individuals, via memory of their locations in the current environment. This in turn allows the highlighting of potential reach locations ‘top-down’, due to their likely association with rewarding or task-relevant LTM individuals or object types. The system is very analogous to the system for biasing the goal places in the agent’s environment to the remembered locations of rewarding or task-relevant objects, as discussed in Section 12.1.1. A memory-guided component to the process of reach selection is fundamental for unsighted agents, who cannot generate a representation of multiple potential reach targets in parallel over the visual field.

**!!The reach affordances SOM as a guide to eye movements** Another use of the reach affordances SOM is in controlling the agent’s eye movements. Say the agent has selected a particular reach goal in the reach affordances SOM, corresponding to some peripheral point on the retina, and is about to execute the associated reach trajectory. Though he is not obliged to foveate the object to be reached for, it is useful to do so—and agents typically do so, early in execution of a reach movement (see e.g. Johansson *et al.*, 2001).

It may be that the operation of foveating the target object can be performed directly within the reach affordances SOM. Here’s a potential model. It makes two assumptions. Firstly, I assume there’s a medium isomorphic with the retina, that can hold commands to foveate

[I presume this links to the orienting SOM somehow.]

**!!The reach affordances SOM as part of the orienting network** As already discussed in Section ?? [WRITE THIS SECTION NOW!blob], an agent with memory of the locations of objects (or object types) in his environment needs a way of *orienting* to an appropriate place when the object representation is activated: that is,

Together with the mechanisms for controlling for the movements of the eyes and head in relation to the body (see Sections 4.2.2.4 and 4.2.3.2), I think it is also a big component of the mechanism that creates ‘stability’ of the agent’s representation of his environment (see e.g. Henderson and Hollingworth, 1999 for discussion).

This idea is consistent with empirical evidence that the spatial attention system subserving ‘object perception’ is the same as that subserving action selection (see e.g. Schneider and Deubel, 2002).

This hooks back to the discussion of the ‘tracking map’ in Section 3.4.3.1. Recall from that section that the tracking map requires a representation of the low-level signatures of objects at each retinal location: this is something that has to be computed in the early visual mechanism, i.e. the retinotopic feature maps. [The idea has to be that the visual training might be *in the retinal feature maps*.]

As originally introduced, the tracking map had a role in learning environment-centred representations of the location of an attended object. Note that this object can be *out of reach*—in fact, in a large environment, it will normally be out of reach. It might be wondered whether the

Thinking about where these retinal structures might be: one possibility is that units in the visual reach affordances SOM might identify the relevant visual retinal patterns. They certainly identify *some* retinal patterns. But note they only identify patterns associated with objects that are ‘within reach’.<sup>3</sup> We have to envisage some other circuit holding visual patterns associated with objects at arbitrary distances. I assume that the circuits are internal to the retinal feature maps. Why would we expect these circuits to learn visual signatures of objects far away? I suggest that the circuit is always trained on objects *within* reach—but since these objects can be of arbitrary sizes, and the visual signature of a small object close up is the same as that of a larger object far away. (Modulo stereopsis, which has to be learned in the orienting network.) Remember that we

---

<sup>3</sup>What counts as ‘within reach’ of course depends on how long the agent’s arm is. See Umiltà *et al.* (2008) for a clear demonstration of this.

just need a way to *bootstrap* learning in the orienting network—and the visual signatures of reach targets (of all sizes) within the agents perispace provide a sufficient mechanism for this.

**Geometric attributes of proto-objects** Somewhere in here, I should also note that proto-objects as represented on the retina have a degree of form to them (see e.g. Driver *et al.*, 1994; Peterson and Kim, 2001). I want to say that this emerges naturally if you assume that tactile rewards are stronger for touches of *the whole palm*—or perhaps better, for touches of the joined fingertips. Better still, the reward could be particularly large not just for ‘gentle’ touches, but for touches that arrive without ‘slip’ sensations on the contacted surface. This internal reward schedule causes the agent to learn reaches that achieve an endpoint with a certain *orientation* of the hand, as well as a certain location. (The reaches will also have some degree of curvature to them, because in their final stages, the hand must drop ‘onto’ the object, in a direction orthogonal to the surface being contacted.)

**Moving targets** Note that visual features representing motion are among the features passed as input to the reach affordances SOM. These features should enable preparation of a hand trajectory to encounter a moving target, of the kind that would be required to catch a ball. There is good evidence that reach actions to moving targets are computed in advance; see e.g. Land and McLeod (2000). It is an attractive property of the current model that it handles stationary targets and moving targets using the same general mechanism.

[Maybe the patterns that track movements of proto-objects on the retina are also used to record movements of moving targets?]

**The reach affordances SOM as a model for visuomotor learning** Units in the reach affordances SOM are classic examples of ‘representations in a hidden layer’ of a neural network. They are the intermediate units in the network that maps visual inputs to motor outputs. They start off having no intrinsic meaning of their own, but over time, as their connections to the visual and motor systems develop, they take on meanings both as representations of visual stimuli and as representations of motor actions. In the learning model just described, the SOM unit that is initially selected to represent ‘the location of a target object’ is selected completely at random—but as learning progresses, it acquires both a visual meaning *and* a motor meaning. In the world of machine vision, algorithms for learning visual representations often use supervised learning: ‘for this visual input, you should return *this* unit’. In the algorithm described in this section, we have a way of identifying ‘*this* unit’ without external intervention. This algorithm will be a model for the other algorithms we describe for learning to map perceptual representations of objects onto their motor affordances—in fact, onto increasingly abstract and high-level affordances, for increasingly abstract and high-level sensorimotor and cognitive operations.

**Allocentric coordinates for reach trajectories** An interesting property of human reach movements is that the hand travels in a roughly straight line in Cartesian space (see e.g. Flanagan and Rao, 1995). I believe learning in the goal reach trajectories SOM will cause hand movements to behave like this. (But it’s just a guess at present.)

**Optimal learning in the goal hand trajectories SOM** More generally, there is good evidence that human reach trajectories are optimised on some metric. This results in several characteristic properties, to do with path curvature and also velocity profile, which has a characteristic bell shape (see e.g. Harris and Wolpert, 1998). I hope that something in the trajectory-optimising system yields this shape as well. One relevant point is that the agent’s internal reward system should probably reward *gentle* touch sensations more than *strong* touch sensations. This will encourage trajectories in which the hand is almost stationary when it reaches the target object. That pushes in the right direction for obtaining a bell-shaped velocity profile.

### 4.2.3.3 Coda: learning in the motor system commands SOM

In Section 4.2.1 I left a promissory note to explain why the motor system commands SOM can be expected to be trained on arrays of low-level motor commands in which there is some structure—that is, in which there are regular correlations between the strengths of motor impulses given to different muscle groups in the arm. As noted there, the motor system commands SOM will only learn sensible representations of commands to the whole ‘arm’ motor system if there are correlations between motor commands represented in its different fields. There are many reasons to expect to see such patterns. Some relate to learning in the hand places SOM. For instance, as discussed in Section 4.2.2.1, training in this SOM encourages the hand to fall into *cyclic* patterns of motor movements. The individual actions in each phase of such a cycle are all distinct; the fact that each of these actions repeats many times creates clear correlations between signals in the different muscle groups associated with the arm. Other reasons to expect correlations relate to learning in the goal hand trajectories SOM. For instance, if this SOM is trained to reach targets at arbitrary locations in peripersonal space, and the optimally rewarding reaches are ‘gentle’ reaches (as suggested at the end of Section 4.2.3.2), we can expect the hand to frequently be *at rest*, or very nearly at rest, in every point in peripersonal space. The patterns of motor commands that generate these resting states will also each be frequently produced. For another thing, the goal hand trajectories SOM encourages the learning of *optimal* trajectories. Often, the most efficient way to move the hand along a given Cartesian trajectory is through one particular combination of joint movements. So these particular combinations of joint movements will also occur frequently, and will therefore be learned by the motor system commands SOM.

Of course, we have to envisage some form of bootstrapping taking place, whereby initially crude representations of complexes of motor movements in the motor system commands SOM enable *some* learning in the subsequent SOMs, which in turn generate better signals from which the motor system commands SOM can train. Whether this kind of bootstrapping happens is a big question!

## 4.3 Evidence for the reach model in parietal/premotor cortex

(Specifically, the dorsodorsal pathway.)

I can definitely refer to Cisek and Kalaska (2005)’s paper showing evidence for decisions about reach targets here. That’s evidence for the goal hand places SOM, I think.

## Chapter 5

# Representations of object geometry, and hierarchical spatial representations

The visual classification of an object involves activity in the ventral processing stream, as discussed in Section 2.7.1, but also in the dorsal processing stream through parietal cortex (see classically Goodale and Milner, 1992, Ungerleider and Haxby, 1994). The dorsal stream computes representations of three-dimensional object shape, that are used to identify motor affordances. These representations can operate somewhat autonomously from ventral representations (), but in normal function they contribute to the representations of shape (Bracci and Op te Beeck, 2016) and orientation (Schendan and Stern, 2007) that are computed in ventral cortex.

In this chapter, I'll introduce a model of the parietal system that computes affordance-based representations of object geometry, and the visual pathway that links to this system. I assume the original affordance-based representations are learned through tactile exploration, and don't involve vision at all. The system doing this 'blind' learning of object geometry is introduced in Sections 5.1–5.6. The remainder of the chapter will consider how visual representations are mapped onto these geometric representations, both directly from early vision (Section 5.4) and indirectly, through links to ventral object representations (Section 5.5 and 5.7).

### 5.1 Affordance-based representations of physical objects

This is Hayim's stuff.

### 5.2 Representations of object types in parietal cortex

The motor system needs a notion of 'object types', so that general rules about motor interactions with objects can be represented. These types roughly line up with the system of natural kinds, whose definitions make reference to motor affordances.

My basic proposal is that an object type within Hayim's system is like the representation of an *environment type* in the hippocampal navigation system (see Section 3.2).

### 5.3 Representation of object dimensions in parietal cortex

Types in the motor system must be 'transformable' to fit the tokens that instantiate them—otherwise they're of no use. So when you classify an object as a type *in the motor system*, you obligatorily execute a transformation if it's necessary.

In here, start by pointing to evidence that agents’ object representations are sensitive to the **major axis** (or the ‘radial axis’) of objects. You can refer to the idea that infants learn to orient the major axes of objects perpendicularly to the line of sight (see e.g. Pereira *et al.*, 2010). Also mention that IT computes axis-based shape representations (see e.g. Hung *et al.*, 2012), and that both dorsal and ventral visual areas compute specialised representations of ‘elongated’ objects, somewhat independently of object type (Bracci and Op de Beeck, 2016).

I already proposed a model of major axes in Section 3.10.1, in the context of environment representations. The key idea there was that the major axis is the direction the agent spends most time travelling in during exploration. This same idea can be extended into the domain of 3D object representations learned through haptic exploration. (I need to put the details here.)

## 5.4 Mapping from early vision to parietal object representations

### 5.4.1 Computation of 3D object type

I assume a SOM-based convolutional network of the kind described in Section 2.7 is used to learn to map retinal features onto 3D object types. The network in parietal cortex is different, in that there’s a clear source of *supervision*: during training, we know what desired 3D object type is. (I assume supervision happens by providing 3D object type as an additional input to the top-level SOM, as discussed in Section 2.7.5.)

### 5.4.2 Computation of shape category transformations: shape and size properties

The SOM-based convolutional network in fact has two outputs: one is 3D object type, another is a set of transformations. Again, this network can learn using supervised methods.

## 5.5 Mappings between dorsal and ventral object representations: types, size and shape

### 5.5.1 Influence of motor types on ventral types

The object types that are represented in the ventral visual pathway are ‘natural kinds’: that means they are defined in ‘functional’ (action-related) terms as well as in terms of their visual properties. I suggest that some of the functional influences come from affordance-based representations of 3D object types in parietal cortex, as discussed in Section 5.2—i.e., *shapes*.<sup>1</sup> I assume these 3D shape representations provide an input to the top-level SOM in ventral cortex, so that object categories learned by this SOM also make reference to shape-based motor affordances. (Having learned this, the input from parietal cortex can be removed, and the type system won’t change: this effect emerges from a SOM’s ability to respond appropriately to partial inputs.)

There is good evidence that agents store a mapping between object types (in IT) and ‘prototypical’ object shapes/affordances (in parietal/premotor cortex). For instance, Jeannerod *et al.* (1999) asked a patient with damage to the grasp pathway in parietal cortex reach for target objects of two kinds: one kind (rectangular blocks) whose category gave no indication of object size; the other (natural implements tools such as lipstick tubes and cups) whose category did roughly indicate size. The patient had difficulty shaping her hand to the former type of object, but less difficulty for the latter type. Hu and Goodale (2000) showed a similar effect in normal subjects. They presented agents with a target object flanked by a second object inducing a perceptual size

---

<sup>1</sup>‘Functional’ can be understood on different levels. At a higher level, different plans or action types are afforded by different object types. I assume this is implemented by having the distribution over possible ‘episodes’ conditioned on ventral object representations (see Chapter 13 for details I think).

illusion originating in ventral cortex. If subjects reached for the target object when it was visible, their grasp preshapes were not affected by the illusion, but if the object disappeared before the reach took place, preshapes were affected, indicating a way for size information from ventral cortex to influence the size parameter of a parietal affordance-based object representation.

There's also evidence that the parietal system for representing objects with different orientations works in tandem with the ventral system that does the same thing. For instance, Schendan and Stern (2007) showed that a set of areas from both dorsal and ventral streams were jointly active in both a static object classification task and a 'mental rotation' task. In the ventral stream, these comprised dorsal occipitotemporal areas adjoining the lateral occipital sulcus (DOT-LOS) and the inferior temporal sulcus (DOT-ITS). In the dorsal stream, these comprised dorsal foci in occipital cortex (DF1) and in the ventral caudal intraparietal sulcus (DF2).

### 5.5.2 Representations of shape and size in the ventral visual pathway

Recall from Sections 5.2 and 5.3 that parietal cortex computes 'transformations' of the prototypical motor type (i.e. a 3D shape representation), which are represented declaratively in their own right in parietal cortex. I assume that there's a function that learns a mapping from ventral cortex object type representations to these transformations. After training, this function represents the 'expected' transformations (of size, and of relative dimensions) for any given category. I assume these can be contrasted with the *actual* transformations (of size and relative dimensions), to give declarative representations of size concepts like 'big' and relative dimensions concepts like 'thin', 'long' in parietal cortex.

I assume there is a function that maps these parietal declarative representations into a corresponding medium in ventral cortex. This medium is unusual in holding properties of objects that are computed as a direct side-effect of object classification, rather than as the result of a subsequent property-level IOR operation.

A circuit for this mapping operation is sketched in Figure ?? . (...)

Note: maybe after the mapping from parietal to ventral size/shape representations has been learned, ventral representations of size and shape can also be computed directly from vision. That is, maybe an agent can look at a long pen, or a big cup, and compute ventral property representations 'long' and 'big'. I'm not sure if that will work.

**Ventral-dorsal mappings for size** We know that the ventral visual system also computes representations of object size. For instance, Konkle and Oliva (2012) show that object representations in occipitotemporal cortex have medial-to-lateral organisation reflecting object size. We also know the ventral visual system's representations of object size can influence the motor system; see again Jeannerod *et al.*'s (1999) experiment showing the difference between known and unknown object manipulation in patients with parietal damage.

Some information about object size is better computed purely within the ventral cortex. We can determine object size based on the size of the retinal region containing the object (that is, of the *attended* retinal region).

Some very subtle results about size representations are found in Fyshe (2015). Maybe also see Bemis and Pyrkänen's papers.

**Ventral-dorsal mappings for shape (specifically 'length')** We also know that the 'elongation' of an object, as computed by the parietal visual stream, exerts an influence on object classification processes in IT. see e.g. Almeida *et al.*'s (2013)

**A link to size/shape representations in language?** (In Section 22.2 I will link the neural circuits computing size/shape and other properties to a model of adjective ordering in language. Kemmerer *et al.*, 2009 give evidence that a difficulty with adjective ordering is associated with damage to inferior parietal cortex, which fits in with the model I'll propose.)

## 5.6 Representations of object parts and relations between objects

### 5.6.1 Object parts

I suggest that ‘parts’ of a 3D object are originally represented in terms of motor affordances. Consider a cup, which has a handle. The affordance-based representation of the cup’s geometry introduced in Section 5.1 identifies various *places* on its surface. One of these places is occupied by an object: a handle. This handle can also be established as an environment. That is, it’s an environment *indexed to the cup environment*. I suggest that an object’s parts are objects that are thus indexed to another object. The geometric relation between the cup and its handle is represented by the operation that maps from the cup environment to the handle environment.

### 5.6.2 Relations between objects

## 5.7 Mappings between dorsal and ventral object representations: object parts and relations between objects

In Section 5.6, I suggested that the location of an object part is originally defined in the affordance-based object representation system. However, this *motor* location can be mapped to a *retinal* visual location. Recall that a purely visual definition of the ‘parts’ of a salient stimulus was introduced in Section 2.11.5. I envisage a learned mapping between these two notions of object parts. (...)

Ditto for relations between objects.

In linguistic terms, these relations are the denotata of *locative prepositions*: e.g. *on*, *around*.

## Chapter 6

# The grasp visuomotor pathway

### 6.1 Evidence for the grasp model in parietal/premotor cortex

(Specifically, in the ventrodorsal pathway, maybe? But we're not talking about manipulatory actions, so that theoretical label isn't perfect...)

### 6.2 Towards a definition of hand/arm motor programmes?

The above analogies only address how to define, select and execute optimal trajectories that transport the hand to a target object. They don't define the kind of 'motor programmes' that are denoted by action verbs. (Except maybe the special verb 'touch'.) So how are motor programmes like *punch*, *squash*, *snatch* etc defined?

I suggest using the approach of Lee-Hand and Knott (2015), in which motor programmes of that kind are defined as *deviations* from the default 'touch' trajectory. Motor programmes defined this way would improve on those in Lee-Hand and Knott, in several ways. Firstly, the touch trajectories would be *optimised*: they wouldn't just be implemented in a feedback controller. (The learned system would be more like a feedforward controller of some kind.) Secondly, the touch trajectories are defined *for all initial states* of the hand/arm, as well as for all possible target locations.

### 6.3 A model of causative motor actions

Possibly this can go here. It sits nicely after the mirror system section, because it can help give an account of actions that *aren't* in the mirror system. (The basic idea there is that since causative actions are defined by their external perceptual effects, which are as easy for an observer to see as for the agent, all that an observer has to be able to in order to recognise a causative action is to recognise that the agent's motor action did indeed cause the observed effect.)

### 6.4 Transitive and intransitive body-centred actions

[This section is probably out of place.]

An analogy with the spatial navigation that's slightly less easy to see concerns how motor goals are defined. In the navigation system, I described two sorts of motor goal. One was where (under some circumstance) a goal location was *intrinsically* associated with reward: see Section 3.3.1. The other was where a goal location is reached because it is currently occupied by some *object* that is intrinsically associated with reward: see Section 3.4.5. I suggest that the former type of motor

goal is that associated with an intransitive action: for instance, the action of shrugging, where the goal motor state (of the shoulder) is simply a point in peripersonal space, or the action of walking, where the goal motor states (of the feet/hands) are simply points in peripersonal space.<sup>1</sup> And I suggest that the latter type of goal is that associated with a transitive action. The point of a transitive action is to interact with some target object: to do so, of course, you have to get your hands on it. (...)

---

<sup>1</sup>In the latter case, of course, these goals are modulated by a central pattern generator, and perhaps also by high-level navigation commands to do with speed and turn angle.

## Chapter 7

# Action execution and action perception: the self-other distinction

### 7.1 Early stages in the action perception pathway

#### 7.1.1 V1 units representing simple motion patterns

The ‘complex cells’ in primary visual cortex are actually of two types: one type respond to a simple visual feature anywhere within a certain area of retina; another type respond to a simple visual feature if it happens to *move* in a certain direction within this area (see again Hubel and Wiesel, 1962). The former type of cell is well modelled by units in the pooling SOMs described in Section 2.7.3. The other type can be modelled by units in a layer of SOMs very similar to pooling SOMs, which I will call **motion pattern SOMs**.

Motion pattern SOMs tile the retina in the same way as pooling SOMs, and take the same inputs: each motion pattern SOM takes input from all the local SOM units in its area of retina, gated as usual by saliency. However, while pooling SOMs learn by temporarily ‘clamping’ their active units, to force given token units to represent consecutive stimuli, motion pattern SOMs have a *recurrent* input, so they learn commonly occurring *sequences* of simple visual stimuli in their local area of retina. The layer of motion pattern SOMs is illustrated in Figure 7.1.

Motion pattern SOMs can learn the kind of moving oriented stimulus that complex cells are thought to encode. This is because oriented stimuli often move smoothly through consecutive locations while maintaining their orientation. However, stimuli that move smoothly through consecutive locations often also change their intrinsic visual features: for instance, a moving oriented stimulus can also smoothly change its orientation. I have not seen any reports of V1 cells with

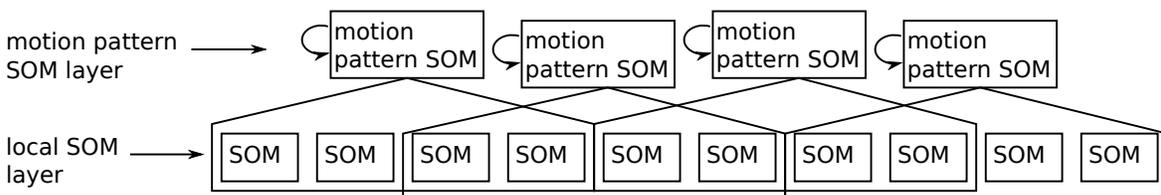


Figure 7.1: A layer of motion pattern SOMs taking input from the local SOMs. Each motion pattern SOM takes input from the local SOMs in a particular region of the retina, and learns to represent temporal sequences of patterns expressed in these SOMs.

this behaviour. They certainly exist further on in the visual pathway, in particular in the areas MT and MST, which are specialised for encoding patterns of motion. (I will talk more about these areas in Section 7.1.) I think their absence from V1 may result from the very small receptive fields of V1 motion-sensitive cells: they are around  $1\text{-}2^\circ$  of visual arc at the fovea (see e.g. Rolls, 2009). MT/MST cells have larger receptive fields, and can therefore encode rotational changes of oriented stimuli as well as directional movements.<sup>1</sup>

I should also note that motion pattern SOMs should learn special representations for *stationary* stimuli. A stationary pattern is a special kind of temporally extended pattern—probably the most common kind, in fact.

### **7.1.2 MT and MST: encoding patterns of motion**

### **7.1.3 STS: joint attention and biological motion recognition**

## **7.2 A model of ‘the mirror system’ for reach/grasp actions**

---

<sup>1</sup>Note the receptive fields for motion pattern SOMs need not be the same as for pooling SOMs, even if this impression is given in Figures 2.8 and 7.1.

## Chapter 8

**Summary: SM mechanisms  
involved in apprehending a simple  
transitive action**

## Part II

# Working memory and long-term memory

## Chapter 9

# Memory representations of individuals, episodes, situations and plans: an introductory model

### 9.1 Motivating ideas

#### 9.1.1 Episodes are perceived in sequentially structured SM routines

WM representations of experienced events have to be *created during experience*. Events take time to occur, so the SM processes through which they are experienced must be similarly extended in time. The founding assumption in our model is that event-perception processes have a well-defined temporal structure—and that the mechanism representing events in WM capitalises on this well-defined structure. In this section we outline what this structure is; for details, see Knott (2012).

We argue that perceiving an episode involves a relatively discrete *sequence* of SM operations. This assumption rests on some well-accepted findings about perceptual processes. Firstly, there is good evidence that focal attention must be allocated to an individual in order to represent it in any detail in the object classification system in inferotemporal cortex (see e.g. Zhang *et al.*, 2011). If an event involves several participant individuals, therefore, the observer must attend to them one by one, rather than in parallel. Secondly, when an event is perceived, participants playing certain semantic roles are recognised first. For transitive events, we argue the AGENT participant must be attended to before the PATIENT (Knott, 2012).<sup>1</sup> If the observer is executing the action, this is because the decision to act must precede selection of a target; if the agent is watching an action, it is because s/he must monitor the agent to identify the intended target (Webb *et al.*, 2010). Thirdly, a representation of the motor action cannot be evoked until the target object has been attended to. In action execution, the agent must activate a representation of the target object before its motor affordances can be computed (Johansson *et al.*, 2001); in action perception, the observer must compute the trajectory of the agent’s hand onto the target (e.g. Oztop *et al.*, 2004). If these assumptions, which are individually quite well accepted, are brought together, an interesting model of event perception emerges, in which apprehending a transitive event involves a sequence of three SM operations: attention to the agent, then attention to the target, then activation of a motor programme. The idea that events have a characteristic temporal structure is certainly present in other models of event perception, in particular that of Reynolds *et al.* (2007). For Reynolds *et al.*, these sequential regularities relate primarily to the structure of an agent’s *movements*: they are the kind of regularities that the ‘biological motion’ system becomes attuned to. In our model, such regularities are encoded within the action representation system, as discrete

---

<sup>1</sup>Our terms ‘agent’ and ‘patient’ refer to Dowty’s (1991) more general concepts ‘proto-agent’ and ‘proto-patient’.

actions. But there is more to an event than an action. In our model, experiencing an event also involves a higher-level sequence, of relatively discrete *SM operations*. One of these is the activation of an action representation. But this operation must be *preceded* by an action of attention to the agent, and then an action of attention to the patient (if there is one). In our model, the notions of agent and patient are in fact *defined* by the serial order of attentional operations in this SM sequence: the (proto-)agent is the first individual attended to; the (proto-)patient is the second.

### 9.1.2 Individuals are perceived in sequentially structured SM routines

Alongside this model of event perception, we also assume that the perception of each participant in an event involves its own canonically-structured sequence of SM operations. It is well established that in order to classify an object, an observer must first direct focal attention to the region of space it occupies. But observers can also attend to a region of space containing a homogeneous group of objects. Walles *et al.* (2014) argue that in between focal attention and object classification there is an intervening attentional operation that selects a spatial *scale* at which the classifier will be deployed, determining whether the classifier identifies the local or global form (Navon, 1977) of the attended stimulus. This operation determines whether a single individual is classified or a homogeneous group of individuals. In summary, perception of an individual involves a SM routine comprising three operations: selection of a salient region of space, then selection of a classification scale (determining whether a singular or plural stimulus will be classified), and finally activation of an object category. Event perception, in turn, is a higher-level sequential SM routine, some of whose elements have their own sequential structure.

### 9.1.3 Individuals and episodes are represented in WM as prepared SM routines

We propose that representations in semantic WM exploit the sequential structure of perceptual processes. Specifically, we propose that WM representations of both individuals and episodes take the form of *prepared sequences* of SM operations. This proposal is attractive for several reasons. For one thing, it offers a clear account of how semantic WM representations can influence SM processing: a prepared sequence of SM operations is an ‘executable’ structure, that can initiate sequentially structured SM activity (including actions). For another thing, it suggests an account of a puzzling recent finding: stimuli held in WM appear to be transiently *reactivated* in SM areas during the delay period (see e.g. Meyers *et al.*, 2008). If WM representations are prepared SM routines, that can be executed in simulation, then active simulation processes could occur during the delay period, resulting in these transient patterns of SM activity. Finally, the proposal places semantic WM representations within a class of neural representation that is relatively well understood. We know a lot about how prepared sequences of attentional or motor operations are represented, because they have been intensively studied, in both animal and human experiments. There is evidence that prepared SM routines are stored in static, declarative patterns of activity, as discussed in detail in Section ???. In our model, semantic representations have this character: they are declarative patterns, that when activated, trigger the execution of a prepared sequence of SM operations.

### 9.1.4 Event representations make use of pointers

Modelling semantic WM representations as prepared sequences suggests a novel account of how semantic roles are bound to participants in representations of events. Our account makes use of four ideas, which we introduce here.

The key idea is that the binding mechanism is implemented as part of the active process of rehearsing SM routines, rather than within a static representational structure. The classic binding problem arises because the SM media representing an individual’s properties (location, shape etc) naturally represent just *one* individual: if the properties of several individuals are represented, it is hard to specify which properties belong to which individual. If a WM event representation

supports the simulation of a sequential SM routine in which representations of agent and patient are active in these media *at different times*, many of these problems go away.

Of course, the event representation must still make reference to both participants, so it can activate these temporally separated representations. The second idea in our binding scheme is that event representations represent participants using *pointers* into the medium representing individuals—and that there are separate pointers for agent and patient. The pointers are active simultaneously in a WM event representation, but they are only *followed sequentially*, when an event is rehearsed. In neural networks terms, agent and patient are coded ‘by place’ in our WM event representations, in separate groups of units. Place coding of this kind is not normally seen as a viable way of implementing role-binding: a simple place-coding scheme suffers from the fact that there is nothing in common between representations of John-as-agent and John-as-patient (for a discussion see Chang, 2002). But if the place-coded representations of agent and patient just hold pointers into the medium representing individuals, which are activated at different times, this problem does not arise.

The third idea in our binding scheme is that the place-coded pointers in WM event representations do not point directly to SM media representing individuals, but rather to a *WM* medium holding representations of individuals. Recall that representations of individuals also have internal structure: we proposed above that the WM representation of an individual is also stored as a prepared, replayable SM routine. In our model, WM representations of recently-perceived individuals are held in a separate WM medium: the agent and patient representations in a WM event point to, and sequentially re-activate, representations within this WM medium. During rehearsal of a WM event, these sequentially reactivated representations create opportunities for secondary rehearsal operations, simulating the steps involved in perceiving the participant individuals. This scheme introduces a measure of hierarchy in the model of role-binding, enabling the representations filling semantic roles to have a degree of internal structure—an important requirement, as noted earlier.

The fourth idea in our binding scheme relates to *how* the agent and patient fields of the WM event medium ‘point’ to particular WM representations of individuals. Our key suggestion is that they point to the *location* of individuals, rather than to their intrinsic properties. Several individuals can have the same intrinsic properties, so referring to these by themselves is not sufficient. However, note that episodes take time to occur, and the locations of their participant individuals might change as they are monitored. (Indeed, the semantics of many events *requires* that participants move.) Our suggestion here is that pointers to the agent and patient are held in the visual ‘tracking maps’ discussed in Section 2.6.4. We envisage an ‘agent’ tracking map that is part of the ‘agent’ field, and a ‘patient’ tracking map that is part of the ‘patient’ field. These maps are initialised when the agent and patient are attended to in succession, and thereafter, they hold ‘pointers’ to the locations of the agent and patient while the event is being monitored.<sup>2</sup>

Our use of pointers in event representations has a lot in common with that of Kriete *et al.* (2013): they also assume event representations include place-coded fields for agent and patient, that holding pointers to a single medium for holding object representations. There are several differences, though. Firstly, we see the media holding pointers as components of an inherently sequential plan, with the ‘agent’ medium representing an earlier operation than the ‘patient’ medium. This view stems from our view of event perception as inherently sequential. (Kriete *et al.* have no account of how the agent and patient fields are initialised during event perception, although it is implicit in their account that the two fields must be initialised at different times.) Secondly, we see the medium that is pointed to by agent and patient representations as holding another layer of plans, rather than directly to object representations. This enables an account of the internal structure of object representations, that parallels the account of the internal structure of event representations. Finally, we take advantage of place-coded agent and patient representations to define *localist* representations of whole episodes, as we now discuss.

---

<sup>2</sup>When an event is completed, or stops being monitored, these tracking maps allow for any changes to locations (and other properties) of the agent and patient to be recorded in static memory. This will be discussed further in Section 12.6.2; for the moment, the key idea is just that agent and patient hold pointers into a WM medium representing recently-attended individuals.

### 9.1.5 Localist representations of events, and an associated probability model

Many current models of cognitive representations have a Bayesian flavour. The brain must often represent some variable in the world which could have different values at different times: for instance, the object that is currently being classified, or an action that is currently being planned. In a Bayesian model, the brain does not attempt to identify the ‘actual’ value of the desired variable in such cases, and represent just that. Rather, it represents a probability distribution over all possible values of the variable. Working with probability distributions has many computational benefits; moreover there is good evidence that the brain does work with probability distributions (see e.g. Kiani and Shadlen, 2009; Pouget *et al.*, 2013).

A Bayesian representation of an event—for instance, a perceived event, or an event the agent intends to carry out—would be expressed as a *probability distribution over possible events*. However, if an individual event is represented as a complex pattern of neural activity, it is hard to represent such a distribution: overlaying several complex patterns is likely to lead to binding problems. There are a number of technical solutions to this problem: for instance, events can be expressed as points in high-dimensional spaces, where they are unlikely to interfere with one another (see e.g. Stewart and Eliasmith, 2012). But even in high-dimensional spaces, it is difficult to overlay large numbers of event representations in a way that keeps them distinct from one another.

Using place-coded representations of the agent and patient of an event opens up another approach, which is to represent events *in localist units*. A localist unit can represent an event through the connections it has into the place-coded medium: it can have separate connections to a representation in the ‘agent’, ‘patient’ and ‘action’ areas in the medium. When this unit is activated, these connections will cause the activation of a complete, complex, place-coded event representation. This kind of localist event representation is often called a ‘convergence zone’ (see e.g. Damasio and Damasio, 1994).

The idea that events are represented as convergence zones is widely invoked by neuroscientists, especially in models of the hippocampus. But the idea does not work computationally, unless there is a means for distinguishing the agent of the event from the patient. Our place-coded model of agent and patient representations supplies this means. (It should be remembered that this place-coded model in turn rests on a particular model of the sequential structure of event perception—so ultimately the localist account of event representations rests on a particular model of event perception.)

Our localist model of event representations is expressed using a self-organising map or **SOM** (Kohonen, 1982). In this model, individual events are represented as localist units in the SOM, and probability distributions are expressed as patterns of activity over these units. A SOM is an unsupervised learning device, that learns localist representations of patterns in the data it is exposed to. Its learning mechanism encourages it to represent frequently occurring patterns in detail, and less frequent patterns in units encoding generalisations; these learning mechanisms create useful representations of episodes, as we discuss in Section ??.

### 9.1.6 Localist representations of ‘situations’

The most derived representations in our cognitive model are representations of ‘situations’. Each situation is associated with a full probability distribution over events: it is a very rich structure, that supports the agent in experiencing the world, and also in representing experiences in long-term memory. In our model, the ‘current situation’ is represented in the hidden layer of a recurrent network that learns to predict the *next event*, given the event that has just occurred, plus a copy of its own previous state. After training, this network predicts a full distribution of possible next events in the SOM holding localist representations of events (exploiting its ability to represent multiple events). This recurrent network is also expressed as a SOM: in this case, a SOM with a recurrent input. Each event that is experienced updates the representation in the SOM, which induces a new distribution over predicted next events. Again, details are given in Section ??.

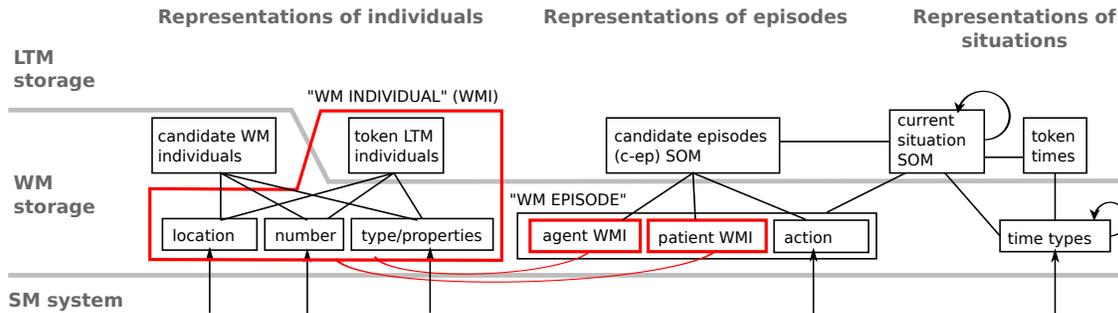


Figure 9.1: Architecture of the model of WM individuals and WM episodes

## 9.2 A circuit for representing individuals, episodes, situations and plans

Our model is illustrated in Figure 9.1. SM media are below the grey line; WM media are above it. The WM system representing individuals is on the left, and that representing events (or ‘episodes’, as we call them here) is on the right. The copy operations implementing pointers are highlighted in red.

The WM medium on the left holds a representation of a single selected individual, a **WM individual**, stored as a prepared sequence of a **location**, a **number** (i.e. classification scale) and a set of perceptual **properties**. These three representations are activated in parallel in the WM medium, but when the prepared sequence is executed or rehearsed, they activate associated first-order representations in the attentional and classification systems one at a time, as discussed above.

The media representing a WM individual provide input to another layer, the **candidate WM individuals (cWM-ind)** layer, which stores *combinations* of location, number and type over a short time period, and thus represents the set of recently-attended individuals. A partially specified WM individual can function as a query to the cWM-ind layer: if we specify a location, we may be able to retrieve an associated type and number (and vice versa). If we can, then the individual retrieved is classed as ‘old’; if we cannot, it is classed as ‘new’. These attributes are recorded in a **status** field of the WM individual, which is not part of the prepared sequence. Queries formed from partially-specified WM individuals can be used to generate expectations about the location or properties of individuals in the current scene, as we discuss below.

The **WM episodes** system is structurally similar to the WM individuals system. It holds a representation of a single selected episode, a ‘WM episode’, stored as a planned sequence of operations activating an **agent**, a **patient** and an **action**. As noted above, the agent and patient media hold content-addressed pointers to representations in the WM individual medium. All the media within a WM episode are active in parallel, but when a WM episode is executed or rehearsed, the representations they point to become active sequentially: the ‘agent’ and ‘patient’ media activate two successive representations in the WM individual medium, and then the ‘action’ medium activates a representation in the (pre)motor system.

The prepared operations in a WM episode also provide input to a layer holding episode representations learned over a longer timespan, the **candidate episodes (c-ep)** layer. This layer is a self-organising map (**SOM**): when exposed to training episodes, it learns to represent episodes as localist units, organised so that similar episodes are close together in the map. Each unit can encode a particular combination of representations in the agent, patient and action media, and thus can represent a complete episode by itself. Note this localist scheme is enabled by our model of binding: the ‘agent’ and ‘patient’ fields of a WM episode index their fillers *by place*, so carry information about both roles and fillers. Clearly, we cannot represent every possible episode using localist units. But that is not the purpose of the c-ep SOM: its role is rather to represent the

episodes that occur frequently, so these can provide a top-down bias on SM processing during experience. Since the c-ep SOM uses localist representations, it can also represent *multiple* expected episodes simultaneously: a useful property, as we will show.

A final component of the network is a layer representing the **current situation**. In our model, this is the hidden layer of a recurrent network that learns to predict the *next episode*, given the episode that has just occurred, plus a copy of its hidden layer at the previous time point. The current situation network learns to predict a *distribution* of possible next episodes in the c-ep SOM (exploiting its ability to represent multiple episodes). (It is somewhat analogous to Reynolds *et al.*'s (2007) recurrent network for event representation, but Reynolds *et al.*'s network predicts the next *component* of an episode, rather than the next discrete episode.)

One useful feature of our model is that the c-ep SOM can learn *generalisations* over episodes. One kind of generalisation is hard coded in the model: the copies of WM individuals created in the agent and patient fields ignore location information, so representations of episodes in the SOM abstract over the location of participants. In our model, expectations about the locations of objects are dealt with in the WM individuals system, as we will illustrate below. This step considerably reduces the combinatorial possibilities that need to be represented in the SOM. But the SOM also learns generalisations of its own. The ability to generalise is a standard feature of learning in SOMs, since episodes that are sufficiently similar will activate the same localist unit. In particular, since the representations of agents and patients providing input to the SOM are distributed, the SOM can learn to abstract away from the properties of token individuals and represent the participants of episodes as types, as we will show.

## 9.3 Properties of units in each medium of the circuit

In this section I'll discuss the sorts of properties we expect neurons in each medium to have.

### 9.3.1 The LTM/WM individuals system

In the system representing individuals, we have three simple WM media that hold object type/RPC, object location, and cardinality. Then we have two convergence zones holding associations between these: a WM individuals medium and a LTM individuals medium. The patterns of activity in all these media change *every time a new individual is attended to*. There are lots of cells like this in PFC: for instance, see Lebedev *et al.* (2004) for PFC cells that only encode the most recent object in a task requiring memory of two objects. (The fact that multiple objects can still be retained bespeaks an 'activity-silent' way of storing objects, in short-term connections; see Stokes *et al.*, 2015.)

An important idea in our model is that the LTM individuals system can represent the agent himself, or an externally perceived individual. We therefore expect to see some components of the LTM individuals system that are specialised in representing 'the self', while others must specialise in representing 'others'.

### 9.3.2 The current WM episode: agent, patient and action media

In the system representing the 'current WM episode', we have three WM media: an agent medium, a patient medium, and an action medium. Thinking just about the agent and patient media: these can be distinguished from media in the WM individuals system, in various ways. Firstly, agent cells respond to the first individual observed, and patient cells to the second individual. Cells of both kinds have certainly been found: for instance, in Warden and Miller's (2007) study of macaque dorsal PFC, during a task when two objects were presented in sequence, and both needed to be retained in WM, 13% of cells were sensitive to the identity only of the first object, and 28% of cells were sensitive to the identity only of the second object. (I'll suggest below that dorsal PFC is not the best place to look for agent and patient cells, and that dorsal PFC cells preferentially represent whole episodes, but this is already something.) Secondly, agent cells should retain their

activity when a second (patient) object is attended to: this makes them unlike the WM individual cells discussed above. I think the cells in Warden and Miller’s study that responded just to the first stimulus represented this stimulus in sustained activity. I think the ‘memory cells’ in Lebedev *et al.*’s (2004) study had that property. (Again, these were recorded from dorsolateral PFC, and I think there are better areas where agent cells might be found.) Thirdly, both agent and patient cells have the property that they *stop* encoding individuals after two individuals have been attended to (in the current task), because that saturates the capacity of the system. There’s evidence of that in Buschman *et al.* (2011).

Somewhere in here: we expect to see specialised representations of ‘the self *as agent*’ and ‘the self *as patient*’—and similarly, specialised representations of others as agents and patients.

### 9.3.3 The candidate episodes buffer

Then we have a convergence zone holding associations between agents/patients and actions: the candidate episodes medium. Here, units represent whole episodes, combining agents and patients (and objects). In our model, the cells in this medium start to respond as soon as the first object is presented (in the agent medium), and the pattern of activity changes when the second object is presented. Individual cells here are very much like the cells discussed in Warden and Miller (2007). Warden and Miller found

### 9.3.4 The situations SOM

Then we have a situations SOM. I suggest that this should encode particular points during sequences of episodes: those are like the dlPFC cells found in

### 9.3.5 The scenarios medium

Then we have a plans/scenarios medium.

## 9.4 Evidence for the media in the hippocampus and medial temporal cortex

### 9.4.1 Preliminaries: models of the structure of hippocampus

I’ll be drawing on several theoretical proposals about internal structure within the hippocampus.

1. Rolls (2010 [cited in Bonnici *et al.*, 2012]) summarises the long-standing idea that ‘pattern separation’ happens in the dentate gyrus and CA3, while ‘pattern completion’ happens in CA3. What is ‘pattern separation’? In our model, it corresponds to the localist encoding that comes for free in a SOM: picking the ‘winning unit’ in a SOM is like creating a ‘well separated pattern’. SOM units have a crucial role in pattern completion, in that they hold the disparate elements of the pattern, as in a convergence zone. I want to argue our SOM model is a high-level picture of what’s happening in dentate gyrus and CA3.

2. Preston and Eichenbaum: posterior hippocampus holds representations of episodes by themselves; anterior hippocampus holds representations of episodes-in-contexts.

3. Zeidman and Maguire (2016): anterior hippocampus is needed to generate detailed representations of scenes or situations.

4. Poppenk *et al.* (2013): ventral/anterior hippocampus holding representations of ‘gist’, while dorsal/posterior hippocampus holds representations of ‘detail’.

### 9.4.2 The LTM/WM individuals system

The focus in hippocampus is on LTM individuals, which ‘enter WM’ simply by becoming active. (...)

Perirhinal cortex and lateral entorhinal cortex hold representations of token individuals. (See references in Preston and Eichenbaum, 2013.) The main result is that you need perirhinal/lateral entorhinal cortex to ‘recognise’ a familiar individual. (Recognition can happen even if you can’t reconstruct where you met that individual, or any episodes involving the individual: that additional information is computed in the hippocampus.)

Parahippocampal and medial entorhinal cortex hold representations of token LTM environments. (This is where the parahippocampal place area is.) This is discussed later, in Section ??.

Ventral/anterior hippocampus holds representations of ‘situations’, which can stand in for times (Preston and Eichenbaum, 2013; Poppenk *et al.*, 2013). (These situation representations will be discussed in Section 9.4.5.)

Hippocampal units could hold associations between individuals, locations and situations: this would implement a memory for what objects were where at what situations/times.

### 9.4.3 The current WM episode: agent, patient and action media

Representations of ‘the current agent’ are held in medial temporal cortex (Hassabis *et al.*, 2013).

There are agent- and patient-specific areas in mid-superior temporal cortex (Frankland and Greene, 2015).

WM representations of the current action near medial temporal cortex? Still to find..

### 9.4.4 The candidate episodes buffer

There’s heaps of evidence the hippocampus stores all manner of relations: see for instance Konkel and Cohen (2009), who review evidence the hippocampus stores relations that can be sequential (B came after A), spatial (B is to the right of A) and associative (A and B co-occurred). Eichenbaum *et al.* (1999) review evidence that the hippocampus can hold representations of arbitrary behavioural actions, if these are relevant for the task.

The tricky thing is to decide whether the relations that are stored correspond to whole episodes. For instance, an object stimulus could just be encoded as an object, but for all we know, ‘the appearance of an object’ could be considered to be an episode. Also, since in our model, the objects in an episode are attended to sequentially, certain sequences of objects (those that participate in an episode) should be represented in single declarative representations, that don’t involve updating. To think about some more!

Localist representations of ‘the current episode’, and of alternative possible episodes: dorsal/posterior hippocampus (Preston and Eichenbaum, 2013; Poppenk *et al.*, 2013). [Has handily strong links to the perirhinal cortex, where object representations are held; see Preston and Eichenbaum, 2013.]

For the idea that recollection of an episode from LTM involves a decision between competing alternative episodes: see Redish and Mizumori (2015). For evidence the hippocampus holds episodes that compete during recollection from LTM: see Wimber *et al.* (2015). I’ll assume the area where episodes compete is the same area that holds localist representations of episodes, because localist representations can easily be made to compete. So again, the candidate episodes SOM is in dorsal/posterior hippocampus.

### 9.4.5 The situations SOM

Evidence that the hippocampus is involved in storing sequences of sentence-like episodes, over a delay of 24 hours: Adlam *et al.* (2005). (This evidence comes from dysfunction: both developmental and adult-onset hippocampal amnesia.)

Representations of episodes-in-contexts (the situation SOM): ventral/anterior hippocampus (Preston and Eichenbaum, 2013; Poppenk *et al.*, 2013). [See Poppenk *et al.* for references to ventral/anterior hippocampus holding representations of ‘gist’, while dorsal/posterior hippocampus holds representations of ‘detail’.] Zeidman and Maguire (2016) review evidence that the anterior hippocampus is needed to generate detailed representations of scenes or situations. If this activity

involves generating a rich distribution over possible actions, it would call not just on the WM episode buffer, but also on the situation SOM.

The context representations here are sensitive to individual times, time types, previous episodes, tasks, and spatial environments. Some units in the situation SOM have associations with specific token times (and thus represent particular moments); others generalise over times (and thus represent generic schemas). The hippocampus is biased towards the token representations.

Interesting evidence about how the hippocampus represents items ‘in context’ comes from an experiment by Horner *et al.* (2012). They asked subjects with variable amounts of hippocampal damage to remember items (words) in contexts (photos of spatial scenes), and later tested them both on item memory (did you see this word?) and item-context memory (which scene did this word appear in?), while monitoring brain activity with MEG. Subjects’ item memory performance did not correlate with hippocampal damage, but item-context performance did correlate (negatively) with hippocampal damage, suggesting that only item-context memory depends on the hippocampus. However, MEG analysis revealed that items are *represented* individually in hippocampus, early in item-context trials: a ‘hippocampus-dependent’ signal representing just the test item emerged around 350ms after its perceptual onset. This signal quickly *merged* into a representation of the item-context association, suggesting that the hippocampal representation of the item functioned to retrieve the context representation by a pattern-completion mechanism.

Elsner and Hommel (2002) found interesting evidence that the hippocampus is involved in storing learned perceptual consequences of motor actions in humans. In a learning phase, subjects produced actions (button-presses with fingers) associated with unique perceptual consequences (auditory tones). In testing phases, they heard a ‘learned’ auditory tone associated with one of the motor actions, mixed in varying proportions with a ‘neutral’ tone not associated with any action. The proportion of learned to neutral tones was found to correlate with PET activity in the hippocampus.<sup>3</sup> This suggests the hippocampus develops special representations of states that are reliably brought about by motor actions. (There is evidence for such representations in rats too; see Corbit and Balleine, 2000.) In our model, these representations are readily interpreted as states of the situation SOM, after it has been updated by a motor action: the updated pattern in the situation SOM is ideally placed to encode the perceptual consequences of actions.

#### 9.4.6 Orthogonally from all this: a mixture of specific and general representations

The hippocampus holds generic individuals and episodes as well as specific ones. (I think: still have to find evidence.)

Some comments about the sparseness of representations in the hippocampus and surrounding regions are also in order here. Some single-neuron studies on humans have found evidence for neurons with highly selective responses to objects, faces or landmarks (see especially Quiroga *et al.*, 2005; Wixted *et al.*, 2014). Other studies find much less selectivity: for instance, Valdez *et al.* (2015) find that on average, neurons in the hippocampus (and amygdala) respond selectively to 25% of all objects presented, without any indication of generalisation. My point here is that a cell that responds to several distinct objects might still be a localist representation. For one thing, it might be a localist unit that represents a type—possibly a very broad type. For another thing, it might be a localist unit that represents something larger than a single object. It could represent an episode involving two different objects: for instance, a person and a landmark. (In this case, it would respond to both a person and a landmark not because it participates in distributed representations of both people and landmarks, but because a person and a landmark are distinct components of the complex semantic entity it represents.) Or it could represent other kinds of complex semantic entity: for instance, an object at a particular time, or an object in a particular situation. (In this case, it might respond to multiple objects in virtue of the current situation, rather than because of anything to do with the object.)<sup>4</sup>

---

<sup>3</sup>And also activity in a frontal area, the supplementary motor area, which we will discuss in Section 9.5.5.4.

<sup>4</sup>If we built this system, I guess we’d have to look at how many SOM units in the relevant media actually do

## 9.5 Evidence for the media in PFC

### 9.5.1 Preliminaries: models of the internal structure of PFC

I'll be drawing on several theoretical proposals about internal structure within the PFC.

1. Badre and D'Esposito: their idea about an organisation into representations of different 'orders': prepared SM operations, first-order rules, second-order rules.
2. O'Reilly: what and where.
3. One about the presence of both specific and general representations, both for individuals and events.

### 9.5.2 The LTM/WM individuals system

Specific and general—but maybe with a focus on general. (See e.g. Gotts *et al.*, 2015.)

Many of these representations are not sustained actively over saccades: see e.g. Stokes (2015); Lebedev *et al.*, 2004). In our model, that is definitional of the WM individuals system. (Whereas representations in the 'agent' and 'patient' fields of a WM episode have to be sustained across saccades.)

Representations of object type, and token objects: ventral PFC (see e.g. Courtney *et al.*, 1996; O'Reilly, 2010; ). Representations of object location, in a range of coordinate systems: dorsal PFC (see e.g. Courtney *et al.*, 1996; O'Reilly, 2010; ).

- Representations of combinations of object type/identity and location:
- Representations of singularity/plurality?

### 9.5.3 The current WM episode: agent, patient and action media

#### 9.5.3.1 Agent

Representations of the agent in medial PFC (Hassabis *et al.*, 2013) including the self as agent (Uddin *et al.*, 2007) with self preferentially represented in ventromedial PFC and others in dorsomedial PFC (see Wagner *et al.*, 2012 for a review).

Medial PFC also represents animate objects; see Martin and Weisberg, 2003.

Medial PFC also represents object identity and object location; see e.g. Chao *et al.*, 2016.

#### 9.5.3.2 Patient

Patient in premotor (and parietal) areas holding affordance/use-based representations of objects and tools (see e.g. Yee *et al.*, 2010; Binkofski and Buxbaum, 2013; Buxbaum and Kalénine, 2010).

#### 9.5.3.3 Prepared action

Prepared action is represented in dorsolateral and ventrolateral PFC and dorsal premotor cortex (Yamagata *et al.*, 2012). Importantly for us, the prepared action signal remains tonically active during both action preparation and action execution (Yamagata *et al.*, 2012). Some components of this premotor signal encode actions at an abstract level that generalises between hands (Gallivan *et al.*, 2013) and between self-generated and externally-generated actions (Ariani *et al.*, 2015).

### 9.5.4 The candidate episodes buffer

'Localist' representations of alternative possible episodes (the candidate episodes buffer): I suggest ventromedial PFC, which computes the expected rewards associated with alternative actions (Domenech and Koechlin, 2015), and dorsolateral PFC, which maps perceptual stimuli onto expected episodes. Dorsolateral PFC contains representations of prepared SM routines (Averbeck *et al.*, 2002). The dorsolateral and ventrolateral representations are the ones that if selected will

---

become active.

actually drive behaviour: so there’s an emphasis on posterior PFC areas (Badre and D’Esposito, 2009).

## 9.5.5 The situations SOM

### 9.5.5.1 Contexts

Representations of contexts, i.e. current states: medial PFC (Navawongse and Eichenbaum, 2013; Preston and Eichenbaum, 2013).

### 9.5.5.2 Prepared SM sequences

Barone and Joseph (1989) is a good reference here.

### 9.5.5.3 First-order rules

A ‘current state’ is like a currently active pattern in the situation SOM: it represents a transition between episodes, rather than an episode. For instance, it could represent a first-order rule (‘if S1, then R1’). On that basis, the PFC representations encoding these are slightly more anterior than those encoding specific episodes (Badre and D’Esposito, 2009).

### 9.5.5.4 Perceptual consequences of actions

Refer back to the Elsner and Hommel (2002) paper described in Section 9.4.5, maybe?

## 9.5.6 The scenarios medium

Representations of static high-level ‘plans’ or ‘tasks’: the anterior parts of frontal cortex (see especially Badre and D’Esposito, 2009). (So situations and episodes are more posterior.) Badre *et al.* (2010) have nice fMRI evidence in humans that ‘second-order’ rules are represented in anterior lateral PFC, while ‘first-order’ rules are represented in more posterior PFC. Second-order tasks would correspond to scenario units; first order tasks would correspond to units in the situation SOM. Stokes *et al.* (2013) also have nice evidence for static high-level plans. They record a population of lateral PFC cells, while monkeys are doing two tasks that require different stimulus-response rules. After a cue indicating the task is presented, they find activity settles down into a low-energy stable state, that is orthogonal to the states associated with both stimuli and responses. These states apparently bias the *dynamics* of PFC to map from stimuli to appropriate responses. Different tasks are represented by different dynamics. This is exactly the role of plan/scenario units in our model. (E.g. Marco’s units.)

## 9.5.7 Orthogonally from all this: a mixture of specific and general representations

The context representations are also a mixture of specific (i.e. occurring at a token time) and general (i.e. of use in behaviour, in multiple situations). Evidence for specific context representations comes from studies showing medial PFC holds representations of ‘token episodes’ which in our model are really token contexts, from which token episodes can be clearly *selected*. Evidence for general context representations include any ‘action rules’, that apply generally in a range of situations, and are thus generic. See again Badre *et al.* (2010) for examples of th. (Note that ‘action rules’ in our system are situation SOM units, that map a perceived episode onto a distribution of expected next episodes.)

## Chapter 10

# Memory representations of individuals, episodes, situations and plans: a detailed model

Section 10.1 introduces the complete architecture of the system for representing individuals, episodes, situations and plans. It's basically two copies of the NN architecture shown in Figure 9.1, with various links connecting them together. The interactions between the two systems give the system as a whole several useful abilities: these are discussed separately in Sections 10.2–???. In each section, we present the ability, giving some examples, and also discuss empirical evidence for the circuit in question.

### 10.1 Architecture of the complete model, and empirical motivation

#### 10.1.1 Architecture

Larger diagram in here.

#### 10.1.2 Empirical support for the architecture

Here, mention the connectivity between the relevant brain regions.

You should talk about the role of the nucleus reuniens in linking hippocampus with medial PFC, and the relevance of this for working memory tasks (see e.g. Griffin, 2015).

### 10.2 Application 1: Consolidation of hippocampal memories

The combined systems can do consolidation of episodic memories in cortex (specifically, in PFC). To do this, I guess plasticity in PFC has to be set very low during waking SM experience, while plasticity in hippocampus is set high, to enable fast learning—but using nicely separated, orthogonal representations of episodes and situations. Then at night, there's a process of pseudorehearsal, where episodes from hippocampus are played to PFC, in an offline mode, interleaved with training data generated by PFC.

### 10.2.1 Evidence for the roles of hippocampus and PFC in consolidation

Describe the standard model (Marr, O'Reilly, people like that).

There are various problems for the standard model, in particular the fact that hippocampus does seem to hold remote memories too. These will be addressed as they arise.

### 10.2.2 How hippocampus encodes token episodes

I suggest that the hippocampus represents each token episode with exactly one unit in the candidate episodes SOM, and each token situation with exactly one unit in the situations SOM. The learning constant for updating weights in the SOMs can be very high, and the Gaussian radius for influencing neighbouring SOM units can be set to very low, so only a single SOM unit is affected. This ensures that the episode, and its associated situation, are represented strongly in hippocampal LTM, and also very 'orthogonally' from all other episodes and situations.

Of course, this policy for selecting SOM units means we will soon run out of SOM units: it needs to go hand-in-hand with a policy about how to *choose* SOM units to represent new incoming episodes/situations. This selection policy is described in Section 10.2.4.

### 10.2.3 Principles for selecting hippocampal material to replay

A big question concerns *which* episodes are replayed from PFC. I can think of a few principles governing this.

#### 10.2.3.1 Sequential structure

One principle is that you don't just replay single episodes: you replay little nuggets of sequential structure. (Might be worth referring to the DeepMind way of doing offline learning here: small nuggets help to break the self-correlations that normally get in the way of reinforcement learning.)

#### 10.2.3.2 Emotional valence

Another principle is that episodes with high emotional valence are replayed preferentially. I suggest this is implemented by having the SOM units that store episodes/situations with high emotional valence learn associations with higher learning rates, so their connections are stronger when they are first learned. See below for an idea about how this can result in their being replayed more often.

### 10.2.4 Implementing a buffer of recent episodes

Another big question relates to how the hippocampus implements a buffer that tends to hold the recent episodes. In the hippocampus, new episodes must ultimately overwrite old ones. Here are a couple of principles which could implement this process.

In Section 10.2.2 I introduced the idea that the hippocampus represents each token episode with exactly one unit in the candidate episodes SOM, and each token situation with exactly one unit in the situations SOM. This ensures that the episode, and its associated situation, are represented nice and 'separately'/'orthogonally' from all other episodes and situations. We first need a principle that selects which SOM units will be selected for an incoming episode/situation. I suggest there's a mechanism that selects the candidate episodes SOM unit with the *weakest connections* into the WM episode buffer, and the situation SOM unit with *the weakest connections* into the WM episode buffer and recurrent context. These processes select units to represent the current incoming episode.

Secondly, assume an offline consolidation mechanism that picks at random from amongst the units in the situations SOM, with a bias towards situations with high emotional valence, and then plays forward a little sequence of episodes from the selected situation. The critical thing is that *each time a SOM unit representing a situation or episode is replayed to PFC, its connections are*

*incrementally weakened*, so it progressively becomes more likely to be the unit chosen to represent the next incoming situation or episode. Since episodes/situations with high emotional valence have higher weights to begin with, they are replayed more times before they fade completely from hippocampus—which gives them an opportunity to be represented as tokens in PFC. (See below.)

### 10.2.5 Learning token episodes/types, and generalisations, in PFC

I'm not going to assume that there are units representing token times, as there are in the hippocampus. I'll assume we just have SOMs, with lots of units. The situation SOM units connect to the elements of 'the triggering WM episode', but also to representations of spatial context, and also to other rich contextual content. (I need to think about what these might be.) The candidate episodes SOM units connect to the elements of the associated episode, but also to rich content in primary SM areas: smells, sounds, colours.

I suggest there are two principles at play here.

Firstly, there's a principle that if you can predict the next episode well, because it conforms to a general rule, you don't need to store it as a token: you can rely on the rule to complete it for you. (This implements a 'constructivist' model of LTM recall.) My guess is that it's still stored as a token episode in the hippocampus, but that during consolidation into PFC, you lose the hippocampal representation, but just make small changes to a general rule in PFC—so the memory of the episode as a token is altogether lost.

Secondly, there's a principle that if the episode occurs *very frequently* the training inputs, the PFC-based situation and candidate episodes SOMs are likely to store it as a token.

Those two principles should fall out of normal SOM learning quite well, I think.

#### 10.2.5.1 Remembering what I was wearing: the role of rehearsal in the WM individuals system

In frequently-revisited emotionally salient memories, you can often remember a lot about the *state* of the participants involved: what people looked like, what they were wearing, and so on.<sup>1</sup> I guess that in order for these memories to survive, we have to envisage that rehearsal of *episodes* in the hippocampus also leads to rehearsal of *WM individuals*. Say the hippocampus also allocates brand-new orthogonalised units for each individual that's encountered—I mean, each *stage* of each individual. That allocation uses the same principles as the allocation for episodes and situations. That means an individual that happens to participate in an emotionally salient episode will be frequently reconsolidated, and its transient state at the time it participated will be well recorded.

Now imagine the copy operation that transfers the properties of a WM individual into the agent or patient field of the WM episode. If there's a stage of an individual that is very frequently transferred, through its role in a frequently reconsolidated episode, *that stage will also get into prefrontal LTM*.

To make this work, we have to assume that the links between the WM individual medium and the agent and patient media are not straightforwardly 1:1 copies: rather they are SOM-like structures. And they don't *just* link to WM-individual media: they also link straight to first-order perceptual representations in SM cortices.

I really like the story this tells about how stages of individuals get represented in LTM. I have a concept of 'the toddler me', 'the boy me', 'the high-school me', 'the student me', 'the young Mozart', and so on: the episodes I choose to rehearse discretise the spatially extended person into a number of distinct snapshots.

### 10.2.6 Remote and generic memories in the hippocampus?

On the face of it, the above account of consolidation has a problem with the finding that very remote memories can still be found in the hippocampus: see e.g. Bonnici *et al.* (2012). Also with

---

<sup>1</sup>In fact, I think this is the place to plug a much more general account of stative properties of LTM individuals. States obtain at particular times. It's also the place to plug a general account of LTM for object locations.

the fact that the hippocampus contains general memories (see e.g. ).

I suggest these remote memories can be accounted for by the fact that the agent, during waking experience, sometimes *remembers* old token memories. When you remember something, I assume it's because of a recall process *distributed over hippocampus and PFC*, drawing seamlessly on a mixture of consolidated and non-consolidated material. Again, you tend to be reminded of emotionally salient episodes: those are the ones that are most strongly represented in hippocampus, and the only sort that are likely to survive as tokens in PFC. My key suggestion is that *when you remember something, you re-install it as an episode/situation in hippocampus*. Consider: the remembering happens 'during waking experience'—and everything in this mode is allocated a new pair of units in the situation SOM and candidate episodes SOM. So now we can envisage a goodly amount of reverberation, where emotionally significant memories bounce backwards and forwards between hippocampus and PFC. I will discuss this more in Section 10.5.3.4.

The above process of re-encoding prefrontal memories in the hippocampus also allows for generic facts to be stored in the hippocampus. On the above model, whenever an agent entertains a generic episode, whether just thinking, or talking, the hippocampus will newly encode this episode, in a set of newly-minted SOM units. We will talk more about entertaining generic propositions in Chapter 21.

### 10.3 Application 2: Learning of optimal SM behaviour

This section covers how an agent can learn to behave optimally in the SM here-and-now. That includes learning how to act, but also learning how to perceive. The mechanisms involved should include reinforcement learning (for learning of motor actions), but also learning of common sequences of episodes (for learning the SM/anticipatory skills needed to optimally sample the environment).

Note that some skill learning happens during sleep, so you wake up doing better at the task: evidence that consolidation is involved in that too.

### 10.4 Application 3: Querying of episodic memories

It can do querying of episodic memories (in both hippocampus and PFC) by language, by representing a query episode first in the WM episode in PFC, then sending it to the WM episode in hippocampus, where a complete episode will be retrieved.

In the model I'm thinking of, there are two media where queries can be expressed, and where responses to queries can be activated. I assume that both queries are executed simultaneously. Looking ahead, I want to make use of the existence of two WM episode/query media in an account of information structure: if one of these media (say the PFC one) retains the query unaltered, while the other (say the hippocampus one) expresses potential responses to the query, then we can easily check (a) the response is indeed a response to the query; and if so, (b) what the 'new' part of the response is. However, there are some potential problems: in particular, people without the hippocampus (like HM) can still answer questions.

However, HM's language is not completely normal, and some of the dysfunctions relate to prosody, which was 'monotone' (see e.g. Mackay *et al.*, 1998), which perhaps indicates problems with information structure. And in more recent research, there has been good evidence that patients with hippocampal damage are impaired at certain WM tasks—in particular those that require the matching of representations containing complex bindings; see Yonelinas (2013) for a review. (If the hippocampus holds a clearly-delineated representation of an episode active for a short period of time, as discussed in Section ??, it is natural to expect cognitive strategies to be able to exploit such representations.) In fact, there is evidence that patients with hippocampal damage are impaired in certain dialogue-level linguistic capabilities, in particular in establishing and making use of a representation of the 'common ground' (Duff and Brown-Schmidt, 2012), which is certainly important in question-answering.

The jury is still out on this one: the crucial experiments have not yet been conducted. However, we predict that patients with hippocampal damage are impaired (a) in signalling information structure in their answers to questions, and (b) when questions are complex, in generating answers that actually match the question. That’s certainly the behaviour we should get in our model if we disable the hippocampal WM-episode medium.

### 10.4.1 Post-retrieval processing

The model here should include ‘post-retrieval processing’, whereby the query episode held in PFC is compared to the result episode in hippocampus, to see that they match.

Evidence for post-retrieval processing: see for instance the review in Ranganath and Knight (2003). They discuss the proposal that activation of (ventrolateral) PFC during episodic memory tasks reflects ‘selection and maintenance of relevant attributes of study items and test cues’ (Ranganath and Paller, 1999a; 1999b; 2000). Meanwhile, dorsolateral PFC acts ‘to monitor and manipulate’ representations retrieved from episodic memory (D’Esposito *et al.*, 2000; Petrides, 1996).

## 10.5 Application 4: The interface with language

You’re just giving the big picture here. More information about NL syntax, and more SM interpretations of syntax, is presented right through the remainder of the document.

### 10.5.1 A SM interpretation of syntactic structure

This section summarises ideas from Knott (2012). Here, you can talk about XP structure.. and then about head-raising, DP raising and so on.

### 10.5.2 A model of sentence generation

Describe: (i) How the clause gets read out, and where the parameters are; (ii) how pointers into the WM system allow for nested sequential routines.

### 10.5.3 Elements of a model of sentence semantics

#### 10.5.3.1 The semantics of generic sentences

Natural language semanticists have no good idea how to model generics (see e.g. Carlson and Pelletier, ). But the semantics of generic comes for free in the current model. That includes generic sentences, but also generic objects (i.e. object types). That’s because we’re doing it all right!

#### 10.5.3.2 The semantics of conditionals

Another element of semantics that’s well modelled is conditionals. These express generalisations—again, often generic in flavour, rather than absolute.

#### 10.5.3.3 Information structure in sentences

Having got the link to syntax in Section 10.5, and the ideas about episodic memory querying in Section 10.4, you can now give a model of how it does given-new information structure, by virtue of that comparison operation. It can use a representation of ‘the current context’ to control behaviour. (This mainly uses the PFC representation of the current situation—one that abstracts somewhat over token times (since ‘now’ is a brand new token time).) Using similar mechanisms, it can constructively re

#### **10.5.3.4 Temporal subordinate clauses**

This covers the ideas in the LTM CogSci paper.

## Chapter 11

# WM representations of spatial and stative propositions

### 11.1 WM representations of spatial propositions

### 11.2 WM representations of predicative propositions

In here, there should be a model of how the WM episode can be extended to model predicative sentences. The main idea is along the lines of your paper at NZLing about predication, supplemented with

## Chapter 12

# LTM for spatial and stative information

In Chapter 3 I described mechanisms that compute spatial representations of environments and spatial locations of objects in environments. In Chapter 5 I described mechanisms that compute spatial representations of 3D object geometry, and of locations within objects (see especially Section 5.7). In those earlier chapters, I discussed how these representations were computed ‘online’, during SM experience. In this chapter, we will describe circuits that store these representations and relationships in LTM.

The basic thrust of this chapter is that these circuits are stored in the same hippocampal area that holds representations of episodes, and their relationships with situations. [I think this needs to be revised.]

### 12.1 LTM for object locations

Key idea here: there’s a SOM whose units hold associations between LTM individuals, LTM environments, situations, and places. The places are specified in an environment-centred frame of reference, so they endure over movements of the agent. (Vision delivers information about object place in retinotopic coordinates; to convert this information to environment-centred information, we can use the orienting SOM described in Section 3.4.3.)

Of course, a place only means something in the presence of an active LTM environment, as discussed just above in Section ??, and objects can occupy different places at different times: hence the need to encode these complex associations.

The SOM is called the **object locations SOM**. Its inputs are shown in Figure 12.1.

#### 12.1.1 Object location memory and trajectory planning

In this section I will briefly return to the topic of navigational planning, which was the subject of Section 3.3. Recall that in Section 3.4.5 we introduced a system that learns a set of ‘goal places’ for a given environment. As described in Section 3.4.5.2, a distribution of activity over the set of goal place units allows the agent to select a trajectory that leads towards a reward state. Often, objects themselves are associated with rewards, rather than locations.<sup>1</sup> Now that we have a model of object location memory, we will discuss how this memory, together with learned associations between objects and rewards, can induce a distribution over goal places.

Say the agent has a mechanism for associating objects with rewards, as well as locations. This circuit would map LTM individuals onto a standard ‘reward’ unit, with the strength of the

---

<sup>1</sup>We have not introduced this idea explicitly: but in Chapters 9 and 10 we motivated the idea that in certain situations, the observer has a bias towards certain objects or object types. This idea will be linked to concepts of reward in Chapter 13.

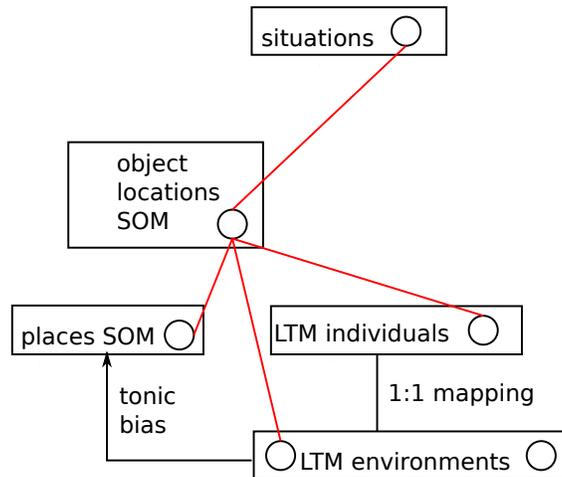


Figure 12.1: The object location memory SOM

connection reflecting the amount of reward, and would map LTM individuals inhibitably onto a standard ‘punishment’ unit, with the strength of inhibition reflecting the amount of associated punishment. A similar circuit can be envisaged for object types, more generally. (We will discuss these systems in more detail in Section ??.) And say the agent also has a mechanism for storing the locations of objects in LTM, of the kind discussed in the current section. When the agent enters some new environment, we would like to activate a set of candidate navigation goals based on the known or expected locations of objects in this environment. This would supply a useful ‘top-down’, or ‘semantic’, component to the agent’s representation of alternative possible navigation goals. In this section I’ll describe a circuit that achieves this. I’ll refer to the LTM individuals system, rather the object types system, but the principle can be applied to both. The circuit involves three steps.

The first step is to activate a distribution over the set of LTM individuals that reflects their associated reward or punishment value. This can easily be done, simply by activating the ‘reward’ and ‘punishment’ units at fixed values: the learned associations of LTM individuals with these units will then induce a pattern of activity over LTM individuals, whereby individuals will be activated in measure of their association with reward, and inhibited in measure of their association with punishment. (I will assume a non-zero baseline of neutral activation for LTM individuals.)

When this is in place, the second step is to activate the object location memory system in parallel. We activate the current LTM environment unit, and the current situation SOM unit, and the distribution over LTM individuals that reflects their reward value. The object location memory system will now activate a distribution of places in the second-order places SOM. (That is, the SOM representing the locations of arbitrary objects.)

The final step is simply to follow the associations between units in the second-order places SOM and goal places. As described in Section 3.4.5, each goal place is associated with a unit in the places SOM. So following these 1:1 associations will induce a pattern of activity in the goal places medium. This will allow the agent to select a trajectory in the current environment, taking into account three fully separable components: firstly his knowledge of the spatial structure of the environment (encoded in ‘regular’ place cells, as biased by the LTM environment unit); secondly, his knowledge of which objects (or object types) that are likely to be in the environment, and where they are likely to be; and thirdly, his knowledge of the rewards (and punishments) associated with different objects (or object types). These three types of knowledge combine completely orthogonally, and can be applied in parallel, which makes for very efficient decision making.<sup>2</sup>

<sup>2</sup>Note, incidentally, that there is also opportunity for *serial* consideration of alternative options in each of the

## 12.2 LTM for object parts and possessions

I want to argue that propositions about one individual ‘having’ another individual are also expressed using the convergence zone that records LTM for object locations. The fields that need to be associated (a LTM environment, a LTM individual, a place and a situation/time) are exactly the same as those for object locations. The only difference is that when you say *John has a dog*, the ‘have’ verb expresses the operation of taking the attended individual (in this case John) and activating the associated environment (John-as-environment).

The identity between the two forms of proposition can be seen in syntactic alternations: for instance *The cake had a ribbon (on it)* alternates with *There was a ribbon on the cake*.

## 12.3 LTM for object properties

Separately from the above circuit for locations/possessions, there is a circuit that remembers the properties of LTM individuals (in different situations, times, and environments, as usual). I want to argue this circuit is completely distinct from the spatial locations circuit. (Even though it’s overlaid in the same patch of hippocampus.) Continuity of objects is ultimately a spatial (spatiotemporal) thing: if an object at a given location has one set of properties at one moment, and another set of properties at the next moment, we definitely want to say that the same spatiotemporally continuous object has changed its properties, and not that the original object has vanished and a different token object taken its place.

### 12.3.1 The dominant property assembly layer: representations of object types, and of properties

Refer back to Section 9.5.2, maybe? This is where you should be able to describe the rich property complex (RPC).

The system of object types is learned in our model by the dominant property assembly (DPA) circuit. The circuit is shown in Figure 12.2.

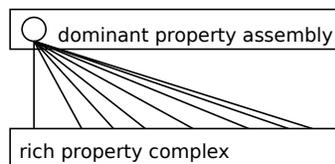


Figure 12.2: The rich property complex (RPC) and the dominant properties assembly (DPA) media. After a property assembly has been selected, property-level IOR can operate: this involves inhibiting the properties in the associated RPC with the selected property assembly.

Include the idea of property-level IOR (Chris and Lech). This is supposed to do properties, but also subtypes.

### 12.3.2 The property memory circuit

Convergence zone units associate LTM individuals property complexes, property assemblies, and situations. This allows individuals to have different properties at different times.

Explain how convergence zone units hold something like stages of LTM individuals.

Say there’s memory at the level of types as well, implementing the ‘categorical bias’ (see ).

---

above cases. The agent can serially consider different objects, perhaps in measure of their likelihood of being in the scene, or can serially scan locations in the environment, considering their likely contents, or finally, can serially evaluate whole trajectories, perhaps simulating each to see where it will lead. I think what we have here is a very interesting framework for implementing a mixture of Kahneman’s ‘fast’ and ‘slow’ thinking.

The SOM is called the **object locations SOM**. Its inputs are shown in Figure 12.3.

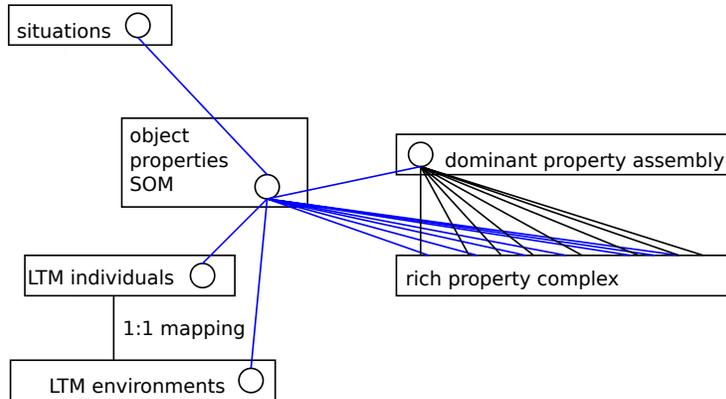


Figure 12.3: The object properties SOM

### 12.3.3 The location memory and property memory circuits combined, in an account of object recognition

Object recognition is a complex thing: you need to make reference to both properties (including types) and location. It is quite possible for two different objects to be indistinguishable in terms of their intrinsic features: for instance, two peas in a pod, or a pair of identical twins. Under these circumstances, the objects are individuated by their physical location. But in other circumstances, objects are identified mainly by their intrinsic properties. If I meet John on the street one day, I recognise him as John because of his intrinsic properties. The only circumstance in which I would have to infer he is a doppelganger of John, and not John himself, is if I know that John is somewhere else.

In the treatment I give here, when the observer establishes an object, generating an RPC and a location, there are two options: either the object is *recognised*, in which case, the object is represented by an existing LTM individual, or a new LTM individual is created to represent the object. Recognition firstly requires that the RPC of the perceived individual is sufficiently similar to that of an existing LTM individual to activate it as a candidate. Having selected a candidate LTM individual, the remembered location of this individual is retrieved in the object location memory circuit, and compared to the location of the currently attended individual, *in the current situation*. If these locations are not the same, the currently attended individual is a doppelganger, and a new LTM individual is created to represent it.

The system implementing the above process combines the circuits for LTM of object properties (Section 12.3.2) and object locations (Section 12.1). It is shown in Figure 12.4.

### 12.3.4 The continuity of LTM individuals

As discussed above, an LTM individual can have different locations at different times, and different properties at different times. What, then, is an LTM individual unit ‘in itself’? What allows it to persist, over changes to its location and properties?

I should say something about object files in here. And perhaps refer to the idea of a tracking map, for implementing object files.

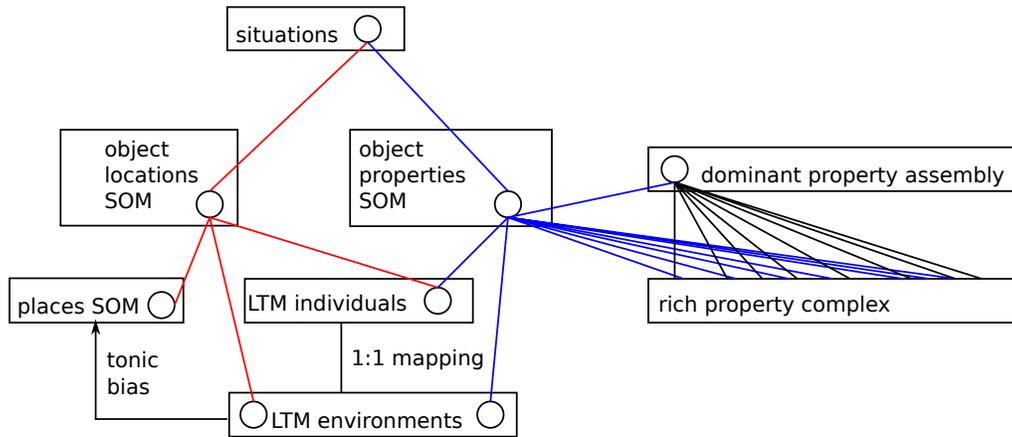


Figure 12.4: The object property memory circuit and object location memory circuits combined

## 12.4 Evidence for the media in the hippocampal region

### 12.4.1 Object location memory

We have already noted the presence of ‘view cells’ in both hippocampus and perirhinal cortex (see Section 3.4). Rolls *et al.* also found cells in these areas that encoded combinations of objects and locations.

Miller *et al.* (2013) show from single-cell recordings that cells in the human hippocampus encode associations between objects and locations, not just during perceptual experience, but also during recall.

### 12.4.2 How is the spatial representation system overlaid on the episodes/situations system?

We have to locate some parts of the spatial representation circuit (place cells, view cells and cells responding to combinations of LTM individuals and places) in the hippocampus, and in the perirhinal cortex. As usual, there is no part of the hippocampus that exclusively subserves spatial representation: however, there’s evidence that dorsal/posterior hippocampus is more specialised for encoding detailed representations of place. In rats there are more place cells in dorsal hippocampus (Jung *et al.*, 1994) and lesions to the dorsal, but not ventral hippocampus impair navigation performance on a water maze (Moser *et al.*, 1995); in humans, posterior (but not anterior) hippocampus is larger in London taxi drivers than controls, and in taxi drivers the size of posterior (but not anterior) hippocampus correlates with time spent in the job (Woollett and Maguire, 2000).

We have already located the candidate episodes SOM (mainly) in dorsal/posterior hippocampus, and the situations SOM (mainly) in ventral/anterior hippocampus. If the dorsal/posterior hippocampus also holds representations of places, and associations between individuals and places, we have a mixture of two populations of cells in this area: those holding holistic episode representations, and those subserving spatial cognition and LTM for object locations. We suggest there is a good reason for situating place cells and cells encoding object location in posterior hippocampus. Recall that posterior hippocampus is located in between parahippocampal cortex (which holds representations of LTM environments, see Section ??) and anterior cortex (which holds representations of situations, see Section 9.4.5). As discussed in Section ??, the SOM that encodes places needs to take input from tonically active environment representations; while the SOM that associates LTM objects with places must also take input from situations and environments.

## **12.5 Evidence for the media in PFC?**

Not sure what to put here yet. Probably there should be something about goal cells here.

## **12.6 Interactions between the spatial representation system and the episodes/situations system**

All these things should go for hippocampus and PFC.

### **12.6.1 Spatial components to situations**

'Situations' (used in the episodic memory system) can include LTM environment representations.

### **12.6.2 Modelling the effects of episodes on their participant individuals**

This is where we do axioms about how objects change as a function of the episodes they participate in.

## Chapter 13

# The reward system in the situation update network

### 13.1 The reward system for action execution

This is a straight implementation of TD learning within the situation update SOM, I think.

### 13.2 The reward system for action perception

I think this boils down to 'learning by surprise'.

### 13.3 Situation updates and plan updates

## Chapter 14

# Cross-modal representations of agents and patients

In this chapter I have to describe the mechanism that links the representation of the agent as physical object-in-the-world, object-of-a-certain-type, initiator-of-actions, possessor-of-plans, and possessor-of-a-body. And also the mechanism that links the representation of the patient as physical-object-in-the-world, object-of-a-certain-type, environment-for-hand-navigation-actions, and alterer-of-hand/arm-plant-dynamics. That's all a precondition for the account of DP-raising in the clause in Section 16.

## Part III

# Language representation mechanisms

## Chapter 15

# A model of clause syntax

## Chapter 16

# A SM interpretation of clause syntax

## Chapter 17

# A SM interpretation of stative sentences

### 17.1 Locative sentences and LTM for object location

In here I'll discuss sentences like *The cup is on the table*. But I will gloss over the semantics of spatial prepositions and the internal structure of PPs: I'll discuss those topics in Chapter 18. (Then I'll have a closer look at sentences containing spatial prepositions in Chapter 19.)

### 17.2 Predicative sentences and LTM for object types

Remember that a WM episode can be set up to represent the sequential process of doing property-level IOR on an object. In this case, since 'agent' (subject) and 'patient' (predicate) 'point' to the same individual, there is a choice about which to read out first, and languages have different conventions about this. (This is my model of 'predicate raising'.)

#### 17.2.1 A look ahead to DP-internal adjectives

Here, just mention that properties of objects can also be expressed within the structure of DPs. (A full story is given in Section 22.2, when we discuss the internal structure of DPs in more detail.)

### 17.3 Possession sentences

Here, discuss the structure of *have*-sentences. These read out material stored in the circuit for object parts/possessions.

### 17.4 Stimulus-experiencer sentences

Here, discuss the structure of stimulus-experiencer sentences. These read out material stored in the circuit for emotional relations.

## Chapter 18

# A SM interpretation of spatial PPs

### 18.1 Syntax of spatial PPs

### 18.2 SM interpretation of the syntax of a spatial PP

#### 18.2.1 WM spatial transitions

I think I need to introduce a third WM medium. We already have WM individuals and WM episodes: we now also need **WM spatial transitions**. I presume these are represented both in the hippocampal region and in PFC, like the other two media.

A key idea is that WM spatial transitions can be created from two separate sources: one relating to spatial representations of navigable environments in the hippocampus, the other relating to spatial representations of manipulable objects in parietal cortex. I assume both these systems compute the same types of spatial transition, and can communicate them to the WM spatial transition medium for WM storage. (This is intended to explain why prepositions like *on*, *in*, etc can describe relations both with objects and with environments.)

The other key idea is that an agent has to be able to *replay* spatial transitions from WM representations.. so they are also in some sense stored as prepared sequences of SM operations.

Evidence of the involvement of left parietal cortex (specifically, supramarginal gyrus) in processing spatial prepositions is reviewed in Struiksma *et al.* (2011). Evidence for the additional involvement of left inferior PFC is given in Tranel and Kemmerer (20). Wu and Chatterjee (2007) also found evidence that lesions in inferior parietal, occipito-parietal and inferior prefrontal cortex (all left-lateralised) were associated with deficits in processing spatial prepositions.

## Chapter 19

# A model of sentences containing spatial PPs

### 19.1 Stative spatial sentences

Here you can model the semantics of *The cup is on the table*. (Introduce the general principle for how to interpret a PP in context.)

### 19.2 Sentences denoting episodes involving movement

Here you can model the semantics of *John walked to the door*. And then, *John put the cup on the table*. (Though this latter one also needs some background on causative syntactic constructions.)

## Chapter 20

# Emotional attitudes towards objects: perception, LTM and language

### 20.1 Preliminaries: Damasio's model of emotions

### 20.2 A model of emotion perception and representation

[This is out of place: it's purely about the SM system.]

### 20.3 LTM representations of emotional relations

[This is out of place: it's purely about the LTM system.]

I also need to find a circuit for expressing propositions about emotions, like *John loves Mary* or *Mary hates biscuits*. Such propositions can also make reference to environments, situations and times. I think I'll propose that the convergence zone in this case is a hybrid, making use of parts of the circuit for episodes, but not generating situation updates, just as it is for predicative sentences (See the discussion of predication in Section 17.2 for a similar idea.) In the case of emotional propositions, I'll say it involves placing things in the 'agent' and 'patient' media (even though they're not normal agents and patients), and the emotion in the 'action' medium (even though it's not really an action). Perhaps the arbitrariness of placement allows the variation that is found between stimulus-experiencer forms (e.g. *Biscuits disgust Mary*, where the object of the emotion is the 'agent') and experiencer-stimulus forms (e.g. *Mary hates biscuits*, where the experiencer is the 'agent').

## Chapter 21

# Quantified propositions

### 21.1 Quantified propositions and the semantic LTM system

### 21.2 Reference to groups of objects perceived separately

This section is a propos of the discussion of object recognition in Section ???. You should be able to recognise two identical objects separately, but still be able to refer to them *later* as a group. I think that's permitted by the quantification system just described.

Say I look at one object and classify it as a dog, and then look elsewhere at another object, and classify that as a dog too. Even though these objects were established separately, I should be able to refer to them collectively, as a group. I already have ideas about how group objects are represented if they are *perceived* as a group: in this case, they are represented as an *LTM environment*, rather than an LTM individual. The RPC returned by the classifier is associated directly with this LTM environment. The LTM environment modulates the medium representing places, as it normally does. But in this case, places are somewhat abstract, but I assume there is a place associated with each individual. My guess is that there is also a parietal saliency map associated with the 'space', which allows for individuals to be attended to one by one, and which allows for the numerosity of individuals to be gauged. This parietal medium has a role in the representation of numbers: both ordinal numbers and numerosities.

If we attend to two identical individuals in a (physical) environment sequentially, we will create separate LTM individuals to represent them, as described in Section 12.3.3. However, we must still be able to form a representation of them as a homogeneous group. (That's demonstrated in the underlined expression in the following example: *I saw a dog. Later I saw another dog. The dogs both looked hungry.*) Therefore, there must be a way of representing the two individuals as an LTM environment—provided they have something in common.

I suggest the way this happens is through a query to LTM. I have a general model of quantified sentences which allows groups of objects to be retrieved. In that model, these groups are always associated with an abstract 'space', which supports sequential attention, and computation of numerosity. I now suggest the retrieved group is also routinely associated with an *LTM environment*—even if only one LTM individual is retrieved. The numerosity computation then identifies how many LTM individuals were retrieved, and the saliency map operations support sequential attention to the different individuals in the group. Crucially, the LTM group can now participate in LTM more generally, either as a (singular) group, or as (plural) individuals.

## Chapter 22

# The internal structure of DPs

You will already have said a little about the internal structure of DPs in Section 10.5.

### 22.1 The syntax of DPs

Cinque (2005) also has a good description of the high-level constituent orders that are attested cross-linguistically. He argues that there's a sequence of XPs hosting Demonstrative, Number, Adjective, and Noun, and that the noun (or actually NP) can raise to the specifier of any of the higher XPs, to give Dem, Num, A, N (with N unmoved); Dem, Num, N, A (with N moving one 'notch'); Dem, N, Num, A (N moving two notches); N, Dem, Num, A (N moving right to the top).

Overlaid on this, we need to consider the positions of multiple adjectives. For adjective ordering inside the DP, Panayidou (2013) gives a pretty good recent guide. The key references I'll use are Cinque (1994) and Cinque (2010), I think.

Cinque's idea is that Ns (or in 2010, NPs), sit at the bottom of a hierarchy of projections introducing (in turn) 'quality, size, shape, colour, provenance'.<sup>1</sup>

### 22.2 A SM interpretation of DP-internal adjectives

Somewhere in here, say that Kemmerer *et al.*, 2009 give evidence that a difficulty with adjective ordering is associated with damage to inferior parietal cortex, which fits in with the model I'll propose (at least as it relates to size and shape adjectives, that have their origin in parietal cortex in my model).

There are also some good neuro papers on adjectives: including Chang *et al.* (2009), Bemis and Pyllkkänen (2012; 2013), Fyshe (2015),

You might cite Belke (2006), but I'm not sure it's quite what the title suggests.

### 22.3 Partitive DPs

In this model, you have to refer to the idea that any object can be reinterpreted as an environment.

*A bowl of cherries, A line of soldiers*

### 22.4 The *kind-of* construction

This should

---

<sup>1</sup>An example of 'provenance' is nationality (e.g. *British*).

## Part IV

# Language processing mechanisms

## Chapter 23

# Sentence generation

## Chapter 24

# Sentence interpretation

# Bibliography

- Achanta, R., Estrada, F., Wills, P., and Süsstrunk, F. (2008). Salient region detection and segmentation. In *Proceedings of Computer Vision and Pattern Recognition*, pages 66–75.
- Achanta, R., Hemami, S., Estrada, F., and Süsstrunk, F. (2009). Frequency-tuned salient region detection. In *Proceedings of Computer Vision and Pattern Recognition*, pages 1597–1604.
- Adlam, A.-L., Vargha-Khadem, F., Mishkin, M., and de Haan, M. (2005). Deferred imitation of action sequences in developmental amnesia. *Journal of Cognitive Neuroscience*, **17**(2), 240–248.
- Almeida, J., Mahon, B., Zapater-Raberov, V., Dziuba, A., Cabaco, T., Marques, J., and Caramazza, A. (2013). Grasping with the eyes: The role of elongation in visual recognition of manipulable objects. *Cognitive, Affective, and Behavioral Neuroscience*, **14**(1), 319–335.
- Amit, E., Mehoudar, E., Trope, Y., and Yovel, G. (2012). Do object-category selective regions in the ventral visual stream represent perceived distance information? *Brain and Cognition*, **80**, 201–213.
- Antolík, J. and Bednar, J. (2011). Development of maps of simple and complex cells in the primary visual cortex. *Frontiers in Computational Neuroscience*, **5**, 17.
- Arbib, M., Bonaiuto, J., Jacobs, S., and Frey, S. (2009). Tool use and the distalization of the end-effector. *Psychological Research*, **73**, 441–462.
- Ariani, G., Wurm, M., and Lingnau, A. (2015). Decoding internally and externally driven movement plans. *Journal of Neuroscience*, **35**(42), 14160–14171.
- Arora, A., Weiss, B., Schurz, M., Aichhorn, M., Wieshofer, R., and Perner, J. (2015). Left inferior-parietal lobe activity in perspective tasks: identity statements. *Frontiers in Human Neuroscience*, **9**, 360.
- Averbeck, B., Chafee, M., Crowe, D., and Georgopoulos, A. (2002). Parallel processing of serial movements in prefrontal cortex. *PNAS*, **99**(20), 13172–13177.
- Badre, D. and D’Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Reviews Neuroscience*, **10**, 659–669.
- Badre, D., Kayser, A., and D’Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. *Neuron*, **66**, 315–326.
- Barone, P. and Joseph, J.-P. (1989). Prefrontal cortex and spatial sequencing in macaque monkey. *Experimental Brain Research*, **78**, 447–464.
- Bemis, D. and Pykkänen, L. (2012). Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cerebral Cortex*, **23**, 1859–1873.
- Bemis, D. and Pykkänen, L. (2013). Combination across domains: an MEG investigation into the relationship between mathematical, pictorial, and linguistic processing. *Frontiers in Psychology*, **3**, article 583.

- Bichot, N., Thompson, K., Rao, S., and Schall, J. (2001). Reliability of macaque frontal eye field neurons signaling saccade targets during visual search. *Journal of Neuroscience*, **21**, 713–725.
- Binkofski, F. and Buxbaum, L. (2013). Two action systems in the human brain. *Brain and Language*, **127**, 222–229.
- Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer, Berlin.
- Bonnici, H., Chadwick, M., Lutti, A., Hassabis, D., Weiskopf, N., and Maguire, E. (2012). Detecting representations of recent and remote autobiographical memories in vmPFC and hippocampus. *Journal of Neuroscience*, **32**(47), 16982–16991.
- Borji, A. and Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**(1), 185–207.
- Borji, A., Cheng, M., Jiang, H., and Li, J. (2014). Salient object detection: A survey. arXiv preprint 1411.5878.
- Bracci, S. and Op de Beeck, H. (2016). Dissociations and associations between shape and category representations in the two visual pathways. *Journal of Neuroscience*, **36**(2), 432–444.
- Braver, T. and Cohen, J. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. In S. Monsell and J. Driver, editors, *Attention and Performance XVIII: Control of cognitive processes*, pages 713–737. MIT Press.
- Bruce, N. and Tsotsos, J. (2007). An information theoretic model of saliency and visual search. In N. Paletta and E. Rome, editors, *Proceedings of WAPCV 2007*, pages 171–183. Springer Verlag, Berlin.
- Buschman, T., Siegel, M., Roy, J., and Miller, E. (2011). Neural substrates of cognitive capacity limitations. *Proceedings of the National Academy of the USA*, **108**(27), 11252–11255.
- Buxbaum, L. and Kal’ene, S. (2010). Action knowledge, visuomotor activation, and embodiment in the two action systems. *Annals of the New York Academy of Sciences*, **1191**, 201–218.
- Cant, J. and Goodale, M. (2007). Attention to form or surface properties modulates different regions of human occipitotemporal cortex. *Cerebral Cortex*, **17**, 713–731.
- Cant, J., Arnott, S., and Goodale, M. (2009). fMR-adaptation reveals separate processing regions for the perception of form and texture in the human ventral stream. *Experimental Brain Research*, **192**, 391–405.
- Caspari, N., Popivanov, I., De Mazière, P., Vanduffel, W., Vogels, R., Orban, G., and Jastorff, J. (2014). Fine-grained stimulus representations in body selective areas of human occipito-temporal cortex. *NeuroImage*, **102**, 484–497.
- Cate, A., Goodale, M., and Köhler, S. (2011). The role of apparent size in building- and object-specific regions of ventral visual cortex. *Brain Research*, **1388**, 109122.
- Cavina-Pratesi, C., Kentridge, R., Heywood, C., and Milner, A. (2010). Separate channels for processing form, texture, and color: Evidence from fmri adaptation and visual object agnosia. *Cerebral Cortex*, **20**, 2319–2332.
- Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, **26**, 609–651.
- Chang, K.-M., Cherkassky, V., Mitchell, T., and Just, M. (2009). Quantitative modeling of the neural representation of adjective-noun phrases to account for fmri activation. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 638–646, Suntec, Singapore.

- Chao, O., Huston, J., Li, J.-S., Wang, A.-L., and de Souza Silva, M. (2016). The medial prefrontal cortex/lateral entorhinal cortex circuit is essential for episodic-like memory and associative object-recognition. *Hippocampus*, **26**, 633–645.
- Chatterjee, A., Thomas, A., Smith, S., and Aguirre, G. (2009). The neural response to facial attractiveness. *Neuropsychology*, **23**(2), 135–143.
- Chersi, F. and Burgess, N. (2015). The cognitive architecture of spatial navigation: Hippocampal and striatal contributions. *Neuron*, **88**, 64–77.
- Cinque, G. (1994). Evidence for partial N-movement in the Romance DP. In G. Cinque, J. Y. Pollock, L. Rizzi, and R. Zanuttini, editors, *Towards Universal Grammar: Studies in Honor of Richard Kayne*. Georgetown University Press, Washington, DC.
- Cisek, P. and Kalaska, J. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action. *Neuron*, **45**, 801–814.
- Colby, C. and Goldberg, M. (1999). Space and attention in parietal cortex. *Annual Review of Neuroscience*, **22**, 319–349.
- Connolly, A., Guntupalli, S., Gors, J., Hanke, M., Halchenko, Y., Wu, Y.-C., Abdi, H., and Haxby, J. (2012). The representation of biological classes in the human brain. *Journal of Neuroscience*, **32**(8), 2608–2616.
- Connolly, A., Sha, L., Guntupalli, S., Oosterhof, N., Halchenko, Y., Nastase, S., Visconti, M., Abdi, H., Jobst, B., Gobbin, I., and Haxby, J. (2016). How the human brain represents perceived dangerousness or ‘predacity’ of animals. *Journal of Neuroscience*, **36**(19), 5373–5384.
- Corbitt, L. and Balleine, B. (2000). The role of the hippocampus in instrumental conditioning. *Journal of Neuroscience*, **20**(11), 4233–4239.
- Courtney, S., Ungerleider, L., Keil, K., and Haxby, J. (1996). Object and spatial visual working memory activate separate neural systems in human cortex. *Cerebral Cortex*, **6**(1), 39–49.
- Coutanche, M. and Thompson-Schill, S. (2015). Creating concepts from converging features in human cortex. *Cerebral Cortex*, **25**, 2584–2593.
- Damasio, A., editor (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace, New York.
- Damasio, A. and Damasio, H. (1994). Cortical systems for retrieval of concrete knowledge: The convergence zone framework. In C. Koch and J. Davis, editors, *Large-scale Neuronal Theories of the Brain*, pages 61–74. MIT Press, Cambridge, MA.
- D’Esposito, M., Postle, B., and Rypma, B. (2000). Prefrontal cortical contributions to working memory: evidence from event-related fmri studie. *Experimental Brain Research*, **133**(1), 3–11.
- Domenech, P. and Koechlin, E. (2015). Executive control and decision-making in the prefrontal cortex. *Current Opinion in Behavioral Sciences*, **1**, 101–106.
- Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, **67**(3), 547–619.
- Driver, J., Baylis, G., Goodrich, S., and Rafal, R. (1994). Axis-based neglect of visual shapes. *Neuropsychologia*, **32**(11), 1353–1365.
- Duan, H. and Wang, X. (2015). Visual attention model based on statistical properties of neuron responses. *Nature Scientific Reports*, **5**, 8873.
- Duff, M. and Brown-Schmidt, S. (2012). The hippocampus and the flexible use and processing of language. *Frontiers in Human Neuroscience*, **6**, e69.

- Duncan, J. and Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review*, **96**, 433–458.
- Eichenbaum, H., Dudchenko, P., Wood, E., Shapiro, M., and Tanila, H. (1999). The hippocampus, memory, review and place cells: Is it spatial memory or a memory space? *Neuron*, **23**, 209–226.
- Ekstrom, A., Kahana, M., Caplan, J., Fields, T., Isham, E., Newman, E., and Fried, I. (2003). Cellular networks underlying human spatial navigation. *Nature*, **184**, 184–187.
- Elsner, B., Hommel, B., Mentschel, C., Drzezga, A., Prinz, W., Conrad, B., and Siebner, H. (2002). Linking actions and their perceivable consequences in the human brain. *NeuroImage*, **17**, 364–372.
- Ferbinteanu, J. and Shapiro, M. (2003). Prospective and retrospective memory coding in the hippocampus. *Neuron*, **40**, 1227–1239.
- Flanagan, J. and Rao, A. (1995). Trajectory adaptation to a nonlinear visuomotor transformation: evidence of motion planning in visually perceived space. *Journal of Neurophysiology*, **74**(5), 2174–2178.
- Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., and Rizzolatti, G. (2005). Parietal lobe: From action organisation to intention understanding. *Science*, **308**, 662–667.
- Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, **3**, 194–200.
- Frankland, S. and Green, J. (2015). An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of the USA*, **112**(37), 11732–11737.
- Freedman, D., Riesenhuber, M., Poggio, T., and Miller, E. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorisation. *Journal of Neuroscience*, **15**, 5235–5246.
- Freud, E., Ganel, T., Shelef, I., Hammer, M., Avidan, G., and Behrmann, M. (2015). Three-dimensional representations of objects in dorsal cortex are dissociable from those in ventral cortex. *Cerebral Cortex*, pages 1–13.
- Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, **11**, 127–138.
- Fyshe, A. (2015). *Corpora and Cognition: The Semantic Composition of Adjectives and Nouns in the Human Brain*. Ph.D. thesis, CMU, School of Computer Science.
- Gabay, S., Kalanthroff, E., Henik, A., and Gronau, N. (2016). Conceptual size representation in ventral visual cortex. *Neuropsychologia*, **81**, 198–206.
- Gallivan, J., McLean, A., Flanagan, J., and Culham, J. (2013). Where one hand meets the other: Limb-specific and action-dependent movement plans decoded from preparatory signals in single human frontoparietal brain areas. *Journal of Neuroscience*, **33**(5), 1991–2008.
- Goodale, M. and Milner, A. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, **15**, 20–25.
- Gorman, C. and Knott, A. (2016). A neural network model of hierarchical category development. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society (CogSci)*, pages 342–347, Philadelphia, PA.
- Gottlieb, J., Kusunoki, M., and Goldberg, M. (1998). The representation of visual salience in monkey parietal cortex. *Nature*, **391**, 481–484.

- Gotts, S., Milleville, S., and Martin, A. (2015). Object identification leads to a conceptual broadening of object representations in lateral prefrontal cortex. *Neuropsychologia*, **76**, 62–78.
- Griffin, A. (2015). Role of the thalamic nucleus reuniens in mediating interactions between the hippocampus and medial prefrontal cortex during spatial working memory. *Frontiers in Systems Neuroscience*, **9**, e29.
- Güçlü, U. and van Gerven, M. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, **35**(27), 10005–10014.
- Guntupalli, J., Hanke, M., Halchenko, Y., Connolly, A., Ramadge, P., and Haxby, J. (2016). A model of representational spaces in human cortex. *Cerebral Cortex*, **26**, 2919–2934.
- Harris, C. and Wolpert, D. (1998). Signal-dependent noise determines motor planning. *Nature*, **394**(6695), 780–784.
- Hassabis, D., Chu, C., Rees, G., Weiskopf, N., Peter, D., and Maguire, E. (2009). Decoding neuronal ensembles in the human hippocampus. *Current Biology*, **19**, 546–554.
- Hassabis, D., Spreng, R., Rusu, A., Robbins, C., Mar, R., and Schacter, D. (2013). Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex*, **24**(8), 1979–1987.
- Haxby, J., Guntupalli, S., Connolly, A., Halchenko, Y., Conroy, B., Gobbini, I., Hanke, M., and Ramadge, P. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, **72**, 404–416.
- Henderson, J. and Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, **50**, 243–271.
- Herbert, C., Ethofer, T., Anders, S., Junghofer, M., Wildgruber, D., Grodd, W., and Kissler, J. (2009). Amygdala activation during reading of emotional adjectives—an advantage for pleasant content. *Scan*, **4**, 35–49.
- Hinton, G. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, **11**(10), 428–434.
- Horner, A., Gadian, D., Fuentemilla, L., Jentschke, S., Vargha-Khadem, F., and Duzel, E. (2012). A rapid, hippocampus-dependent, item-memory signal that initiates context memory in humans. *Current Biology*, **22**, a2012.
- Hu, Y. and Goodale, M. (2000). Grasping after a delay shifts size-scaling from absolute to relative metrics. *Journal of Cognitive Neuroscience*, **12**(5), 856–868.
- Hubel, D. and Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cats visual cortex. *Journal of Physiology*, **160**, 106–154.
- Hung, C.-C., Carlson, E., and Connor, C. (2012). Medial axis shape coding in macaque inferotemporal cortex. *Neuron*, **74**, 1099–1113.
- Iordan, M., Greene, M., Beck, D., and Fei-Fei, L. (2016). Typicality sharpens category representations in object-selective cortex. *NeuroImage*, **134**, 170–179.
- Ishizu, T. and Zeki, S. (2011). Toward a brain-based theory of beauty. *PLoS One*, **6**(7), e21852.
- Ito, H., Zhang, S.-J., Witter, M., Moser, E., and Moser, M.-B. (2015). A prefrontalthalamohippocampal circuit for goal-directed spatial navigation. *Nature*, **522**, 50–59.
- Itti, L. and Baldi, P. (2009). Bayesian surprise attracts attention. *Vision Research*, **49**, 1295–1306.

- Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews—Neuroscience*, **2**, 1–11.
- Jacobs, J., Kahana, M., Ekstrom, A., Mollison, M., and Fried, I. (2010). A sense of direction in human entorhinal cortex. *Proceedings of the National Academy of Sciences of the USA*, **107**(14), 6487–6492.
- Jeannerod, M. (1999). To act or not to act: Perspectives on the representation of actions. *Quarterly Journal of Experimental Psychology: A*, **52**(1), 1–29.
- Jeannerod, M., Decety, J., and Michel, F. (1999). Impairment of grasping movements following a bilateral posterior parietal lesion. *Neuropsychologia*, **32**(4), 369–380.
- Johansson, R., Westling, G., Backstrom, A., and Flanagan, J. (2001). Eye-hand coordination in object manipulation. *Journal of Neuroscience*, **21**(17), 6917–6932.
- Johnson, A. and Redish, D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, **27**(45), 12176–12189.
- Jung, M., Wiener, S., and McNaughton, B. (1994). Comparison of spatial firing characteristics of units in dorsal and ventral hippocampus of the rat. *Journal of Neuroscience*, **14**, 7347–7356.
- Kahneman, D., Treisman, A., and Gibbs, B. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, **24**, 175–219.
- Kawabata, H. and Zeki, S. (2004). Neural correlates of beauty. *Journal of Neurophysiology*, **91**, 1699–1705.
- Kemmerer, D., Tranel, D., and Zdanczyk, C. (2009). Knowledge of the semantic constraints on adjective order can be selectively impaired. *Journal of Neurolinguistics*, **22**(1), 91–108.
- Keysers, C. and Perrett, D. (2004). Demystifying social cognition: A Hebbian perspective. *Trends in Cognitive Sciences*, **8**(11), 501–507.
- Kiani, R. and Shadlen, M. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, **324**, 759–764.
- King, P., Zylberberg, J., and DeWeese, M. (2013). Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of V1. *Journal of Neuroscience*, **33**, 5475–5485.
- Knott, A. (2012). *Sensorimotor Cognition and Natural Language Syntax*. MIT Press, Cambridge, MA.
- Koch, C. and Ullman, S. (1985). Shifts in underlying visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, **4**(4), 219–227.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, **43**, 59–69.
- Komatsu, H. and Ideura, Y. (1993). Relationships between color, shape, and pattern selectivities of neurons in the inferior temporal cortex of the monkey. *Journal of Neurophysiology*, **70**, 677–694.
- Konkel, A. and Cohen, N. (2009). Relational memory and the hippocampus: representations and methods. *Frontiers in Neuroscience*, **3**(2), 166–174.
- Konkle, T. and Oliva, A. (2012). A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, **74**(6), 1114–1124.

- Kourtzi, Z. and Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital cortex. *Science*, **293**, 1506–1509.
- Kriegeskorte, N., Mur, M., Ruff, D., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, **60**, 1120–1141.
- Kriete, T., Noelle, D., Cohen, J., and O'Reilly, R. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of the USA*, **110**(41), 1120–1141.
- Krizhevsky, A. and Sutskever, I Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1106–1114.
- Kropff, E., Carmichael, J., Moser, M.-B., and Moser, E. (2015). Speed cells in the medial entorhinal cortex. *Nature*, **523**, 419–424.
- Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**(1), 79–86.
- Land, M. and McLeod, P. (2000). From eye movements to actions: how batsmen hit the ball. *Nature Neuroscience*, **3**(12), 1340–1345.
- Le Cun, Y. and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 255–258. MIT Press, Cambridge, MA.
- Lebedev, M., Messinger, A., Kralik, J., and Wise, S. (2004). Representation of attended versus remembered locations in prefrontal cortex. *PLoS Biology*, **2**(11), e365.
- Lee-Hand, J. and Knott, A. (2015). A neural network model of causative actions. *Frontiers in Neurorobotics*, **9**, Article 4.
- Lefebvre, G. and Garcia, C. (2008). A probabilistic self-organizing map for facial recognition. In *19th International Conference on Pattern Recognition*, Florida.
- Liddle, M. (2010). *Some neuro-computational investigations into the reviewing of object-files*. Ph.D. thesis, Department of Computer Science, University of Otago.
- Liu, N., Wang, J., and Gong, Y. (2015). Deep self-organizing map for visual classification. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.
- Liu, T., Yuan, Z., Sun, J., and Shum, H.-Y. (2011). Learning to detect a salient object. *IEEE Transactions on Software Engineering*, **33**(2), 353–367.
- Logothetis, N. and Sheinberg, D. (1996). Visual object recognition. *Annual Review of Neuroscience*, **19**, 577–621.
- MacKay, D., Burke, D., and Stewart, R. (1998). H.M.s language production deficits: Implications for relations between memory, semantic binding, and the hippocampal system. *Journal of Memory and Language*, **38**, 28–69.
- Macoir, J., Laforce, R., Brisson, M., and Wilson, M. (2015). Preservation of lexical-semantic knowledge of adjectives in the semantic variant of primary progressive aphasia: Implications for theoretical models of semantic memory. *Journal of Neurolinguistics*, **34**, 1–14.
- Martin, A. and Weisberg, J. (2003). Neural foundations for understanding social and mechanical concepts. *Cognitive Neuropsychology*, **20**(3/4/5/6), 575–587.

- Meyers, E., Freedman, D., Kreiman, G., Miller, E., and Poggio, T. (2008). Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of Neurophysiology*, **100**, 1407–1419.
- Miller, E. and Cohen, J. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, **24**, 167–202.
- Miller, J., Neufang, M., Solway, A., Brandt, A., Trippel, M., Mader, I., Hefft, S., Merkow, M., Polyn, S., Jacobs, J., Kahana, M., and Andreas Schulze-Bonhage, A. (2013). Neural activity in human hippocampal formation reveals the spatial context of retrieved memories. *Science*, **342**(6162), 11111114.
- Moore, T. and Armstrong, K. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, **421**(6921), 370–373.
- Morita, T., Kochiyama, T., Okada, T., Yonekura, Y., Matsumura, M., and Sadato, N. (2004). The neural substrates of conscious color perception demonstrated using fMRI. *NeuroImage*, **21**, 1665–1673.
- Moser, M., Moser, E., Forrest, E., Andersen, P., and Morris, R. (1995). Spatial learning with a minislab in the dorsal hippocampus. *Proceedings of the National Academy of Sciences of the USA*, **92**(21), 9697–9701.
- Murata, A., Gallese, V., Luppino, G., Kaseda, M., and Sakata, H. (2000). Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area AIP. *Journal of Neurophysiology*, **83**(5), 2580–2601.
- Murphey, D., Yoshor, D., and Beauchamp, M. (2008). Perception matches selectivity in the human anterior color center. *Current Biology*, **18**(3), 216–220.
- Navawongse, R. and Eichenbaum, H. (2013). Distinct pathways for rule-based retrieval and spatial mapping of memory representations in hippocampal neurons. *Journal of Neuroscience*, **33**(3), 1002–1013.
- Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, **9**, 353–383.
- Nguyen, A., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Computer Vision and Pattern Recognition (CVPR 15)*, *IEEE, 2015*.
- Nieder, A. and Miller, E. (2004). A parieto-frontal network for visual numerical information in the monkey. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(19), 7457–7462.
- Nobre, A., Gitelman, D., Dias, E., and Mesulam, M. (2000). Covert visual spatial orienting and saccades: Overlapping neural systems. *NeuroImage*, **11**, 210–216.
- Öhman, A. (2005). The role of the amygdala in human fear: Automatic detection of threat. *Psychoneuroendocrinology*, **30**, 953–958.
- Olson, I. and Marshuetz, C. (2005). Facial attractiveness is appraised in a glance. *Emotion*, **5**(4), 498–502.
- O’Reilly, R. (2010). The What and How of prefrontal cortical organization. *Trends in Neurosciences*, **33**, 355–361.
- Oztop, E., Bradley, N., and Arbib, M. (2004). Infant grasp learning: a computational model. *Experimental Brain Research*, **158**, 480–503.

- Pereira, A., James, K., Jones, S., and Smith, L. (2010). Early biases and developmental changes in self-generated object views. *Journal of Vision*, **10**, 1–13.
- Perrett, D., Harries, M., Bevan, R., Thomas, S., Benson, P., Mistlin, A., Chitty, A., Hiatenen, J., and Ortega, J. (1989). Frameworks of analysis for the neural representation of animate objects and actions. *Journal of Experimental Biology*, **146**, 87–113.
- Pessoa, L. (2005). To what extent are emotional visual stimuli processed without attention and awareness? *Current Opinion in Neurobiology*, **15**, 188–196.
- Peterson, M. and Kim, J. (2001). On what is bound in figures and grounds. *Visual Cognition*, **8**(3/4/5), 329–348.
- Petrides, M. (1996). Lateral frontal cortical contribution to memory. *Seminars in the Neurosciences*, **8**, 57–63.
- Poppenk, J., Evensmoen, H., Moscovitch, M., and Nadel, L. (2013). Long-axis specialization of the human hippocampus. *Trends in Cognitive Sciences*, **17**(5), 230–240.
- Posner, M., Rafal, R., Choate, L., and Vaughn, J. (1984). Inhibition of return: Neural basis and function. *Cognitive Neuropsychology*, **2**, 211–228.
- Pouget, A. and Sejnowski, T. (1997). Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience*, **9**(2), 222–237.
- Pouget, A., Beck, J., Ma, W.-J., and Latham, P. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, **16**(9), 1170–1178.
- Preston, A. and Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, **23**, R764R773.
- Pylyshyn, Z. and Storm, R. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, **3**, 179–197.
- Quiroga, R., Reddy, L., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, **435**, 1102–1107.
- Ranganath, C. and Knight, R. (2003). Prefrontal cortex and episodic memory: Integrating findings from neuropsychology and functional brain imaging. In E. Wilding, A. Parker, and T. Bussey, editors, *Memory encoding and retrieval: a cognitive neuroscience perspective*, pages 1–14. Psychology Press, New York.
- Ranganath, C. and Paller, K. (1999a). Frontal brain activity during episodic and semantic retrieval: insights from event-related potentials. *Journal of Cognitive Neuroscience*, **11**(6), 598–609.
- Ranganath, C. and Paller, K. (1999b). Frontal brain potentials during recognition are modulated by requirements to retrieve perceptual detail. *Neuron*, **22**(3), 605–613.
- Ranganath, C. and Paller, K. (2000). Neural correlates of memory retrieval and evaluation. *Cognitive Brain Research*, **9**(2), 209–222.
- Rao, R. and Ballard, D. (1996). A computational model of spatial representations that explains object-centred neglect in parietal patients. In *Proceedings of Computational Neuroscience*.
- Redish, D. (2016). Vicarious trial and error. *Nature Reviews Neuroscience*, **17**, 147–159.
- Redish, D. and Mizumori, S. (2015). Memory and decision making. *Neurobiology of learning and memory*, **117**, 1–3.
- Reynolds, J., Zacks, J., and Braver, T. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science*, **31**, 613–643.

- Rolls, E. (2009). The neurophysiology and computational mechanisms of object representation. In S. Dickinson, A. Leonardis, B. Schiele, and M. Tarr, editors, *Object Categorization: Computer and Human Vision Perspectives*, pages 257–287. Cambridge University Press, Cambridge, UK.
- Rolls, E., Xiang, J., and Franco, L. (2005). Object, space, and object-space representations in the primate hippocampus. *Journal of Neurophysiology*, **94**, 833–844.
- Rolls, E., Grabenhorst, F., and Deco, G. (2010). Decision-making, errors, and confidence in the brain. *Journal of Neurophysiology*, **104**, 2359–2374.
- Sapir, A., Hayes, A., Henik, A., Danziger, S., and Rafal, R. (2004). Parietal lobe lesions disrupt saccadic remapping of inhibitory location tagging. *Journal of Cognitive Neuroscience*, **16**(4), 503–509.
- Sağlam, M., Lehnen, N., and Glasauer, S. (2011). Optimal control of natural eye-head movements minimizes the impact of noise. *Journal of Neuroscience*, **31**(45), 16185–16193.
- Schendan, H. and Stern, C. (2007). Mental rotation and object categorization share a common network of prefrontal and dorsal and ventral regions of posterior cortex. *NeuroImage*, **35**, 1264–1277.
- Schmolesky, M. (2007). The primary visual cortex. In: *Webvision: The Organization of the Retina and Visual System*. Online textbook (accessed 2016).
- Schneider, W. and Deubel, H. (2002). Selection-for-perception and selection-for-spatial-motor-action are coupled by visual attention: A review of recent findings and new evidence from stimulus-driven saccade control. In W. Prinz and B. Hommel, editors, *Attention and Performance XIX: Common Mechanisms in Perception and Action*, pages 609–627. Oxford University Press.
- Schultz, W., Dayan, P., and Montague, P. (1997). A neural substrate of prediction and reward. *Science*, **275**, 1593–1599.
- Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, **61**(1–2), 39–91.
- Stewart, T. and Eliasmith, C. (2012). Compositionality and biologically plausible models. In M. Werning and W. Hinzen, editors, *The Oxford Handbook of Compositionality*. Oxford University Press, New York.
- Stokes, M. (2015). ‘Activity-silent’ working memory in prefrontal cortex: a dynamic coding framework. *Trends in Cognitive Sciences*, **19**(7), 394–405.
- Stokes, S. (2013). The impact of phonological neighbourhood density on typical and atypical emerging lexicons. *Journal of Child Language*, **CJO2013**, doi:10.1017/S030050091300010X.
- Strickert, M. and Hammer, B. (2005). Merge SOM for temporal data. *Neurocomputing*, **64**, 39–71.
- Struiksma, M., Noordzij, M., Neggers, S., Bosker, W., and Postma, A. (2011). Spatial language processing in the blind: Evidence for a supramodal representation and cortical reorganization. *PLoS One*, **6**(9), e24253.
- Sullivan and de Sa, V. (2012). A temporal trace and som-based model of complex cell development. *Neurocomputing*, **58–60**, 827–833.
- Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, **19**, 103–139.
- Tanaka, K. (1997). Mechanisms of visual object recognition: Monkey and human studies. *Current Opinion in Neurobiology*, **7**(4), 523–529.

- Tipper, S., Lortie, C., and Baylis, G. (1992). Selective reaching: Evidence for action-centered attention. *Journal of Experimental Psychology: Human Perception and Performance*, **18**, 891–905.
- Tranel, D. and Kemmerer, D. (20). Neuroanatomical correlates of locative prepositions. *Cognitive Neuropsychology*, **21**(7), 719–749.
- Tsao, D., Vanduffel, W., Sasaki, Y., Fize, D., Knutsen, T., Mandeville, J., Wald, L., Dale, A., Rosen, B., Van Essen, D., Livingstone, M., Orban, G., and Tootell, R. (2003). Stereopsis activates V3A and caudal intraparietal areas in macaques and humans. *Neuron*, **39**, 555–568.
- Uddin, L., Iacoboni, M., Lange, C., and Keenan, J. (2007). The self and social cognition: the role of cortical midline structures and mirror neurons. *Trends in Cognitive Sciences*, **11**(4), 153–157.
- Umiltà, M., Escola, L., Intskirveli, I., Grammont, F., Rochat, M., Caruana, F., Jezzini, A., Gallese, V., and Rizzolatti, G. (2008). When pliers become fingers in the monkey motor system. *PNAS*, **105**(6), 2209–2213.
- Ungerleider, L. and Haxby, J. (1994). ‘what’ and ‘where’ in the human brain. *Current Opinion in Neurobiology*, **4**, 157–165.
- Valdez, A., Papesh, M., Treiman, D., Smith, K., Goldinger, S., and Steinmetz, P. (2015). Distributed representation of visual objects by single neurons in the human brain. *Journal of Neuroscience*, **35**(13), 5180–5186.
- van Beers, R. (2007). The sources of variability in saccadic eye movements. *Journal of Neuroscience*, **27**, 8757–8770.
- Vanetti, M., Gallo, I., and Nodari, A. (2013). Unsupervised feature learning using self-organizing maps. In *Proceedings of the International Conference on Computer Vision Theory and Applications - Volume 1: VISAPP*, pages 596–601.
- Verfaillie, K., De Troy, A., and Van Rensbergen, J. (1994). Transsaccadic integration of biological motion. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **20**(3), 649–670.
- Wagner, D., Haxby, J., and Heatherton, T. (2012). The representation of self and person knowledge in the medial prefrontal cortex. *Wiley interdisciplinary reviews. Cognitive science*, **3**(4), 451–470.
- Walles, H., Knott, A., and Robins, A. (2008). A model of cardinality blindness in inferotemporal cortex. *Biological Cybernetics*, **98**(5), 427–437.
- Walles, H., Robins, A., and Knott, A. (2014). A perceptually grounded model of the singular-plural distinction. *Language and Cognition*, **6**, 1–43.
- Warden, M. and Miller, E. (2007). The representation of multiple objects in prefrontal neuronal delay activity. *Cerebral Cortex*, **17**, i41–i50.
- Watson, D., Young, A., and Andrews, T. (2016). Spatial properties of objects predict patterns of neural response in the ventral visual pathway. *NeuroImage*, **126**, 173–183.
- Webb, A., Knott, A., and MacAskill, M. (2010). Eye movements during transitive action observation have sequential structure. *Acta Psychologica*, **133**, 51–56.
- Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V., and Rizzolatti, G. (2003). Both of us disgusted in *my* insula: The common neural basis of seeing and feeling disgust. *Neuron*, **40**, 655–664.

- Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., and Anderson, M. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature Neuroscience*, **18**(4).
- Wiskott, L. and Sejnowski, T. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, **14**, 715–770.
- Wixted, J., Squire, L., Jang, Y., Papesh, M., Goldinger, S., Kuhn, J., Smith, K., Treiman, D., and Steinmetz, P. (2014). Sparse and distributed coding of episodic memory in neurons of the human hippocampus. *Proceedings of the National Academy of Sciences of the USA*, **111**(26), 96219626.
- Wolfe, J. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, **1**(2), 202–238.
- Wolpert, D. and Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, **11**, 1317–1329.
- Woollett, K. and Maguire, E. (2000). Acquiring “the knowledge” of london’s layout drives structural brain changes. *Current Biology*, **21**, 2109–2114.
- Wu, D., Waller, S., and Chatterjee, A. (2007). The functional neuroanatomy of thematic role and locative relational knowledge. *Journal of Cognitive Neuroscience*, **19**(9), 1542–1555.
- Yamagata, T., Nakayama, Y., Tanji, J., and Hoshi, E. (2012). Distinct information representation and processing for goal-directed behavior in the dorsolateral and ventrolateral prefrontal cortex and the dorsal premotor cortex. *Journal of Neuroscience*, **32**(37), 12934–12949.
- Yee, E., Drucker, D., and Thompson-Schill, S. (2010). fMRI-adaptation evidence of overlapping neural representations for objects related in function or manipulation. *NeuroImage*, **50**(2), 753763.
- Yonelinas, A. (2013). The hippocampus supports high-resolution binding in the service of perception, working memory and long-term memory. *Behavioral Brain Research*, **254**, 34–44.
- Zeidman, P. and Maguire, E. (2016). Anterior hippocampus: the anatomy of perception, imagination and episodic memory. *Nature Reviews Neuroscience*, **17**, 173–182.
- Zeiler, M. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision—ECCV 2014*, pages 818–833.
- Zhang, Y., Meyers, E., Bichot, N., Serre, T., Poggio, T., and Desimone, R. (2011). Object decoding with attention in inferior temporal cortex. *Proceedings of the National Academy of Sciences of the USA*, **108**(21), 8850–8855.