
Computational modelling of hand-centred visual representations in the primate brain



Juan Manuel Galeazzi González
St. Cross College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy in Experimental Psychology

Trinity 2015

This thesis is specially dedicated to my beloved parents, my sisters and my lovely wife for their unconditional love and unswerving support, and to the memory of my late uncle Jesús José González Batriz who was like a big brother to me.

Acknowledgements

I would like to express my immense gratitude to everyone who supported this doctoral project.

Firstly, I would like to thank my supervisor and friend, Dr. Simon Stringer, without his expert guidance and persistent encouragement this thesis would not have been possible. His admirable commitment and dedication to the intellectual growth and well-being of the members of the lab speaks volumes about his generosity and kind-heartedness. Similarly, I would also like to thank my second co-supervisor Dr. Loredana Minini for sharing her knowledge and expertise on visuomotor neuroscience and all the stimulating discussions about the ventral and dorsal visual systems. I would also like to thank Prof. Glyn Humphreys and Prof. Charles Spence for their valuable feedback during my *transfer of status* and *confirmation of status* examinations.

I would like to express my gratitude to the Mexican National Council of Science and Technology (Consejo Nacional de Ciencia y Tecnología, CONACYT), the Youth Institute of Jalisco (Instituto Jalisciense de la Juventud Jalisco, IJJ), the Secretariat of Public Education (Secretaría de Educación Pública, SEP) and the Oxford Foundation for Theoretical Neuroscience and Artificial Intelligence for their very generous assistance in funding this doctoral project.

I would also like to thank everyone in the lab: Dan, Ben, James, Bedeho, Hector, Akihiro, Harry, Hannah, Irina, Daniel, Nasir, David and Mihaela. It was a great pleasure to meet you all and a great honour to work together. I am

especially thankful to Ben and Dan for making sure our computer system ran as smoothly as humanly possible. Thank you Dan also for your leadership and consistent support to everyone in the lab, and particularly for introducing me to the intricate beauty of cricket. Thank you Hector for your companionship, we shared many moments of frustration specially during the countless hours of debugging codes and searching parameters. In particular, I am profoundly grateful to Bedeho, not only you helped me on countless occasions teaching me some of your programming ninja skills, but your friendship made my stay in Oxford incredibly pleasant and intellectually stimulating. I will treasure all our lively debates and irreverent conversations ranging from discussing the merits and limitations of socioeconomic policies and ideologies to the most mundane topics of conversation. We have not found yet a single topic in which we agree, however that did not stop us from developing a deep bond based on mutual respect and a shared passion in the search for knowledge.

Lastly, I would like to thank to the most important people in my entire life and to whom this thesis is especially dedicated: My wife and my family. Mom and Dad, I owe you everything, if there is anything good or worthy of praise in my character is because I learned it from you. Your dedication, infinite kindness and unconditional love have always inspired me to give the best of myself. Gaby and Ale, thank you for all your love and support, I know that we can always count on each other. I am incredibly proud of the amazing women you have become, although in my eyes you will always be my little sisters. Laura you are my one and true love, you have always been there for me, even in the most trying of moments. I am infinitely thankful for your love and your confidence in me, specially on those moments when I am full of self-doubt. You are my main motivation to keep moving forward and I would simply not have written this thesis without you by my side!

Abstract

Neurons have been found in various areas of the primate brain that respond to the location of objects with respect to the hand. In this thesis we present a number of self-organising theories of how these hand-centred visual receptive fields could develop during visually guided learning in unsupervised neural network models. Experimental chapters 3 and 4 show the development of hand-centred representations under two different learning hypotheses *Continuous Transformation Learning* and *Trace learning*, using a neural network model of the primate visual system, VisNet. In Chapter 5 the trace learning hypothesis is advanced by gradually increasing the realism in the visual training conditions. Results are shown using realistic visual scenes consisting of multiple targets presented at the same time around the hand. In Chapter 6 this learning hypothesis was further explored using natural eye and head movements recorded from human participants. This required a new time accurate differential formulation of the VisNet model to faithfully represent the temporal dynamics of the recorded gaze changes. These results provide an important step in showing how localised, hand-centred receptive fields could emerge under more ecologically realistic training conditions. In Chapter 7 we present a new self-organising model of how hand-centred representations may develop using a proprioceptive signal representing the position of the hand. In these simulations we first develop a body-centred representation of the location of the visual targets by trace learning. I then showed how these body-centred representations of the visual targets can be combined with a proprioceptive representation of the location of the hand in order to develop hand-centred representations. This thesis concludes with a discussion of the main findings and some suggestions for future work.

Contents

List of Figures	vii
List of Tables	xi
1 Introduction	1
1.1 The primate visual system: A tale of two streams	2
1.1.1 Vision for perception: From the eye to the cortex	4
1.1.2 Vision for action: The dorsal stream	6
1.2 Coordinate frames in the primate parietal areas: The posterior parietal cortex and premotor areas	10
1.2.1 Hierarchy of coordinate systems	14
1.2.2 Gain modulated responses and heterogeneous representations	17
1.2.3 Visuospatial and visuomotor deficits associated with lesions in pos- terior parietal areas	18
1.3 Hand-centred representations in the brain	22
1.3.1 Encoding reach vectors in hand-centred coordinates	23
1.3.2 Body-part centred representations of visual space: Hand-centred re- ceptive fields	25
1.3.3 The roles of vision and proprioception in the localisation of the hand	31
1.4 Computational modelling of reference frame transformations and body-part centred representations	37

1.5	Overview of the thesis: Self-organization of hand-centred representations . . .	41
2	Methods: Competitive networks, VisNet and metrics	46
2.1	Competitive learning	47
2.2	Visnet	50
2.2.1	Learning in VisNet: Continuous transformation learning and trace learning	55
2.2.2	Information Analysis	56
2.2.3	Selection and optimization of parameters in VisNet	60
3	The development of hand-centred visual representations by Continuous Transformation learning	63
3.1	Introduction	64
3.1.1	CT learning hypothesis	67
3.2	Method	71
3.2.1	Model architecture and parameters	71
3.2.2	Stimuli	73
3.2.3	Training	73
3.2.4	Learning	75
3.2.5	Testing	76
3.3	Results	76
3.3.1	Baseline simulation with three hand-centred target locations	76
3.3.2	Capacity study: increasing the number of hand-centred target locations	83
3.3.3	Performance of the model using a SOM architecture	86
3.3.4	Effect of diluting connectivity in a SOM architecture and Competitive networks	91
3.4	Discussion	96

4	The development of hand-centred visual representations by trace learning	99
4.1	Introduction	100
4.2	Hypothesis	102
4.3	Method	104
4.3.1	Model architecture and parameters	104
4.3.2	Trace learning	105
4.3.3	Stimuli	106
4.3.4	Training	108
4.3.5	Testing	109
4.4	Results	109
4.4.1	Baseline simulation with three hand-centred target locations	109
4.4.2	Performance of model with different numbers of training epochs	115
4.4.3	Performance of the model with larger retinal shifts	118
4.4.4	Capacity study: increasing the number of hand-centred target locations	121
4.4.5	Performance of the model using a SOM architecture	126
4.5	Discussion	133
5	The development of visual hand-centred representations under more realistic training conditions	136
5.1	Introduction	137
5.2	Methods	140
5.2.1	Stimuli	141
5.2.2	Model parameters	143
5.2.3	Training	143
5.2.4	Testing	144
5.3	Results	146
5.4	Discussion	150

5.5	Training the network with more realistic natural objects presented around the hand with more natural textured backgrounds	152
5.6	Methods	156
5.6.1	Stimuli	156
5.6.2	Model parameters	157
5.6.3	Training	158
5.6.4	Testing	159
5.7	Results	162
5.8	Discussion	171
5.9	Assessing the selectivity of neurons for object shape within their localised hand-centred receptive fields	174
5.9.1	Method	177
5.9.2	Stimuli	177
5.9.3	Training and testing	177
5.9.4	Results	180
5.9.5	Discussion	181
5.10	Training the network with natural scenes	186
5.11	Methods	188
5.11.1	Visual stimuli	188
5.11.2	Model parameters	189
5.11.3	Training	189
5.11.4	Testing	191
5.12	Results	191
5.13	Discussion	199
6	The visual development of hand-centred receptive fields in a time-accurate model of VisNet trained with experimentally recorded human gaze changes	203
6.1	Introduction	205

6.2	Methods	209
6.2.1	Experimental Task	209
6.3	Differential model of VisNet	212
6.3.1	Model architecture	214
6.3.2	Activation	215
6.3.3	Trace Learning	216
6.3.4	Numerical Simulation	218
6.3.5	Preprocessing of stimuli for VisNet	218
6.3.6	Training and Testing	220
6.4	Results	221
6.5	Discussion	227
7	The potential role of proprioception in the development of hand-centred representations	231
7.1	Introduction: Overarching network architecture of hand-centred representations	232
7.1.1	Overarching network architecture of hand-centred representations	235
7.2	The development of visual neurons representing the locations of visual objects in a body (trunk) centred frame of reference	238
7.2.1	Body-centred representations of visual targets in the brain	239
7.2.2	Hypothesis	242
7.2.3	Model Architecture	244
7.2.4	Analysis	248
7.2.5	Training the Network	252
7.2.6	Testing the Network	255
7.2.7	Results	257
7.2.8	Summary: Body-centred representations	264

7.3	The development of visual neurons representing the hand-centred locations of objects through combining inputs from body-centred visual neurons and proprioceptive signals representing the body-centred position of the hand	266
7.3.1	Stimuli and Methods	268
7.3.2	Model architecture and parameters	268
7.3.3	Training	271
7.3.4	Testing	274
7.3.5	Results	274
7.3.6	Discussion	281
8	Discussion	285
8.1	Summary of main findings	285
8.2	Future work	293
	References	304

List of Figures

1.1	Basic schematic of the ventral and dorsal visual pathways.	3
1.2	Basic schematic of the visuomotor pathway.	10
1.3	Parcellation IPS.	11
1.4	Parietal subregions in the macaque and the human brain.	12
1.5	Computing the motor error in a common reference frame.	15
1.6	Hand-centred receptive fields.	29
2.1	Architecture of a competitive network.	47
2.2	VisNet architecture.	51
3.1	CT learning hypothesis	68
3.2	Image sequences	74
3.3	Response profiles of three neurons in the top (fourth) layer of VisNet before training.	77
3.4	Response profiles of same three neurons in the top (fourth) layer of VisNet after training.	78
3.5	Information analysis for the first test case	81
3.6	Image of the hand with 8 target locations.	84
3.7	Information analysis for increasing number of target locations	85
3.8	Information analysis for increasing number of target locations using SOMs	90

3.9	Information analysis for all densities using competitive network with further reduced connectivity	94
3.10	Information analysis for all densities using SOMs with further reduced connectivity	95
4.1	Image sequences	107
4.2	Response profiles of three neurons in the top (fourth) layer of VisNet before training.	111
4.3	Response profiles of same three neurons in the top (fourth) layer of VisNet after training.	112
4.4	Information analysis for the first test case	113
4.5	Information analysis for different training epochs	117
4.6	Information analysis for simulations with larger saccades	120
4.7	Image of the hand with 10 target locations.	122
4.8	Information analysis for simulations with increased number of target locations	123
4.9	Information analysis comparing competitive networks with SOMs	129
4.10	Information analysis comparing competitive networks with SOMs	130
4.11	Information analysis with diluted connectivity	132
5.1	Training stimuli showing the hand with a pair of target locations presented simultaneously.	145
5.2	Testing stimuli.	146
5.3	Single cell information analysis.	148
5.4	Multiple cell information analysis.	149
5.5	Realistic training images with natural objects and textured backgrounds. .	160
5.6	Testing with a plain background and one textured stimulus.	161
5.7	Response profiles of six neurons in the top (fourth) layer of VisNet before training.	164

5.8	Response profiles of same six neurons in the top (fourth) layer of VisNet after training.	165
5.9	Information analysis for the six test hand-centred locations.	168
5.10	Tracing back the synaptic connections from a trained output cell to the input Gabor filters.	170
5.11	Pool of objects used for training and testing.	178
5.12	Reduction in the shape selectivity of neurons as the network is trained on increasing numbers of natural objects presented in the same hand centred location.	181
5.13	Realistic training images used for simulations with the hand presented against natural backgrounds during training.	190
5.14	Testing with a plain background and one textured stimulus.	192
5.15	Response profiles of five neurons in the top (fourth) layer of VisNet before training.	194
5.16	Response profiles of same five neurons in the top (fourth) layer of VisNet after training.	195
5.17	Information analysis for the five test hand-centred locations.	197
5.18	Tracing back the synaptic connections from a trained output cell to the input Gabor filters.	198
6.1	Gaze changes during one of the trials and sample of preprocessed frame . .	213
6.2	Response profiles of three neurons in the top (fourth) layer of VisNet before training.	223
6.3	Response profiles of same three neurons in the top (fourth) layer of VisNet after training.	224
6.4	Information analysis for the three test hand-centered locations	225

7.1	Overarching architecture of the neural network model incorporating proprioceptive signals.	236
7.2	Representing the location of the visual target in body-centred coordinates.	239
7.3	Schematic examples of eye and head movements while an object is stationary.	254
7.4	Schematic example of four gaze configurations during testing.	256
7.5	Firing rate responses during testing.	259
7.6	Population analysis of the frame of reference of the output neurons.	260
7.7	Population analysis of the receptive field size and location of the output neurons.	261
7.8	Schematic examples of hand-object configurations using for training and testing.	272
7.9	Feed-forward synaptic weight matrices from the proprioceptive hand position input layer and the visual target position input layer onto the intermediate combination layer before and after training.	275
7.10	Firing rates of cells in the combination layer before and after training.	276
7.11	Feed-forward synaptic weight matrix from the combination layer to the final output layer before and after training.	277
7.12	Firing rates of the final output layer before and after training.	278

List of Tables

3.1	Network dimensions	72
3.2	Lateral inhibition parameters	72
3.3	Sigmoid parameters	73
3.4	Lateral inhibition and excitation parameters for the SOM	89
3.5	Network dimensions	93
4.1	Network dimensions	104
4.2	Lateral inhibition parameters	104
4.3	Sigmoid parameters	105
4.4	Scaled up retina: Network dimensions.	119
4.5	Distribution of responses as the number of spatial locations is increased. . .	125
4.6	Sigmoid parameters	127
5.1	Training stimuli consisting of object pairs.	143
5.2	Network dimensions	158
6.1	Parameters used in the computer simulations.	217
7.1	Parameters of self-organising model of body-centred visual representations	248
7.2	Parameters of self-organising overarching network model of hand-centred visual responses	270

The brain alone remained a cardinal necessity. Only one other part of the body had a strong case for survival, and that was the hand, “teacher and agent of the brain”.

—H.G. Wells, *The war of the worlds*.

1

Introduction

The aim of this thesis is to use artificial neural networks to provide a theoretical account of the development of neurons in the primate brain that encode the location of visual targets in a hand-centred frame of reference. The self-organising neural network models presented in this thesis learn to represent the location of objects in hand-centred coordinates through visually guided learning using biologically plausible, unsupervised learning mechanisms. This chapter begins with an overview of the main components and processing stages of the primate visual system. Subsequently, we present a summary of the relevant research regarding the physiological properties and potential behavioural role of hand-centred visual representations. An overview of previous modelling work relevant to the development of non-retinal reference frames is also reviewed. Lastly, an overview of the modelling work explored in this thesis is provided.

1.1 The primate visual system: A tale of two streams

Ungerleider and Mishkin (1982) suggested that visual information was processed through mainly two independent and highly specialized pathways; the ventral and the dorsal visual pathways. The ventral visual pathway (sometimes characterized as the ‘What’ pathway) would project to the temporal cortex and was thought to be mainly involved in processing visual features that were useful for facilitating object identification and recognition (e.g. colour and shape). On the other hand, the dorsal visual pathway (sometimes characterized as the ‘Where’ or ‘How’ pathway) would project to the parietal cortex and would mainly contain cells responding to motion and velocity in order to help representing the location of objects in space.

However, after about a decade, this strict functional characterization of the visual streams was shown to be insufficient to account for some of the evidence coming from lesions studies and neurophysiological research (Milner and Goodale, 1993). A few years later, Goodale and Milner (1992) proposed a new distinction between the two pathways that was described more in terms of the purpose of their outputs rather than differences in the specialization of the visual inputs. In other words, both pathways would process visual information about the shapes and location of objects, however they would process and represent this information differently given that they were serving different functional purposes. Under this new framework the role of the dorsal stream would be characterized as ‘vision for action’, while the role of the ventral stream would be ‘vision for perception’.

Even though this distinction has been further challenged (Glover, 2004; Himmelbach and Karnath, 2005), there seems to be broad consensus and overall acceptance of this anatomical and functional characterization (Milner and Goodale, 2008). However, it is also agreed that this subdivision is less strict than originally proposed and there is actually significant intercommunication between the pathways (Takemura et al., 2015). In Figure 1.1 we find a simplified schematic representation of the two streams.

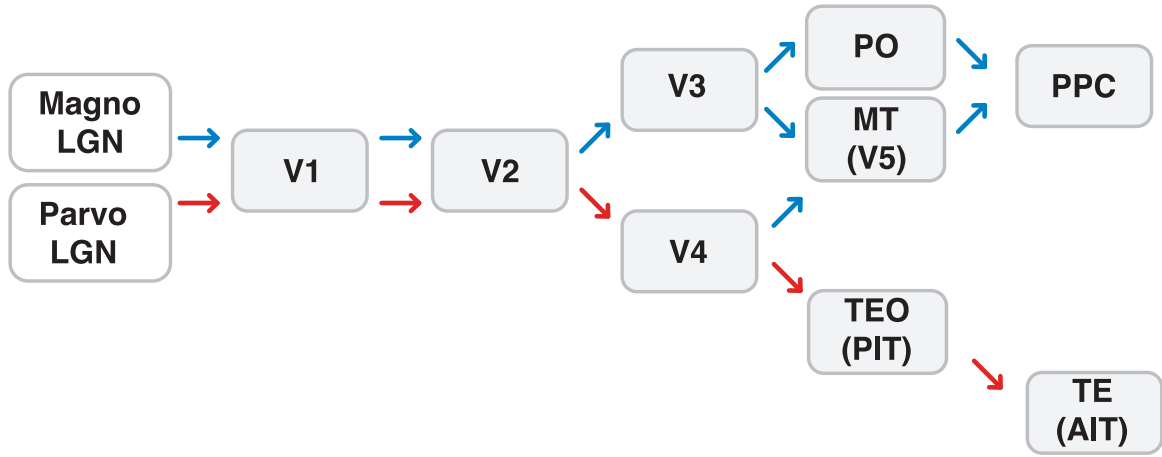


Figure 1.1: Basic schematic of the ventral and dorsal visual pathways. The figure illustrates the main components of visual processing along the ventral and dorsal visual pathways. The red lines correspond to the main afferent projections of the ventral stream and the blue lines represent the afferent main projections in the dorsal stream. This depiction constitutes a simplified abstraction of the visual processing stages and their connectivity. Further details about the subdivisions and connectivity in these areas can be found in Felleman and Van Essen (1991) and Kravitz et al. (2011, 2013).

1.1.1 Vision for perception: From the eye to the cortex

The eyes receive light from the external world and channel it into the retina. The retina then converts this photon energy into electrochemical signals that travel through the optic nerve into the lateral geniculate nucleus (LGN) (Kandel et al., 2000). The LGN is composed of 6 layers that preserve the alignment of the retinotopic receptive fields (Felleman and Van Essen, 1991). Layers 1 and 2 are mostly composed of magnocellular cells that have monochromatic responses, with low spatial resolution and high temporal resolution, while layers 3 to 6 are mostly composed of parvocellular cells that are more sensitive to color and possess a higher spatial resolution and lower temporal resolution (Ferrera et al., 1994).

The primary visual area (V1), also known as the ‘striate cortex’ (the rest of the visual cortex is known as the ‘extrastriate cortex’), is located at the back of the occipital lobe and receives sensory inputs from both the magnocellular and parvocellular cells of the LGN. As most cortical areas, V1 can be divided into 6 layers, which can be further subdivided into sublaminae. The distinction between the magnocellular pathway and the parvocellular seems largely maintained in V1 with laminae $4A$ and $4C_\beta$ receiving projections mostly from the magnocellular cells and laminae $4B$ and $4C_\alpha$ receiving projections mostly from parvocellular cells (Fitzpatrick et al., 1985). It was originally thought that the ventral and dorsal streams were simply an extension of the parvocellular and magnocellular processing pathways (Livingstone and Hubel, 1984), however, it has been shown that the ventral pathway receives inputs from both systems (Merigan and Maunsell, 1993).

V1 is composed of ‘simple’ cells and ‘complex’ cells that are sensitive to the orientation and direction of bars and edges (Hubel and Wiesel, 1962, 1968). The V1 simple cells compute a weighted sum of the inputs, generally modelled as ‘Gabor-filters’, while the complex cells sum the outputs from a subset of simple cells with similar properties but for different positions or phases (Lennie, 2003).

V1 projects primarily to V2, but also projects to other subcortical regions, including sending a feedback signal to the LGN (Felleman and Van Essen, 1991). V2 neurons are also tuned to the orientation and direction of edges, but they are also sensitive to subjective contours (Ffytche and Zeki, 1996; Von der Heydt et al., 1990). V2 cells also have larger receptive fields than V1 cells (Gattass et al., 1981). Area V2 is organized in an alternating pattern of thin and thick stripes. The ‘thick stripes’ in V2 seem to receive information relevant to motion from the 4B laminae in V1 and project mostly to V3, while the ‘thin stripes’ and ‘inter-stripes’ receive their input from layers 2 and 3 in V1 (Shipp and Zeki, 2002).

As we continue within the ventral hierarchy, neurons in V4 seem to respond to increasingly complex combinations of angles and curvatures (Pasupathy and Connor, 2001, 2002) as well as being selectively tuned to the perceived colour (Zeki, 1973). Area V4 receives its inputs predominantly from V2, however it also receives connections from V3 and MT. Most of the outputs from area V4 project to the inferior temporal cortex (IT).

Within IT we continue to find a hierarchical gradient of selective responses to increas-

ingly complex object features. In the posterior inferior temporal cortex (PIT/TEO) we find cells with selectivity to simple forms and shapes (Kobatake and Tanaka, 1994) and as we move anteriorly along IT, at latter stages of the processing hierarchy, we find cells with increasing complexity in their response properties within the anterior inferior temporal cortex (AIT/TE). Single-unit recordings have found cells in AIT/TE that respond maximally to objects or faces (Gross, 2008; Ito et al., 1995). Moreover, these responses seem to be invariant to the size of the stimuli and retinal location.

Up to this point we have focused primarily on the hierarchical anatomical connectivity of the ventral visual stream. In the following subsection we will put more emphasis on the connectivity and processing stages of the dorsal visual pathway.

1.1.2 Vision for action: The dorsal stream

The dominant theories of visual processing in the brain seem to agree on the anatomical division of two main processing streams with a certain degree of functional specialization. While the ventral stream is thought to be predominantly engaged in perceptual processes and invariant object recognition, the dorsal stream seem to be predominantly involved in the visual guidance of movements (Milner and Goodale, 2008). However, it is important to notice that there is less consensus on the different stages of processing along the dorsal stream and their functional roles.

The dorsal stream was originally viewed as an occito-parietal circuit that received most

of its input from the magnocellular layers in LGN (Livingstone and Hubel, 1984). This pathway was initially proposed to mostly process information about the motion of the input features. These motion representations would be processed in a layered hierarchy of processing stages, analogous to the hierarchical stages of processing in the ventral stream. The inputs provided by the magnocellular pathway would project to area V1, and motion-related information would be projected to V2. The hierarchy of motion processing would continue with significant projections from V2 and V3 to Area MT (medial temporal), where cells have been reported to be selective to a variety of features of 2D motion (Maunsell and Van Essen, 1983). Cells in MT will then project to the medial superior temporal region (MST) where more detailed and complex representations of 3D motion and optic flow are formed (Duffy, 1998; Graziano et al., 1994a). Subsequently, this information would continue to be processed along different regions of the posterior parietal cortex (PPC).

However, more recently, it has been suggested that the dorsal stream may be subdivided into at least three different processing streams (Kravitz et al., 2011). These further subdivisions share some of the known occipito-parietal connectivity of the initial stages of the dorsal stream. As we advance through the parietal areas we would then find an anatomical and functional division into three major pathways:

- ***The parietoprefrontal pathway:*** Composed mainly of parietal areas LIP, VIP, MT and MST and prefrontal areas 8A and 46. This processing pathway would be mostly engaged in spatial working memory and top-down control of eye movements.

- ***The parietopremotor pathway:*** This pathway is composed by two separate tracks. The first receives most of its input from areas V6A and MIP and projects mainly to dorsal premotor areas. While the second arises mainly from VIP and areas in the rostral (or anterior) inferior parietal lobe (i.e. PFG/PF, Pandya and Seltzer (1982) or 7a/7b, Lewis and Van Essen (2000)) projecting to the ventral premotor areas. The components of the parietopremotor pathway are mostly involved in visual guidance of movements, representation of peripersonal space and coordinate frame transformations.
- ***The parietomedial temporal pathway:*** This pathway connects the caudal inferior parietal lobe (cIPL) with the medial temporal lobe (MTL). It is thought to be in charge of processing complex spatial information relevant for navigation and representations of extrapersonal space.

In this thesis we will focus primarily on the processing taking place along the parietopremotor pathway, which connects the occipito-parietal components of the dorsal stream to different regions of the posterior parietal cortex (PPC) and premotor areas. Furthermore, the PPC not only receives visual input from LGN and early visual areas, but additionally receives visual, somatosensory and auditory information from the superior colliculus through the pulvinar. In Figure 1.2 we show a scheme that emphasises the connections between the main components of visuomotor processing. For a recent review of parietofrontal processing see Caminiti et al. (2015).

Despite the intricacy and complexity in the connectivity, we also seem to find within the dorsal pathway a hierarchical processing of information (Perry and Fallah, 2014), similar to the ordered stages of layered processing in the ventral visual stream. As part of the visual processing required to facilitate the planning and execution of action, it has been hypothesized that one of the main functions of the parietopremotor pathway of the dorsal stream is to transform the representations of visual objects into a coordinate system that is more suitable to the effector of the movement (Flanders et al., 1992; Soechting and Flanders, 1992). Therefore, in this pathway we seem to find a hierarchy of coordinate systems where we begin with retinocentric or eye-centred representations in early visuomotor areas and as we move forward to later stages of processing we find other body-centred and body-part centred representations (Flanders et al., 1992; Lacquaniti et al., 1995; McIntyre et al., 1998; Soechting and Flanders, 1992).

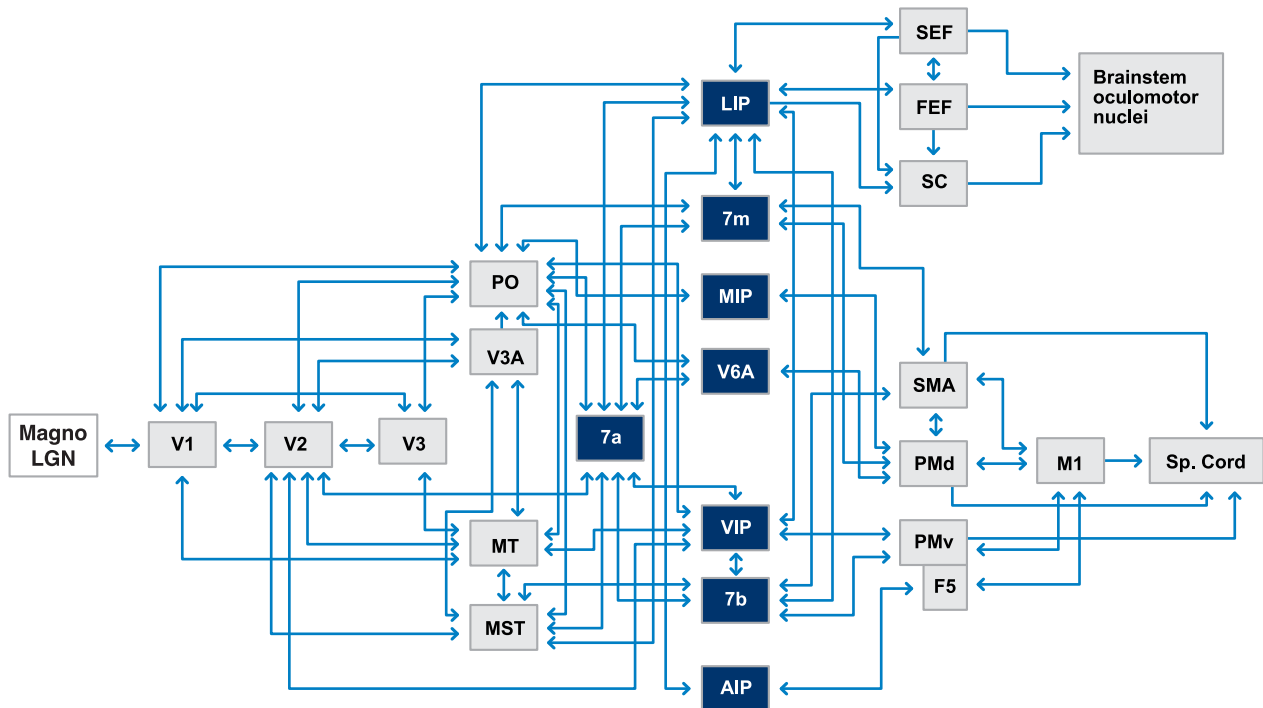


Figure 1.2: Basic schematic of the visuomotor pathway. This figure shows the main components of visuomotor processing in the macaque brain. Areas from the posterior parietal cortex are shown in the dark boxes. The arrows correspond to the main synaptic projections, including back-projections. Adapted from Graziano and Gross (1998).

1.2 Coordinate frames in the primate parietal areas:

The posterior parietal cortex and premotor areas

Since the pioneering work of Mountcastle et al. (1975), our understanding of the different functions of the PPC and its relevant role in visually guided actions has been expanded and extensively explored over the last few decades. As part of the parietopremotor pathway of the dorsal stream, this region receives various sensory inputs (predominantly from early

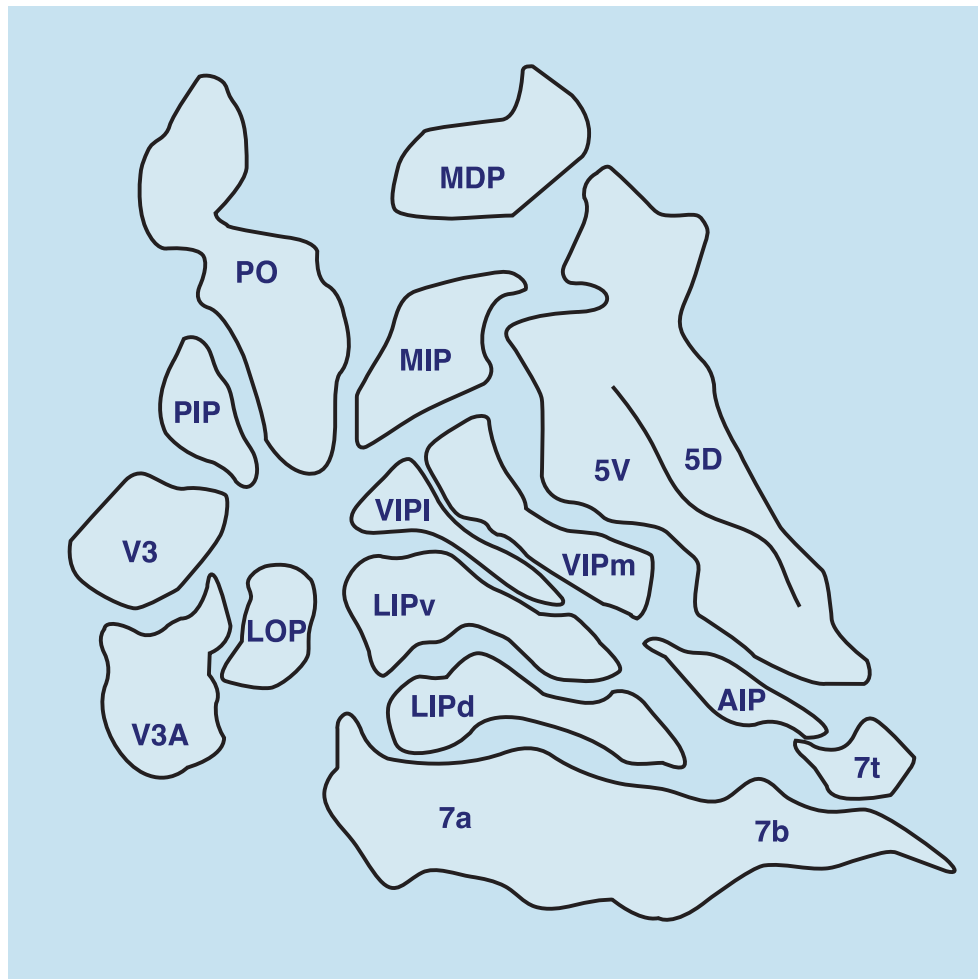


Figure 1.3: Parcellation of the IPS. The figure illustrates the main anatomical subdivisions of the intraparietal sulcus (IPS). Adapted from Lewis and Van Essen (2000).

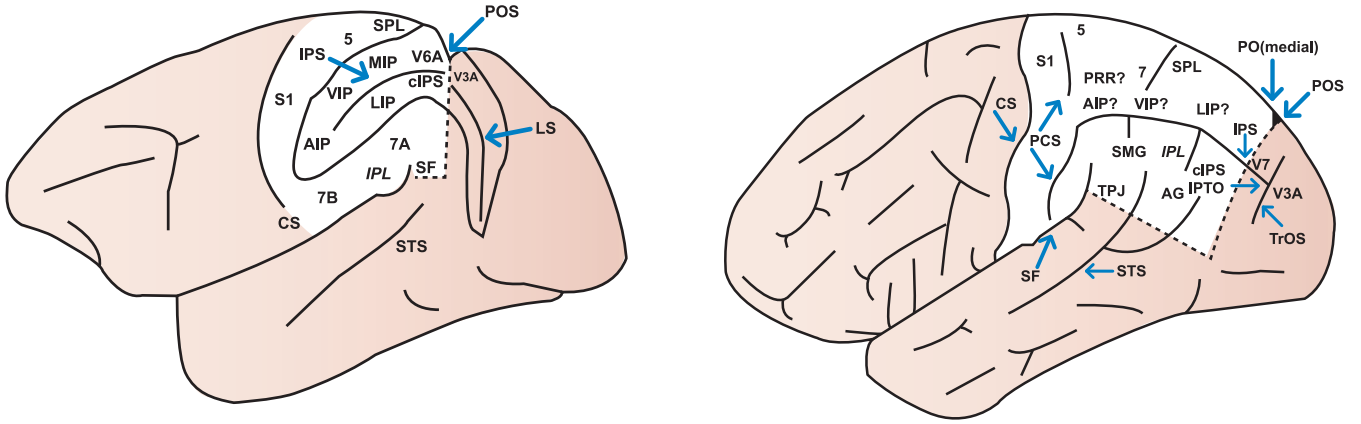


Figure 1.4: Parietal subregions in the macaque and the human brain. A lateral view of the macaque brain (left panel) and the human brain (right panel). Adapted from Culham and Kanwisher (2001).

visual cortical areas) and projects its outputs to several premotor and motor areas within the parietofrontal network (Wise et al., 1997). Single-unit recording studies have shown that both sensory and motor related activity is present in these different areas (Andersen et al., 1987) and it has been suggested that they play a key role in the sensorimotor transformations relevant for goal-directed movements.

In Figure 1.3 we show a schematic representation of the anatomical subdivisions of the intraparietal sulcus (IPS) (Lewis and Van Essen, 2000) and in Figure 1.4 we show a comparison between the human and monkey parietal lobes according to Culham and Kanwisher (2001). In addition to the anatomical subdivisions, numerous studies have shown that there is also functional specialization for specific actions and processes among these different areas. For example, the lateral intraparietal area (LIP) has been suggested

to play a major role in the production of saccades, helping to focus the gaze toward a visual target (Barash et al., 1991a,b), while the parietal reach region (PRR) seems to be more active when reaches are planned (Snyder et al., 2000). Anatomically, PRR includes portions of the medial intraparietal area (MIP) and area V6A. This region receives direct visual inputs from early visual areas and projects directly to premotor areas, providing one of the most immediate pathways of visual information into the premotor cortex (Galletti et al., 2001). The anterior intraparietal area (AIP) seems to be critically engaged in the visual control of grasping behaviour (Murata et al., 2000), and the ventral intraparietal area (VIP) has been suggested to be strongly involved in the direction of heading and the processing of self-motion (Bremmer et al., 2002).

As mentioned earlier, the different components of the parietopremotor pathway are thought to perform the necessary coordinate transformations required in the planning and guidance of motor behaviour (Milner and Goodale, 2008). For example, in order to reach with our hand for an object in space, the motor system could calculate a reach vector as the difference between the initial position of the effector and the position of the target or the endpoint of the movement (i.e. motor error) (Buneo and Andersen, 2006). To compute this reach vector or ‘motor error’, it has been suggested that the representation of the location of the hand and the location of the target could be directly subtracted using a common frame of reference (Andersen, 1995). In Figure 1.5 we show how the motor error could be computed using an eye-centred frame of reference or a body-centred frame of

reference. On the one hand, the retinocentric location of visual targets could be recoded from eye-centred coordinates into a body-centred frame of reference which would be a more natural frame of reference for the proprioceptive representation of the limbs (Buneo et al., 2009). This would mean that a body-centred visual representation of the target could be matched to the body-centred location of the hand in order to compute the difference in a common frame of reference ($M_B = T_B - H_B$). On the other hand, the position of the limbs could be recoded from a body-centred map into an eye-centred coordinate frame. This could also allow us to compute the motor error in a common eye-centred frame of reference ($M_E = T_E - H_E$). However, it is less clear how this eye-centred encoding could be later converted into a motor plan.

Alternatively, the motor error could be explicitly represented using a body-part centred frame of reference (in this case a hand-centred frame of reference). A reach vector in hand-centred coordinates would be equivalent to the motor error represented by the black line in Figure 1.5 ($M_H = T_H$).

1.2.1 Hierarchy of coordinate systems

It has been suggested that sensorimotor transformations along the parietopremotor circuit may occur in ordered stages where visual information is initially represented in a reference frame of the receptor (e.g. retinotopic) and later transformed gradually into representations that would ultimately be encoded in the reference frame of the effector (Flanders et al.,

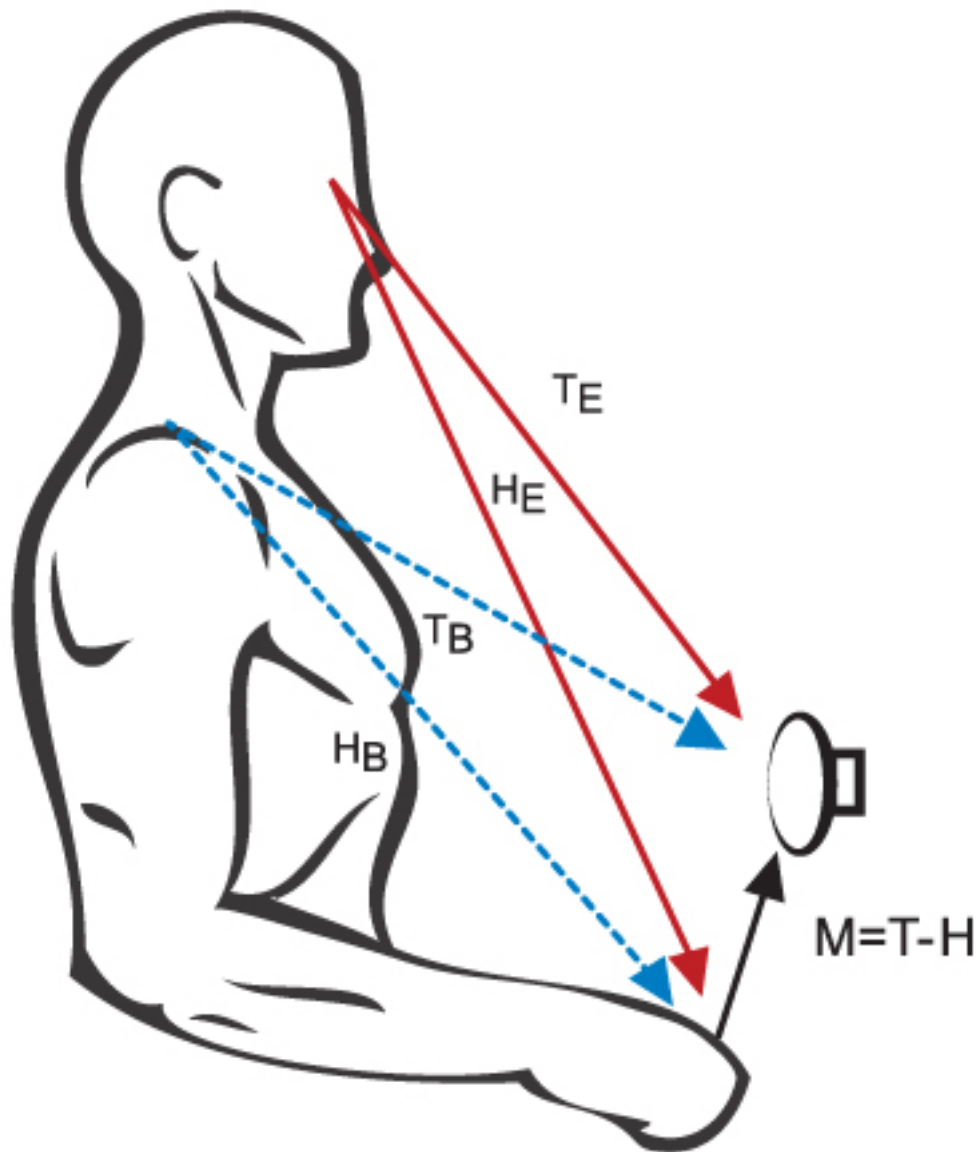


Figure 1.5: Computing the motor error in a common reference frame. The motor error (M) can be efficiently computed as the difference between the position of the target (T) and the position of the effector of the movement (e.g. the hand, H) using a common frame of reference. As shown in the figure this vector can be computed using an eye-centred representation of the hand and the target (i.e. red lines) or in body-centred coordinates (i.e. blue lines). For example, if the hand is not visible, we can compute the motor error using a body-centred representation of the position of the hand and subtract it from a body-centred representation of the visual target. Alternatively, the location of the hand could be recoded into eye-centred coordinates.

1992; Soechting and Flanders, 1992). Neurons in the posterior parietal cortex and adjacent areas seem to combine visual information with positional signals representing the location of the eyes, head and limbs to develop a variety of non-retinal representations (Snyder et al., 1998).

The position of a visual target in space has been shown to be computed in different egocentric frames of reference along the PPC. In single cell recordings, we generally find a gradient of responses from predominantly eye-centred representations in early visuomotor areas to different non-retinal egocentric representations in more parietofrontal areas. For example, head-centred (Duhamel et al., 1997; Galletti et al., 1993), body-centred (Brochier et al., 1995; Lacquaniti et al., 1995) and hand-centred representations (Bremner and Andersen, 2012)).

For example, in the context of reaching, it was found that as part of the initial stages of the visuomotor processing for reaching to a visual target the PRR encoded the reach vectors in eye-centred coordinates (Batista et al., 1999). As the recording sites are moved towards cortical area 5 in the superior parietal lobe (SPL), intermediate representations of both eye and hand-centred coordinates were reported, with the proportion of neurons coding the reach vectors in a purely hand-centred reference frame increasing nearer the cortical surface of area 5 (Buneo and Andersen, 2006) and predominantly in area 5d (Bremner and Andersen, 2012).

1.2.2 Gain modulated responses and heterogeneous representations

A number of neurophysiological studies have also reported a variety of cells in different visuomotor regions of the posterior parietal cortex that seem to be mainly eye-centred but that are also gain modulated by a variety of positional signals from different parts of the body. These populations include cells gain modulated by eye-position (Andersen et al., 1985; Galletti et al., 1995), head position (Brochier et al., 1995) as well as mixed representations and compound gain fields (Batista et al., 2007; Chang et al., 2009; Chang and Snyder, 2010; McGuire and Sabes, 2011). It has been suggested that these gain modulated cells could be used to compute the reach vectors using a common eye-centred reference frame (Cohen and Andersen, 2002). This means that an explicit representation in a body-centred spatial map might not be necessary given that this information could be directly deduced from the activation of these gain modulated cells.

Another theoretical approach has proposed that most cells in the PPC may in fact be mostly composed of mixed and heterogeneous representations that cannot be strictly classified under a specific frame of reference and from which reach vectors could be still flexibly computed (Pouget et al., 2000, 2002).

The heterogeneity of cell response properties found in the PPC can be due partly to methodological reasons (Bremner and Andersen, 2012) as well as the large cortical areas from which these cells responses are collected. Additionally, it can also be argued that

the variety of representations and different types of transformations are not necessarily mutually exclusive, in fact redundancies in motor planning could be exploited by the motor system (Latash, 2012).

Nevertheless, the modular gradient of responses going from eye to body-centred and body-part centred representations strongly suggests that at least in some parts of the parietopremotor pathway the location of visual targets are encoded into a body-centred or body-part centred frame of reference. Moreover, there are subsets of cells that seem to explicitly represent reach vectors and target locations in body-centred and body-part centred reference frames (Bremner and Andersen, 2012; Brothie et al., 1995; Duhamel et al., 1997; Galletti et al., 1993; Lacquaniti et al., 1995).

In this thesis we will present a number of self-organising hypotheses of how the brain could be forming explicit representations of the location of visual targets using a hand-centred frame of reference. These representations would provide information of the location of objects with respect to the hand, which could play a role in hand-object manipulations.

1.2.3 Visuospatial and visuomotor deficits associated with lesions in posterior parietal areas

Unilateral neglect is a neuropsychological impairment that occurs approximately in 50% of stroke patients (Bickerton et al., 2011). This condition is commonly caused by damage in the right hemisphere of the brain, typically within the posterior parietal cortex

(PPC) (Bowen et al., 1999). One of the most intriguing and distinctive deficits in these patients is their inability to represent and respond to stimuli appearing on their contralateral side (Heilman and Valenstein, 2000). Since some of the early clinical accounts of neglect (Jackson, 1875) a large collection of different symptoms have been reported (for a review see Buxbaum et al., 2004). For example, neglect can be specific to how space is represented in particular tasks (Horner et al., 1989), or it may appear in contrasting sensory modalities. There have been reports of patients who neglect the space near their bodies (i.e. peripersonal space) but show almost no deficit for space beyond their reach (i.e. extrapersonal) (Halligan and Marshall, 1991). Conversely, there are reports of patients who fail to represent space far away from their bodies but not near (Vuilleumier et al., 1998). The heterogeneity of the observed symptoms in neglect can allow us to gain an insight on how we encode visuospatial information. For example, there seems to be substantial evidence and broad agreement that peripersonal space is processed differently from visual-spatial information in extrapersonal space (Brockmole et al., 2013). Imaging studies have suggested that activity within the ventral visual pathway is associated with far space processing, while the activation of dorsal areas is mostly related to near space processing (Weiss et al., 2000). Dissociations in how the brain processes visuospatial information in either the ventral or dorsal systems seem to partially account for some of the observed variability between the patients. In other cases though, the distinction between ventral and dorsal representations of space seems less clear. For example, neglect may

appear in the contralesional side with respect to the body (i.e. egocentric) as well as in the contralesional side with respect to individual objects (allocentric or object-centred). In principle, this suggests another important dissociation between egocentric and allocentric deficits (Humphreys et al., 2013). This dissociation seems to be consistent with lesion studies showing that allocentric deficits appear in posterior-temporal lesions (related to more ventral injuries), while egocentric deficits correlate with injuries in the perisylvian network (Karnath and Rorden, 2012). Nevertheless, it is not entirely clear whether the dissociations between allocentric and egocentric spatial encoding are necessary or sufficient to account for the variability between these symptoms. There are other theoretical accounts that suggest that both egocentric and allocentric deficits in fact reflect different aspects of the same disorder and can even be accounted within the same spatial frame of reference (Pouget and Driver, 2000).

Extinction is another related deficit in which patients are unable to detect a contralesional stimulus when it is presented simultaneously with a second stimulus in the ipsilesional side. These patients can only detect a stimulus when it is presented alone (for a review on the relation between neglect and extinction in different sensory modalities see Brozzoli et al. (2006)). Studies on patients with tactile extinction have shown interesting crossmodal effects that provide an insight into the underlying mechanisms of multisensory representation of peri-personal and particularly peri-hand space (di Pellegrino et al., 1997; Làdavas et al., 1998, 2000). For example, it has been reported in patients with tactile

extinction that the detection of a contralesional tactile stimulus varies as a function of the proximity of a visual stimulus in the ipsilesional side (Farnè et al., 2005). In other words, the closer a visual stimulus is presented to the right hand, the tactile extinction is more severe.

Apraxia is another type of complex disorder that can be caused by lesions in posterior parietal and premotor areas (for a review see Petreska et al., 2007). There are six main types of apraxia that are associated with effects on the hands and arms (Wheaton and Hallett, 2007):

- ***Limb-kinetic apraxia:*** Deficits in grasping and fine movements of fingers.
- ***Conceptual apraxia:*** Deficits associated to the representation and usage of tools.
- ***Ideational apraxia:*** Inability to properly sequence the steps required to solve a task.
- ***Verbalmotor dissociation apraxia:*** Deficit to responding accordingly to verbal commands signalling the execution of movements.
- ***Tactile apraxia:*** Selective deficits involving touch.
- ***Ideomotor apraxia:*** Problems in imitating and pantomiming the appropriate use of tools. This may include deficits in orientation and other spatio-temporal properties in movement execution.

Studies in patients with brain lesions in parieto-premotor areas are often used as a way of providing an insight into the main components and properties of visuomotor and visuospatial processing. The variety of symptoms allows us to identify the role and contribution of the different ways in which we represent our bodies and our surroundings.

1.3 Hand-centred representations in the brain

Our hands play a crucial role in the way we perceive, explore, manipulate and interact with the world. Unsurprisingly, the hands concentrate a higher density of sensory receptors than rest of the arms and torso (Penfield and Rasmussen, 1950). Not only do the hands provide the brain with a large array of sensory information about our surroundings (i.e. shape, texture, size, temperature, weight etc.) but they enable us to develop an impressive dexterity that allows us to reach and manipulate objects and the use of tools.

The development of hand function has played a significant role in the evolution of hominids. In fact, some of the morphological changes of the hand seem to be closely related to changes in cranial capacity and brain development (for a review see Jones and Lederman, 2006). Consistent with this, cortical regions devoted to sensorimotor processing like the PPC became further differentiated and expanded in anthropoid primates and specially most enlarged in humans (Kaas, 2004).

In order to move our hand towards a particular object or target location we need to represent the information about the location of the hand and the location of the object. As

we discussed earlier, the brain seems to represent spatial information about the location of objects in a variety of coordinate systems or frames of reference. These representations seem to include a hand-centred spatial map, which is thought to be relevant for object directed actions involving the use of the hand.

1.3.1 Encoding reach vectors in hand-centred coordinates

As we mentioned earlier, it has been suggested that sensorimotor transformations along the parieto-frontal circuit may occur in ordered stages where sensory information is initially represented in the reference frame of the receptor (e.g. retinotopic) and later transformed gradually into representations that would ultimately be encoded in the reference frame of the effector (Flanders et al., 1992). Compatible with this view, it has been found in primates, that as part of the initial stages of the visuomotor processing for reaching to a visual target, early visuomotor areas including PRR seem to be encoding the reach vectors in eye-centred coordinates (Batista et al., 1999). However, as the recording sites are moved towards cortical area 5 in the superior parietal lobe (SPL), intermediate representations in both eye and hand-centred coordinates were reported, with the proportion of neurons coding the reach vectors in a purely hand-centred reference frame increasing nearer the cortical surface of area 5 (Buneo and Andersen, 2006). Moreover, a variety of cell responses have been reported also in the dorsal premotor area (PMd), including limb-centred cells (Batista et al., 2007) as well as cells that seem to encode the spatial relation of the eye,

the hand and the target (Pesaran et al., 2006).

More recently, Bremner and Andersen (2012) reported that cells in Area 5d encoded the locations of a visual target using an explicit hand-centred frame of reference at a single neuron level. The authors used a center-out task in which they presented a number of hand-target configurations arising from different combinations of 4 different gaze fixations, 4 different hand positions (-20° , -10° , 0° , or 10°) and 4 different target positions (-20° , -10° , 0° , or 10°), which gave a total of 64 possible combinations. The population analysis showed that a hand-centred frame of reference was the predominant representation in this brain area. It was found that 41% of the cells were tuned to a specific hand-object configuration, and in most cases (66%) the representation was maintained across the different gaze fixations, hand positions and target positions (i.e. an invariant or inseparable T+H response). This means that these cells would fire maximally irrespective of where on the retina this particular hand-object configuration appears and regardless of the target location and the hand location with respect to the body. In other words, these cells are encoding the location of the visual target with respect to the hand, invariant to hand position and the retinal location of the hand-target configuration.

Interestingly, a more detailed inspection of the temporal progression of the neuronal responses in area 5d suggests that these cells may have in fact interesting dynamical properties. Before a target is presented, the cells seem to represent the location of the hand with respect to the gaze location and after a target is presented the cells switch their responses

to encode the position of the target in hand-centred coordinates (Bremner and Andersen, 2014). However, as the authors recognized, given that it is not possible to know if the animals were computing a reach plan towards the location of the gaze during the fixation period, the cells could be always effectively representing a current reach plan.

1.3.2 Body-part centred representations of visual space: Hand-centred receptive fields

Non-human primate studies

A number of neurophysiological studies in both human and non-human primates have also reported other regions of the brain that seem to show selective responses for regions of space surrounding parts of the body like the hand. These hand centred receptive fields seem to be anchored to the position of the hand and relatively unaffected by changes in gaze and hand position with respect to the trunk. However, unlike most of the studies reported in Section 1.3.1, much of this research did not present a thorough analysis to unequivocally disentangle the underlying frame of reference used by these cells. Furthermore, these responses were investigated using completely different tasks to the center-out reach tasks performed by Bremner and Andersen (2012). In fact, as we will discuss, in most of these studies the participants were not required to actively interact with the stimuli (e.g. perform a reach toward the target location) but actually the hand remained mostly passive, while different stimuli were presented near the hand and the corresponding responses were recorded.

Different electrophysiological recordings in macaques have reported neurons with localised and selective responses to stimuli shown near the body or a part of the body. For example, a single-unit recording study in cortical area 7 by Hyvärinen and Poranen (1974) showed neurons with selective responses when sensory stimuli were presented in localised regions near the body. Most of these cells were active during object-directed motor behaviour and could be activated by both visual and tactile stimulation.

A few years later, Rizzolatti et al. (1981) reported similar findings in a single cell study of the periarculate cortex. In this study, two types of cell responses were reported. A subset of cells had large receptive fields that responded to visual stimuli presented far from the animal and a subset of cells that responded to visual stimuli presented near the skin. These neurons had spatially congruent bimodal receptive fields (i.e. visual and tactile) and their visual responding regions could be found anchored to different parts of the body (e.g. around the hand, mouth and face) and their response properties did not seem to change with eye movements (Rizzolatti et al., 1988). Additional recordings in area F2 and area F4 have similarly shown cells with body-part centred receptive fields (including hand-centred visual RFs) that represent peri-personal space independently of eye movements (Fogassi et al., 1996, 1999).

Duhamel et al. (1998) also reported neurons in the ventral intraparietal area (VIP) with bimodal (i.e. visual and tactile) and spatially congruent receptive fields (RFs). In this study however, the authors were also interested in identifying the underlying frame

of reference of these representations. During the task, the monkeys had to fixate to 9 different locations on a 3 by 3 grid. For each of these fixation points, a visual stimulus would be presented in different parts of the screen. The locations of the stimuli were organized on a 7 by 7 grid. After the analysis, the authors found subpopulations of cells with eye-centred representations while other cells seemed to encode the location of each stimulus in a head-centred frame of reference. It has been hypothesized that VIP could be representing an intermediate stage in the transformation of eye-centred to body-centred or body-part centred representations (Graziano et al., 2004).

Similarly, Graziano and others further investigated the properties of bimodal and multimodal neurons that seem to respond to the location of objects near the body. On a first study, a subset of cells in areas 6, 7b, VIP, and the putamen (Graziano and Gross, 1993) was shown to have tactile RFs aligned with a visual receptive field that extended from the skin to about 10 to 20 cm. Body-part centred, including hand-centred, visuo-tactile receptive fields have also been reported in the ventral premotor area (PMv) (Graziano, 1999; Graziano and Gross, 1998; Graziano et al., 1997, 1994b) in the polysensory zone (PZ) (Graziano et al., 2004). Some of these brain regions are monosynaptically interconnected and it has been suggested they could form part of a "defensive system" to avoid potentially dangerous objects (Graziano and Cooke, 2006; Graziano et al., 2004). Consistent with this notion, high levels of stimulation in some of these regions can elicit defensive or protective behaviours (Graziano and Cooke, 2006; Graziano et al., 2004).

The tactile receptive fields are evidently invariant to changes in posture and independent of retinal location, given that they are receiving signals from receptors in the skin. However, the visual receptive fields of these cells are also anchored to the hand. In other words, if the arm or the hand is moved, the visual receptive field moves together with the hand. More importantly, the visual receptive field remains the same independently of where the eyes are moving. Even though these studies did not systematically analyse the underlying frame of reference used by these cells, the response profiles of these neurons strongly suggest that they are representing the location of the visual targets using a hand-centred frame of reference.

Neurons in these regions seem capable of binding visual information of the space near the hand in order to explicitly represent specific spatial configurations of the hand and a visual target, without the target necessarily touching the skin. Such neurons effectively represent the locations of visual targets in a hand-centred frame of reference. These explicit representations of the space between the object and the hand, or a potential hand-centred reach vector, are thought to be useful for hand-object interactions. In this thesis we will show how artificial neurons may learn to develop visual receptive fields that are anchored to specific positions with respect to the hand. In Figure 1.6 we show a schematic representation of a localised hand-centred receptive field, based on the results reported by Graziano and Gross (1998) with macaques.

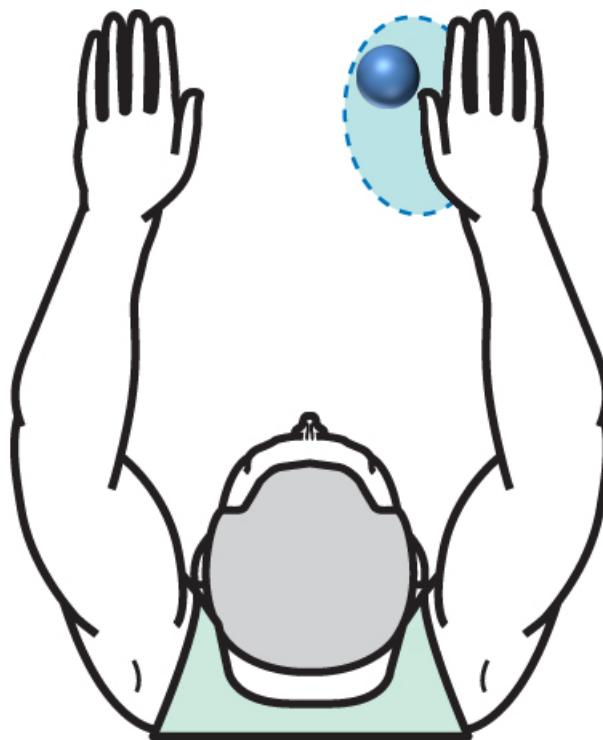


Figure 1.6: Hand-centred receptive fields. Schematic representation of a localised hand-centred visual receptive field. The image is based on the visuo-tactile responses reported by Graziano and Gross (1998).

Human studies

A number of neurophysiological and behavioural studies with human subjects have similarly shown evidence of hand-centred encoding of the location of visual objects. For a comprehensive review on hand-centred representations see Makin et al. (2008) and Holmes (2013). For example, Makin et al. (2009) applied transcranial magnetic stimulation (TMS) to the primary motor cortex of the participants in a task in which they were required to push a button with their index finger when a signal was given. During the trials, a distractor ball would fall near the participant's hand. A single TMS pulse was applied between 40 and 120 ms after the ball was presented to elicit motor evoked potentials (MEP) in the hand muscles. The corticospinal excitability was used as a measure to assess the effect of the approaching ball. This study presented 7 different experiments that provided evidence of how visual spatial information is encoded in hand-centred coordinates. For example, it showed a significant suppression of corticospinal excitability when the ball was approaching near the hand as opposed to the ball falling away. Moreover, these hand-centred effects were tested across different eye, hand and target configurations. Other behavioural studies with humans have similarly suggested the existence of hand-centred and action-centred representations during a variety of tasks including pointing to target locations (Chieffi et al., 1999; Tipper et al., 1992), the planning of reaching movements (Gordon et al., 1994) and grasping actions (Brozzoli et al., 2009). Consistent with these findings, different neuroimaging studies have also provided evidence for the hand-centred

coding of visual space (Brozzoli et al., 2011; Gentile et al., 2011; Makin et al., 2007).

1.3.3 The roles of vision and proprioception in the localisation of the hand

Area 5 and other parieto-premotor regions seem to heavily rely on vision to encode the location of the hand. In some cases the visual input seem to override the proprioceptive representation when the two signals are conflicting. For example, it has been shown in non-human primates that occluding the vision of the arm reduces the way visual receptive fields shift together with changes in arm position (Graziano, 1999). Moreover in this same study, some neurons in PMv with visual receptive fields anchored to the arm, could remap their responses to the visible position of a fake stuffed arm (see Fig 2 in Graziano, 1999). In another study in Area 5, Graziano et al. (2000) explored in more detail the visual representation of the arm location. In this study the arm of the monkey was placed under a plate, keeping it out of sight while a realistic fake arm was located on top of the plate and made visible to the animal. This study showed that a visual sense of a realistic fake arm can also generate significant activation in Area 5 (25%) and mostly in MIP cells (35%), (Graziano and Botvinick, 2002; Graziano et al., 2000) when the seen position and proprioceptive location were congruent. In the same study, other objects of similar size to the arm but completely different shapes were placed on the plate (e.g. a white rectangle, a piece of fruit etc.). However, these objects did not seem to generate the same effect on

the cell responses.

Other studies with macaques have also shown that vision of the animal's own hand directly affects the firing responses of grasping neurons in premotor cortices (F1 and F5, Fadiga et al., 2013). There is also evidence involving visual limb position signals for direct visuomotor transformations between PRR and Area 5 (Buneo et al., 2002). Moreover, a portion of V6A neurons have been reported to have an increased activity exclusively for reaching in light conditions and strongly modulated by the sight of the arm (Bosco et al., 2010).

In humans, the superior parieto-occipital sulcus (sPOS), considered as an homologue of the PRR, responds significantly more during direct visual reaching which involves the view of the hand (Filimon et al., 2009). In fact, merely placing a hand near objects biases visual processing towards the magnocellular visual pathway (Derrington and Lennie, 1984), giving rise to a range of attentional effects in the visual representation of the space near the hand (Gozli et al., 2014, 2012).

Positional information of the location of the hand is integrated using inputs from different sensory modalities, including vision and proprioception. Generally, the information from the different sources is congruent helping to make our estimates more precise. The localisation of the hand is diminished when the visual and proprioceptive signals are incongruent. Many studies have explored the complex interaction of visual and proprioceptive signals at different stages of motor planning and the generation of reaching movements,

and it still remains unclear how and when visual and proprioceptive inputs converge in visuomotor processing (Graziano, 1999; Rossetti et al., 1995; Sober and Sabes, 2003, 2005).

It has been suggested that the hand location is optimally integrated using information from vision and proprioception. However, it has been shown that the visual signals seem to weigh more strongly than the proprioceptive representation when the two signals are in conflict. For example, Holmes and Spence (2005) used mirrors to weight the influence of visual and proprioceptive signals showing that the participants relied significantly more on the visual cue of the hand. Moreover, this effect was increased during active movements as well as with longer duration of visual exposure. However, in a more recent study using the MIRAGE system, it has been argued that subjects might first rely on a visual cue of the hand and with the passage of time rely more on their proprioceptive representations (Bellan et al., 2015).

Hand location, hand ownership and visuo-tactile interactions have been shown to be disrupted in normal human subjects using prisms, dummy hands, mirror-induced illusions and a variety of techniques that cause a conflict between the visual representation of the location of the hand and the proprioceptive representation of hand location (Maravita et al., 2002; Pavani et al., 2000). For example, an illusion box that provides the participants with images of their hand conflicting with the sensed position of the limb affected the awareness of the real hand (Newport and Gilpin, 2011). Differences in motor evoked potentials (MEP) in experimental tasks using a dummy hand in different orientations suggest that

visual information of the hand laterality is also used by the motor system (Makin et al., 2015). Moreover, hand-centred representations in PPC and premotor cortex have also been remapped to a prosthetic hand during a rubber-hand illusion (Brozzoli et al., 2012).

A number of behavioural experiments have similarly shown improvements in the accuracy of reaches and hand-object interactions when vision of the hand is available or a reduction in accuracy when visual feedback is removed (Desmurget et al., 1997; Holmes and Spence, 2005; Miall et al., 1995; Saunders and Knill, 2003, 2005; Sober and Sabes, 2003). As mentioned earlier (Section 1.2.3), studies in patients with contralateral tactile extinction seem to also suggest that a visual signal of the hand position is more relevant than the proprioceptive signals in affecting cross-modal extinction (Làdavas et al., 2000).

Cells encoding the position of a target in a hand-centred reference frame have been mostly reported in the parietal cortex and premotor areas. However, it is well known that the PPC and premotor areas not only receive bottom-up visual signals (Caminiti et al., 1996; Galletti et al., 1996), but they also receive afferent signals from the somatosensory cortex and top-down motor signals (Prevosto et al., 2011; Sakata et al., 1973). Experimental studies have shown that hand-centred cells in the PPC and Area 5 can maintain their firing properties when guided only by motor signals or proprioceptive signals in the absence of visual input of the hand and target (Buneo and Andersen, 2006, 2012). Moreover, the delayed-reach paradigm used in many of these studies (Batista and Andersen, 2001; Buneo and Andersen, 2006) showed that these neurons continued to maintain their

activity when the visual target disappeared from view. This suggests that the hand-centred representations may receive additional proprioceptive input signals specifying the position of the hand, as well as signals conveying a memory of the target location. These additional position signals could be used to update hand-centred representations in the PPC in the absence of visual input.

As mentioned earlier, a considerable number of studies have manipulated systematically the availability and congruency between different modalities, affecting and biasing in different ways the localization of our hands as well as reaching movements. Most of these studies suggest that the hand position can be significantly biased by vision, suggesting that visual information is weighted more strongly when the different signals are in conflict (Holmes and Spence, 2005). Conversely, other studies have suggested that proprioceptive signals might be dominant under different paradigms (van Beers et al., 2002) and at different moments (Bellan et al., 2015). To reconcile these findings, it has been suggested that the reliability and congruency of the sensory modalities in which the limbs and targets are perceived could play a role in determining the type of signal or representation that dominates. It is therefore generally agreed that visual and proprioceptive signals encoding the location of the hand may have distinct roles and weigh differently at different stages of goal-directed movements. When available, visual information of the hand configuration seems to be relevant for the encoding and initial planning of the reach vectors, while proprioceptive signals of the hand position seem to be relevant for transforming a reach plan

into the appropriate motor signals (Sarlegna and Sainburg, 2009).

The proprioceptive localization of the hand is generally more precise at distances closer to the shoulder (van Beers et al., 1998). Localizing the hand using purely proprioceptive signals requires constantly computing and combining the angles of the joints. The representations of the locations of the visual targets with respect to the hand must also be continually updated using the new computed position of the hand. Relying on processing of internal proprioceptive signals will introduce mild error in the estimated hand-centred locations of the targets as well as taking longer to compute. Moreover, different studies have shown that an internal estimation of the location of the arm starts to drift in the absence of a visual cue from the hand (Block, 1890; Craske and Crawshaw, 1975; Paillard and Brouchon, 1968; Wann and Ibrahim, 1992). For these reasons, other authors have also suggested that visually derived hand-centred representations, especially in PMv, might be useful for providing more rapid and accurate information of the hand-object configuration for the control of rapid actions (Makin et al., 2012).

In this thesis, we will initially explore a couple of alternative learning mechanisms by which hand-centred representations can be developed using a purely visual input representation of the hand and nearby target objects. Towards the end of the thesis we will present a theoretical hypothesis of how hand-centred representations could also be developed using a proprioceptive localisation of the hand.

1.4 Computational modelling of reference frame transformations and body-part centred representations

A variety of neural network models have been proposed to reflect the different stages of coordinate transformations and explain some of the response properties found in some neurons of the PPC and premotor areas. For example, Zipser and Andersen (1988) produced an influential artificial neural network model intended to represent a subset of cells in area 7a that were able to transform visual input from eye centred coordinates into head-centred coordinates. The model was composed of two populations of input neurons, one providing visual information of the target in eye-centred coordinates and the other eye-position signals. This input layer projected synaptic connections onto a hidden layer, and the hidden layer projected synaptic connections onto the final output layer. After training the network using backpropagation of error, hidden layer neurons developed visual receptive fields that were gain modulated by eye position, while the output layer was able to encode the location of the target in head-centred coordinates, as reported for cells in area 7a.

Similarly, Chang et al. (2009) incorporated a hand-position signal to the Zipser and Andersen (1988) neural network model, developing gain-modulated cells in the hidden layer and hand-centred cells in the output layer. The cells in the hidden layer of the model replicated the response properties of a subset of cells in PRR that were gain modulated by eye position and hand position. Also inspired by the Zipser and Andersen (1988) model, Xing et al. (2000) explored how different architectural combinations of input and output

layers could lead to a variety of reference frame representations and the integration of multimodal signals.

However, the necessity of a hierarchy of processing stages as well as the nature and role of gain fields in sensorimotor transformations has been challenged (Blohm and Crawford, 2009; Colby and Goldberg, 1999; Pouget and Sejnowski, 1997). Other computational approaches have suggested a different way of implementing these transformations more flexibly and robustly, with neurons encoding information in a mixture of representations (Blohm et al., 2009; Pouget and Sejnowski, 1997). For example, Pouget and Sejnowski (1997) showed that a hierarchy of processing stages in different frames of reference might not be required and produced a two-layered model showing that neurons behaving like basis function units could provide an immediate read-out of multiple frames of reference. This leads to the prediction that most cells in the PPC may in fact be mostly composed of mixed and heterogeneous representations from which reach vectors can still be flexibly computed (Pouget et al., 2000, 2002). Consistent with this view, some physiological reports have shown a mixture of cell responses at the different processing stages. For example, cells in the PRR, which were originally reported as being mostly eye-centred and part of the early stages of visuomotor transformations, include cells with a wide range of heterogeneous and idiosyncratic responses including hand-centred representations (Chang and Snyder, 2010).

Despite the computational advantages of these different theoretical efforts, most of these models use a version of the back-propagation of error algorithm (Werbos, 1974). These

approaches have been largely regarded as biologically implausible (Rolls, 2008) because the information to change the synaptic weights is not available locally in the presynaptic terminal and therefore cannot provide an account of how learning may operate in the cortex. Mazzoni et al. (1991) identified this limitation in the Zipser and Andersen (1988) model and made an effort to produce a more biologically plausible approach. This model used a different learning mechanism that would use a global error for the entire population instead of adjusting the synaptic weights of individual neurons using back-propagation of error. However, this learning mechanism would still require a supervised signal that provides information about the desired output and compares it with the actual output. Thus, this approach does not avoid some of the most fundamental theoretical limitations of supervised learning algorithms.

More recently a few self-organising approaches to the development of head-centred and body-centred representations have been put forward. For example, Spratling (2009) explored a temporal binding hypothesis for developing head-centred and body-centred cells. In this model a hierarchy of two separate conjunctive and disjunctive layers are used to develop head-centred and body-centred responses. In the first stage of processing head-centred output cells are developed from retinocentric representations and gain modulated signals representing the position of the eyes. The model then uses this head-centred output to learn combinations of different head-centred locations to develop output cells in a body-centred reference frame. More recently, colleagues from our research centre developed a self-

organising learning hypothesis for the development of head-centered visual representations (Mender and Stringer, 2013) incorporating eye position gain fields in the input layer and using a trace learning mechanism.

Nevertheless, we are still lacking fully unsupervised models that explain how localised *hand-centred* visual receptive fields can arise naturally in the different parts of the parieto-frontal circuit in a plausible self-organising way. Magosso et al. (2008, 2010) recently proposed a neural network model that learns to integrate visual and tactile peri-hand representations. However, this model already assumes the pre-existence of purely visual hand-centred cells and purely tactile hand-centred cells in the input layer. The authors simply combine these independent representations into an output layer using Hebbian learning in order to develop bimodal cells. Therefore, despite using a biologically plausible learning mechanism, this model does not provide a satisfying theoretical account for how these hand-centred representations are developed in the first place. Similarly, another connectionist approach has proposed that the integration of the different modalities is due to the correlation of the input representations (i.e. vision, touch and proprioception) (Botvinick et al., 1998). However this simplified model does not provide a self-organising account of the reference frame transformations required to integrate these different signals.

In this thesis we provide a number of self-organising theories of how *hand-centred* representations of visual space may develop through visually guided learning using artificial neural networks. Such hand-centred representations are thought to play a role in hand-

object interactions. Nevertheless, movement dynamics and motor control are complex processes that integrate a wide range of sensory and motor information in order to plan and control movement trajectories. This thesis does not examine the dynamics or required specifications of force, velocity and direction to produce the appropriate arm movements. Computational approaches to movement dynamics and motor control have been comprehensively discussed elsewhere (Bays and Wolpert, 2007; Miall and Wolpert, 1996; Scott, 2012; Thomaschke et al., 2012; Todorov and Jordan, 2002; Wolpert, 1997; Wolpert and Ghahramani, 2000).

1.5 Overview of the thesis: Self-organization of hand-centred representations

Recent neurophysiological research has shown that there are some cells that seem to represent reach vectors and target locations using an explicit hand-centred frame of reference (Bremner and Andersen, 2012). Similarly, multiple studies have found bimodal cells that seem to represent a localised region of visual space near the hand (i.e. peri-hand space). In both cases, these cells would respond to a visual object within the hand-centred receptive field, irrespective of where in the retina this spatial configuration is seen as well as where the hand is located with respect to the body. Some of these representations are thought to be relevant for facilitating hand-object interactions. This introductory chapter has presented an overview of the relevant literature that provided the foundations for our

theoretical work.

In this thesis we will present a number of self-organising neural network models of how hand-centred visual receptive fields could develop during unsupervised visually guided learning. Chapter 2 will discuss the principles of competition and self-organization in artificial neural networks. These principles are applied throughout this thesis. Additionally, in this chapter we provide a detailed description of a neural network model, VisNet, of the primate visual system.

The first couple of experimental chapters (3 and 4), show the development of hand-centred representations under two different learning mechanisms. In Chapter 3 we present our first self-organising hypothesis of how VisNet may learn to develop hand-centered visual representations in the output layer using a learning mechanism called *Continuous Transformation* (CT) learning (Stringer et al., 2006). This learning mechanism relies on the spatial similarity of the input patterns as a particular spatial configuration of hand and object shifts over the retina. One of the initial findings in this chapter is that CT learning seems to be able to develop output neurons that respond to specific locations of a visual target object with respect to the hand. However, when the number of hand-centred training locations is increased, CT learning decreases its spatial selectivity. This provided an early indication that CT learning might fail to develop localised hand-centred receptive fields when the model is trained with targets covering many hand-centred locations.

On the other hand, in Chapter 4 we present a second self-organising hypothesis using a

learning mechanism called *trace learning* (Földiák, 1991; Rolls, 1992). Here we incorporated a memory trace of recent neuronal activity which encourages neurons to learn to respond to input patterns that tend to occur close together in time. If there are periods of time in which the eyes tend to shift about a visual scene rapidly while the hand is near an object, trace learning will bind together successive retinal images corresponding to the same hand-object configuration, thus forming localised hand-centred output representations. The results presented in Chapter 4 show how the network is capable of generating output cells that are tuned to specific target locations in a hand-centred frame of reference. Thus, the network is able to develop localised hand-centred receptive fields even as the number of hand-object configurations increases during training.

In the following chapters the trace learning hypothesis is advanced by gradually increasing the realism of the visual training conditions. In Chapter 5, in an advance on the work done in the previous chapters, we trained the network using more natural visual scenes consisting of multiple targets presented at the same time around the hand instead of just an isolated object in a particular location with respect to the hand. In this chapter we presented three sets of experiments, and we introduced the principles of *statistical decoupling* that allows the output cells of the network to develop single localised receptive fields even when the network is always trained with more than one object near the hand. First, VisNet was trained on idealised images consisting of pairs of circular objects presented simultaneously near the hand. In a second study, VisNet was trained with com-

puter generated images consisting of a realistic image of a hand and a variety of natural objects, presented within different textured backgrounds during training. The network is then tested with just one textured object near the hand in order to verify if the output cells are capable of building hand-centered representations with a single localised receptive field. In an additional simulation, we examined how some of the cells with hand-centered receptive fields decrease their shape selectivity and start responding to a larger proportion of objects within a localised region of hand-centered space as the number of objects presented in overlapping locations during training increases. Lastly, in the third study we explored the same learning principles using natural visual scenes collected by volunteers. These results provide an important step in showing how single, localised, hand-centered receptive fields could emerge under more ecologically realistic visual training conditions.

The ecological plausibility of our trace learning hypothesis was further explored in Chapter 6 using eye and head movements recorded from human participants as they completed a jigsaw. A head-mounted camera captured images of the hand and jigsaw, while eye movements were recorded by an eye-tracking device. This combination of data allowed us to reconstruct the retinal images seen as humans undertook the jigsaw task. These retinal images were then fed into a new time-accurate differential formulation of the VisNet model in order to faithfully represent the temporal dynamics of the recorded gaze changes. These results provide an important step in demonstrating how localised, hand-centred receptive fields can emerge through visual experience that is shaped by the natural eye and head

movements of human participants.

In Chapter 7 we present a new self-organising model of how hand-centred representations may develop using a proprioceptive input representation of the position of the hand with respect to the body. In these simulations we show how trace learning can first develop a body-centred representation of the locations of visual targets by incorporating gain-modulated signals that provide information about the position of the eyes and the head. Subsequently, we show how these body-centred representations of the location of visual targets can be used in combination with the proprioceptive representation of the location of the hand in order to develop hand-centred output representations.

Lastly, in Chapter 8 this thesis concludes with a discussion of the main findings and some suggestions for future work.

Essentially, all models are wrong, but some are useful.

—George E. P. Box, *Empirical Model-Building and Response Surfaces*

2

Methods: Competitive networks, VisNet and metrics

This chapter provides the details of the self-organising principles that were utilized in this thesis in order to develop cells with hand-centered visual representations. We begin by introducing the principles of competition and self-organization that are applied throughout the thesis. Subsequently we provide a detailed description of the VisNet architecture and its dynamics, as well as the information analysis metrics applied to interpret the results. VisNet was used for the simulation studies presented in Chapters 3, 4, 5 and 6. This chapter provides the basic computational principles of the original version of VisNet. Each of the adaptations to the original architecture and dynamics will be discussed in each of the experimental chapters. In Chapter 7 we present a new architecture composed of competitive networks.

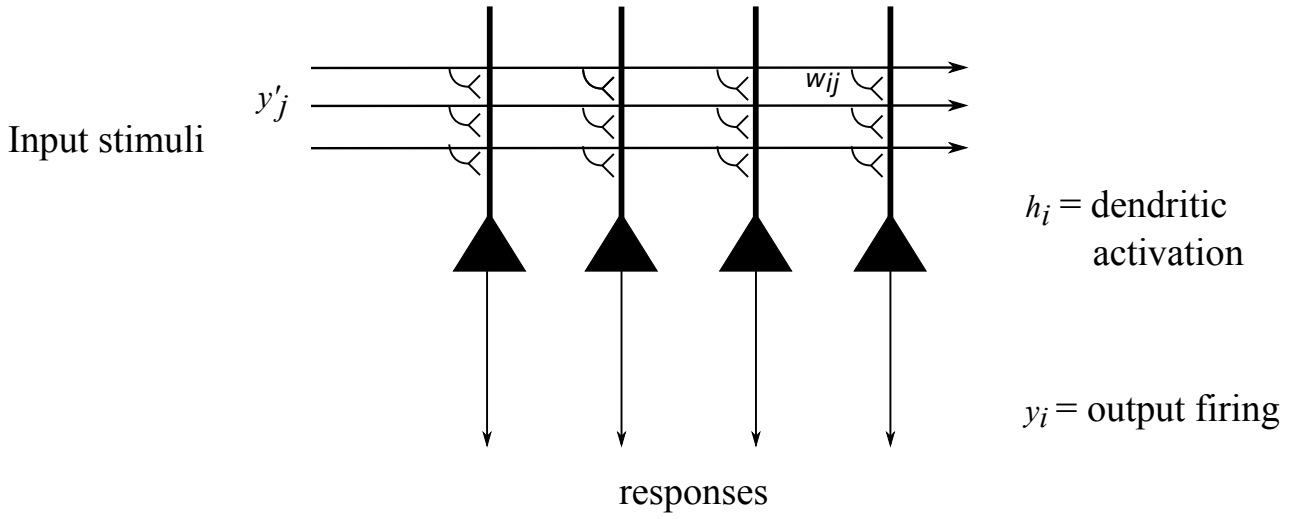


Figure 2.1: Architecture of a competitive network. Adapted from Rolls and Treves (1998).

2.1 Competitive learning

All of the neural networks models used in this thesis incorporate the self-organising principles of competitive learning. Competitive networks have been applied extensively as a powerful and biologically plausible unsupervised learning algorithm that allows cells to group input patterns that share general features (Rolls and Treves, 1998; Rumelhart and Zipser, 1985). What makes competitive networks particularly attractive as models of cortical processing is that they are capable of finding and learning these common features based on the underlying similarities and statistics of the input patterns, without an external 'teacher' signal that forces the network to respond in a pre-specified or desired way. In other words, they learn, self-organize and discover the common features and structure of their environment by directly interacting with it. If the structure and statistics of the

environment change, these models will eventually adapt and self-organize to represent the new statistics of the input space.

Figure 2.1 provides an example of a simple competitive network where a layer of input neurons (i.e. represented by the three horizontal arrows) project their signals to a single layer of output neurons (i.e. represented by the four vertical neurons) through modifiable synapses w_{ij} .

When an input pattern is presented to the network, the activation h_i of each neuron i in the output layer is calculated as the linear sum of the inputs r_j from afferent neurons weighted by the corresponding synaptic weights w_{ij} . This is computed according to

$$h_i = \sum_j w_{ij} y_j \quad (2.1)$$

where y_j is the firing rate of neuron j , and w_{ij} is the strength of the synapse from neuron j to neuron i .

The output neurons are activated and compete with each other, for example by means of mutual inhibition. In the brain, mutual inhibition would be mediated by inhibitory interneurons. Subsequently, the output cell (i.e. winner takes all) or subset of cells (graded competition) with the strongest activation will win the competition and get its afferent synaptic connections strengthened using typically a Hebb learning rule

$$\delta w_{ij} = \alpha y_i y_j \quad (2.2)$$

where δw_{ij} is the change in the synaptic weight w_{ij} , α is the learning rate, y_i is the firing-rate of the post-synaptic neuron i , and y_j is the firing-rate of the pre-synaptic neuron j .

In order to restrict and limit the growth of each neuron's synaptic weight vector, its length is renormalised at the end of each timestep during training. This is an important step in competitive learning (Hertz et al., 1991) and is implemented by rescaling the synaptic weight vectors making sure that for each output cell i we have

$$\sqrt{\sum_j (w_{ij})^2} = 1. \quad (2.3)$$

where the sum is over all input cells j .

After learning, the same output neuron or subset of output neurons will be more likely to be activated when the same input pattern is presented. Novel patterns that share a high degree of overlap with the learned input vector will be more likely to activate the same subset of output cells, while dissimilar input patterns are likely to activate different cells in the network.

Due to their capacity for categorizing similar input vectors, competitive networks have often been used as a pattern classification or categorization algorithm. This categorization allows for a more compressed representation of the input vectors, which is useful for removing redundancies and creating sparse representations of the input space that are more manageable for later stages of processing (Rolls and Treves, 1998). Similarly, given their

capacity for clustering input vectors, competitive networks are also useful for separating out dissimilar input vectors, which can be useful for orthogonalizing the input space and preventing interference between partially overlapping input vectors.

2.2 Visnet

The experiments presented in Chapters 3,4,5 and 6 of this thesis were conducted using VisNet (See Figure 2.2). VisNet is a well-established, self-organising, artificial neural network model that was initially developed as a model of invariant object recognition and a biologically plausible approximation of visual processing in the ventral stream (Wallis and Rolls, 1997). However, it was later recognised that the architecture of VisNet captured the hierarchical organization of competitive neuronal layers that is common in both the dorsal and ventral visual systems, and could thus be applied to explain the development of cell properties in the dorsal stream (Galeazzi et al., *submitted*, Galeazzi et al., 2013, 2015; Rolls and Stringer, 2007; Rolls and Webb, 2014).

The computational similarities between the hierarchy of transform invariant visual representations in the ventral stream and the hierarchy of ego-centric representations that are invariant to changes in gaze and posture in the dorsal stream has also been identified and applied to other models of the development of non-retinal, ego-centric reference frames in the dorsal visual stream (Spratling, 2009).

The VisNet model consists of a hierarchical series of four feedforward layers of compet-

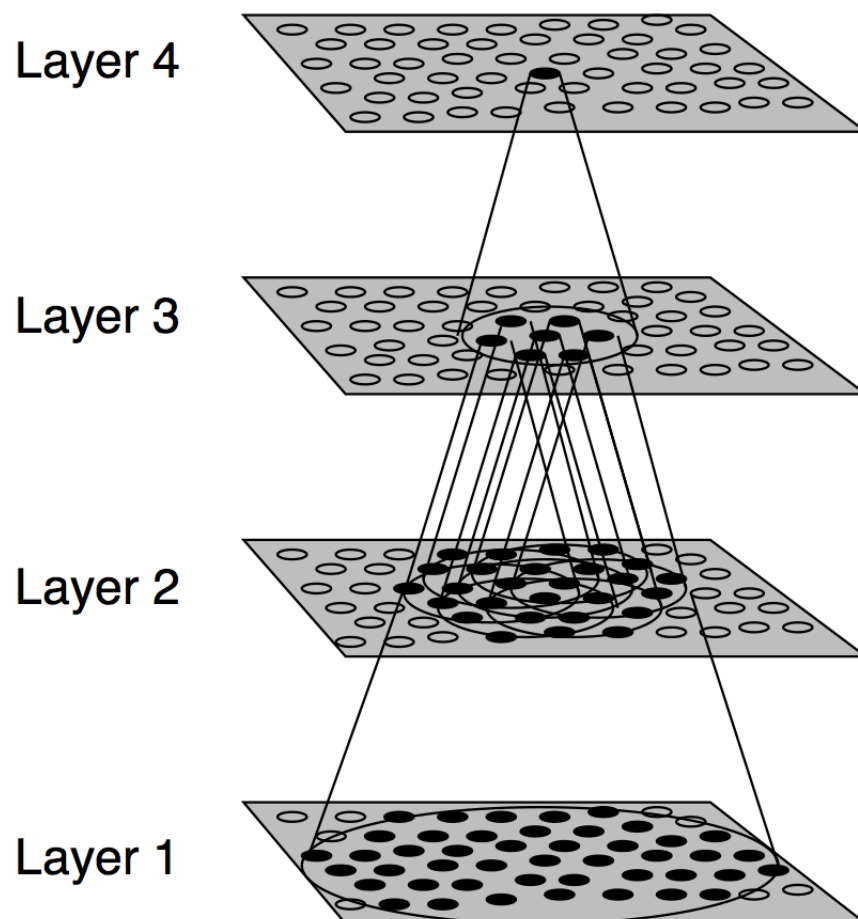


Figure 2.2: VisNet architecture. Stylized image of the VisNet four-layered network. The architecture of the network shows a hierarchical organization. This architecture has been used to simulate visual processing in both the ventral and the dorsal visual systems. Convergence through the network is designed to provide fourth-layer neurons with information from across the entire input retina. From Rolls and Treves (1998).

itive networks. Within each neuronal layer there is lateral competition between neurons implemented by local graded inhibition. During training, there is associative learning at the feedforward synaptic connections between the successive layers of neurons (See Figure 2.2). In VisNet, natural visual images are first passed through an array of Gabor filters that mimic the response properties of V1 simple cells. Subsequently the output of the Gabor filters are fed to the first layer of the network architecture. The forward connections to individual cells are derived from a topologically corresponding region of the preceding layer, using a Gaussian distribution of connection probabilities. These distributions are defined by a radius which will contain approximately 67% of the connections from the preceding layer. This leads to an increase in the receptive field size of neurons through successive layers of the network hierarchy. The typical network dimensions used are shown in Table 4.1, however throughout the thesis we adapted and scaled-up some of the original dimensions. More detailed descriptions can be found in Rolls (2008).

The simulations in this thesis were conducted utilizing an updated version of the original VisNet model (Rolls and Milward, 2000; Wallis and Rolls, 1997). One of the most recent modifications to the network is related to the preprocessing of the input images. Previous research with VisNet used a Difference of Gaussians function (DoG) as the input filters. More recent work, however, utilizes banks of Gabor input filters (Galeazzi et al., 2013; Rolls, 2012b; Rolls and Webb, 2014). The Gabor functions have been suggested to provide a better fit to some of the properties of the receptive fields of simple cells (Jones and

Palmer, 1987). Therefore, before the stimuli are presented to layer 1 of VisNet, they are pre-processed by an initial layer of Gabor filters representing V1 simple cells, where each x, y -location contains a bank of Gabor filter outputs corresponding to a hypercolumn generated by

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi\frac{x'}{\lambda} + \psi\right) \quad (2.4)$$

$$x' = x \cos \theta + y \sin \theta \quad (2.5)$$

$$y' = -x \sin \theta + y \cos \theta \quad (2.6)$$

for all combinations of $\lambda = 2, \gamma = 0.5, \sigma = 0.56\lambda, \theta \in \{0, \pi/4, \pi/2, 3\pi/4\}$ and $\psi \in \{0, \pi, -\pi/2, \pi/2\}$.

The activation h_i of each neuron i in the network is set equal to a linear sum of the inputs y_j from afferent neurons j in the previous layer weighted by the corresponding synaptic weights w_{ij} . That is,

$$h_i = \sum_j w_{ij} y_j \quad (2.7)$$

where y_j is the firing rate of neuron j , and w_{ij} is the strength of the synapse from neuron j to neuron i .

Within each layer competition is graded rather than winner-take-all, and is implemented in two stages. First, to implement lateral inhibition the activations of neurons within a layer are convolved with a spatial filter, I , where δ controls the contrast and σ controls the

width, and a and b index the distance away from the centre of the filter

$$I_{a,b} = \begin{cases} -\delta e^{-\frac{a^2+b^2}{\sigma^2}} & \text{if } a \neq 0 \text{ or } b \neq 0, \\ 1 - \sum_{\substack{a \neq 0 \\ b \neq 0}} I_{a,b} & \text{if } a = 0 \text{ and } b = 0. \end{cases} \quad (2.8)$$

Typical lateral inhibition parameters are given in Tables 3.2 and 4.2

Next, contrast enhancement is applied by means of a sigmoid activation function

$$y = f^{sigmoid}(r) = \frac{1}{1 + e^{-2\beta(r-\alpha)}} \quad (2.9)$$

where r is the activation (or firing rate) after lateral inhibition, y is the firing rate after contrast enhancement, and α and β are the sigmoid threshold and slope respectively. The parameters α and β are constant within each layer, although α is adjusted to control the sparseness of the firing rates. The sparseness a of the firing within a layer can be defined, by extending the binary notion of the proportion of neurons that are firing, as

$$a = \frac{\left(\sum_{i=1}^N y_i / N\right)^2}{\sum_{i=1}^N y_i^2 / N} \quad (2.10)$$

where y_i is the firing rate of the i th neuron in the set of N neurons (Rolls and Treves, 1990, 1998). For the simplified case of neurons with binarised firing rates = 0/1, the sparseness is the proportion $\in [0, 1]$ of neurons that are active. For example, to set the sparseness to, say, 5%, the threshold is set to the value of the 95th percentile point of the activations

within the layer. Typical parameters for the sigmoid activation function are shown in Tables 3.3 and 4.3.

2.2.1 Learning in VisNet: Continuous transformation learning and trace learning

All the learning rules used in this thesis are associative local learning rules, which means that the terms used to modify the synaptic weights are available in the pre and post-synaptic neurons. The models are also unsupervised in that no target output signal is used to guide the learning. Therefore, this provides our models with a higher biological plausibility than models that use supervised error-correction learning rules. For these simulations we will be using two different learning rules: The standard Hebb rule (Hebb, 1949) and the trace learning rule (Földiák, 1991; Rolls, 1992) to adjust the strengths of the feed-forward synaptic connections between successive neuronal layers during training.

Continuous Transformation (CT) learning and *Trace learning* are two different invariance learning mechanisms that we use to develop hand-centred visual representations. They have quite different properties and make different assumptions about the statistics of visual input from the world (Perry et al., 2010; Stringer et al., 2006). We will compare and contrast the operation of the network model under both learning mechanisms.

In Chapter 3 we will present our first self-organising hypothesis of how VisNet may develop hand-centered visual representations in the output layer using CT learning (Stringer

et al., 2006), which relies on the spatial similarity of the input patterns as a particular spatial configuration of hand and object shifts over the retina. Experimental evidence for CT learning in the brain has been provided by the psychophysical study of Perry et al. (2006).

In Chapter 4 we present a second self-organising hypothesis using trace learning (Földiák, 1991; Rolls, 1992). In this learning mechanism we incorporate a memory trace of recent neuronal activity into the trace learning rule, which encourages neurons to learn to respond to input patterns that tend to occur close together in time. Experimental evidence for trace learning has been provided by Cox et al. (2005) and Li and DiCarlo (2008). The neuronal bases for temporal trace learning have been previously discussed by Rolls (1992) and Wallis and Rolls (1997). More recently, Evans and Stringer (2012) have shown how trace learning may arise naturally in a spiking neural network with long synaptic time constants. Altogether, these studies provide strong evidence for how the brain might develop invariant representations by exploiting temporal continuity in the natural statistics of the visual world.

2.2.2 Information Analysis

In addition to showing the response profiles of individual cells before and after training, most of the chapters in this thesis (Chapter 3, 4, 5–6) will be using information analysis measures in order to provide a global measure of network performance in our simulations.

Since the study by Rolls and Millward (2000), information analysis has been used extensively as a measure of performance in VisNet.

Information theory was developed by Shannon in order to quantify the information that was transmitted in the context of signal processing (Shannon, 1948), but it has since then been applied in various disciplines to quantify all forms of communication. For a detailed description of information theory and its applications see MacKay (2003).

In the context of neurophysiology it has been used to quantify the amount of information that is encoded in firing of individual neurons and in the activity across populations of neurons. In VisNet these measures have been applied to quantify the amount of information that is available from the response r of a single cell about a stimulus s that has been shown. In other words, these measures have been used in VisNet to measure how well an output cell has learned to recognize or respond selectively to a particular stimulus even as it is rotated or translated across the retina.

In the context of the hand-centred representations explored in this thesis, we are interested in developing cells that are able to respond selectively to a specific spatial configuration of the hand and a target object invariantly across different retinal locations. We can test for this by quantifying the amount of information that the output cells carry regarding a particular hand-object configuration regardless of retinal location.

During testing we present the network the same hand-object spatial configuration in different retinal locations. The firing rates of the output neurons of the network are com-

puted for each such retinal view. Single and multiple cell information theoretic measures were used to assess the network's performance. Both measures help to determine whether individual cells in the output layer are able to respond selectively to a specific hand-object configuration, i.e. the target location in a hand-centred frame of reference, over a number of different retinal locations.

If an output neuron responds selectively to a spatial configuration, and the cell responds to this configuration across different retinal locations, then the cell will convey maximal single cell information. The amount of information carried by a single cell about a stimulus such as a specific hand-object configuration is computed using the following formula

$$I(s, R) = \sum_{r \in R} P(r|s) \log_2 \frac{P(r|s)}{P(r)} \quad (2.11)$$

where r corresponds to the single response to a single stimulus. The stimulus-specific information $I(s, R)$ is the amount of information the set of responses R of a single cell has about a specific stimulus (i.e. target location with respect to the hand) s . The set of responses R corresponds to the firing rate y of a cell to each of the stimuli presented in all five retinal locations. Further details of how the single cell information is calculated are provided by Rolls and Treves (1998) and Rolls and Milward (2000).

The maximum single cell information possible is

$$\text{Max. single cell info.} = \log_2(\text{Number of stimuli}), \quad (2.12)$$

where the number of stimuli is the number of different spatial configurations of the hand and target object. For example, with three hand-object configurations the maximum single cell information is 1.58 bits. This is achieved when the output cell responds selectively to just one of the three spatial configurations, and responds to that spatial configuration over all retinal positions.

On the other hand, the multiple cell information computes the average amount of information about which stimulus was presented obtained from the responses of all the output cells. This procedure is used to verify whether, across the population of cells, there is information about all of the different hand-object configurations that have been shown. Procedures for calculating the multiple cell information measure have been previously described in more detail (see Rolls and Milward, 2000; Rolls and Treves, 1998; Rolls et al., 1997). In brief, from a single presentation of a stimulus, we calculate the average amount of information obtained from the responses of all the cells regarding which stimulus is shown. This is achieved through a decoding procedure that estimates which stimulus s' gives rise to the particular firing rate response vector on each trial. A probability table of the real stimuli s and the decoded stimuli s' is then constructed. From this probability table, the mutual information is calculated as

$$I(S, S') = \sum_{s, s'} P(s, s') \log_2 \frac{P(s, s')}{P(s)P(s')}. \quad (2.13)$$

Multiple cell information values are calculated for the subset of cells which, according to

the single cell analysis, have the most information about which stimulus (i.e. hand-object configuration) is shown. In particular, the multiple cell information is calculated from five cells for each stimulus that had the most single cell information about that stimulus. For example, in simulations with three target locations this results in a population of 15 cells. Previous research (Stringer and Rolls, 2000) found this to be a sufficiently large subset of cells to demonstrate that shift invariant representations of each stimulus presented during testing were formed, and that each stimulus could be uniquely identified.

2.2.3 Selection and optimization of parameters in VisNet

In this section we provide an overview of the relevant parameters explored in this thesis. In practical terms, it is not possible to search the entire parameter space. However, extensive research with VisNet has shown that the best performance of the network is achieved with a set of initial parameters as found by Rolls and Milward (2000). Therefore, a pragmatic approach was taken guided by previous results, as well as the modellers' knowledge of how the system is likely to behave given certain adjustments.

Typical lateral inhibition parameters are given in Tables 3.2 and 4.2. Typical parameters for the slope of the sigmoid activation function are shown in Tables 3.3 and 4.3. The purpose of implementing lateral inhibition and contrast enhancement is to introduce lateral competition within the competitive layers as well as enabling different subsets of neurons to encode different inputs. In competitive networks, lateral competition can be

implemented through inhibitory interneurons. However, VisNet implements lateral competition by convolving the activation of neurons with a spatial filter described in Equation 2.8. Throughout the thesis these parameters remained mostly unchanged and based on the typical parameters employed in VisNet research (Rolls and Milward, 2000).

The typical network dimensions used in this thesis are shown in Tables 3.1 and 4.1, however throughout the thesis we adapted and scaled-up some of the original dimensions and connectivity. This type of re-scaling has been done in recent experiments with VisNet (Eguchi et al., 2015; Rolls and Webb, 2014). The re-scaling of the layers helps to improve the resolution of the signal and thus the performance of the model, particularly when very similar or highly overlapping images have to be dissociated. This re-scaling was particularly implemented in Chapter 5 and 6, where our simulations introduced more realistic images and a wider range of retinal shifts. The existing network size was not sufficient to implement this, particularly because the small retina generated low resolution outputs and thus a more noisy signal to the subsequent layers of processing. On the other hand, in Chapters 3 and 4, the numbers of connections between the layers were modified to explore the effects of further dilution in the network connectivity. However, any modifications to the initial set of parameters will be specified in each subsection.

In Chapter 6 we implement a novel time-accurate differential version of the VisNet model, in which for the first time, the retinal shifts of the network are guided according to data recorded from the natural eye-movements of human subjects collected with an

eye-tracker during a hand-object manipulation task. This differential implementation also required a scaling-up of the original dimensions, as well as introducing parameters for the activation time constants and the trace time constant. However, the rest of the parameters for lateral inhibition and contrast enhancement were not modified. All of the chapters provide information about the specific parameters used for the simulations.

I shall venture to affirm, as a general proposition, which admits of no exception, that the knowledge of this relation is not, in any instance, attained by reasoning a priori; but arises entirely from experience, when we find that any particular objects are constantly conjoined with each other . . . all these arguments are founded on the supposition of that resemblance

—David Hume, *Enquiries Concerning The Human Understanding and Concerning The Principles of Morals* (1737)

3

The development of hand-centred visual representations by Continuous Transformation learning

This chapter explores the ability of VisNet to develop hand-centred representations in the output layer using a learning mechanism called Continuous Transformation (CT) learning (Stringer et al., 2006). CT learning relies on the spatial similarity of successive visual images as they shift smoothly across the retina. CT learning uses the biologically plausible Hebbian learning rule for the modification of its synaptic weights. The results of our initial simulation showed that CT learning can encourage output neurons to learn to respond to specific locations of a visual target object with respect to the hand. However, when the density of trained target locations was increased, CT learning decreases its spatial selectivity and starts to respond to contiguous locations. The performance of VisNet is

evaluated using self-organising maps and more diluted connectivity.

3.1 Introduction

In the previous chapters we briefly introduced the two main learning mechanisms that we would be exploring in this thesis to provide a self-organising account of how hand-centred representations may develop during visually guided learning. As mentioned earlier, CT learning and trace learning have distinct properties and make different assumptions about the way we can exploit the statistics of the visual input (Perry et al., 2010; Stringer et al., 2006).

In order to examine these two learning mechanisms we used a well-established neural network model of visual processing in the primate brain, called VisNet (Rolls and Milward, 2000; Wallis and Rolls, 1997). One of the main virtues of VisNet in contrast with some of the models described in the introduction of this thesis, is that in VisNet learning is completely unsupervised (i.e. there is no external teaching signal to tell the network what the output should be). Instead of supervised algorithms (e.g. backpropagation of error), we used more biologically plausible learning rules where the synaptic weights are updated locally by associative Hebbian learning (Hebb, 1949) or Hebb-like rules with a memory trace in the case of the *trace learning* rule (Földiák, 1991; Rolls, 1992). In this Chapter we begin by exploring the CT learning mechanism (Perry et al., 2010; Stringer et al., 2006).

As discussed previously, VisNet has been used to simulate visual processes occurring

in ventral as well as in the dorsal stream (Rolls and Stringer, 2007; Rolls and Webb, 2014). Both visual streams seem to share architectural similarities, each consisting of a hierarchical series of neuronal layers with competition mediated by inhibitory interneurons. Similarly, it has been suggested that both pathways might be undertaking similar computations (Spratling, 2009). For example, in the ventral stream we find a hierarchy of object representations, starting with selective responses to the orientation of bars in the simple cells of V1 (Hubel and Wiesel, 1962, 1968), subjective contours in V2 (Ffytche and Zeki, 1996), combinations of angles and curvatures in V4 (Pasupathy and Connor, 2001), simple shapes in TEO (Kobatake and Tanaka, 1994) and at latter stages of the processing hierarchy, along the inferior temporal cortex, we find cells with increasing complexity in their response properties. For example, cells that seem to respond to objects or faces, invariant to size and retinal location (Ito et al., 1995). Similarly, within the dorsal pathway we find a hierarchical organization for processing motion (Perry and Fallah, 2014) as well as a hierarchy of coordinate systems, where we begin with retinocentric or eye-centred representations and in the latter stages of processing we find body-centred and body-part centred representations (Flanders et al., 1992; Lacquaniti et al., 1995; McIntyre et al., 1998; Soechting and Flanders, 1992). Some of these body-centred or body-part centred representations in latter stages of visuomotor processing could be also interpreted as being invariant to changes in retinal locations and across different postures (Spratling, 2009).

VisNet incorporates this hierarchical organization of the visual system and implements

it using series of competitive layers. Moreover, different studies using VisNet have shown that unsupervised learning mechanisms can be successfully applied to develop cells in the output layer of the network that show some of the high-level invariant representations reported in the ventral stream (Rolls, 2012b; Rolls and Milward, 2000) as well as some of the high-level representations found in the dorsal stream (Rolls and Stringer, 2007; Rolls and Webb, 2014). CT learning has been one of the main self-organising learning mechanisms that has been successfully used in VisNet to develop high-level, invariant representations in the cells of the output layer.

In this chapter we present our first self-organising hypothesis of how hand-centred representations may develop using CT learning. CT learning exploits the spatial similarity across different transforms (e.g. changes in retinal location) of the same visual stimulus (Perry et al., 2010; Stringer et al., 2006). We hypothesize that adjacent retinal locations of an individual hand-centred configuration (i.e. a hand and an object near the hand) comprise a particular subset of low-level visual features that will share a higher degree of similarity and spatial overlap than the retinal views of two completely different hand-centred configurations. CT learning may then exploit the spatial similarity between different views of a particular hand-centred configuration to develop a representation of this configuration across different retinal locations. We assumed that the different views of the same hand-centred spatial configuration could be generated by smooth shifts of this configuration across the retina. This process is repeated through the hierarchically arranged

architecture of VisNet which allows the cells in the higher layers to increase their level of invariance across the entire visual field. In other words, the cells in the output layer would learn to respond to a specific hand-object configuration irrespective of where in the retina this configuration appears. The plausibility of our CT learning hypothesis is explored in the computer simulations described below.

3.1.1 CT learning hypothesis

CT learning relies on the spatial similarity of successive visual images, as a visual object or set of bound objects shift smoothly across the retina. An idealised example of the process by which CT learning guides the development of output neurons with translation invariant responses is shown in Fig. 3.1. In this Figure we show an input layer with feedforward projections into an output layer. The synaptic weights are initially set to random values just like in VisNet. When an input pattern is presented to the network (in this case the view of a hand-centred configuration), this generates an activation of a subset of input cells (i.e. the three shaded cells in retinal location 1). This activation is transmitted through the synaptic connections to stimulate the cells in the output layer. An output cell (or a subset of output cells) in the output layer will win the competition and consequently will have its weight strengthened by an associative learning rule (represented by the solid lines between the layers).

A gradual change in the position of the visual stimulus (i.e. a hand-object configuration)

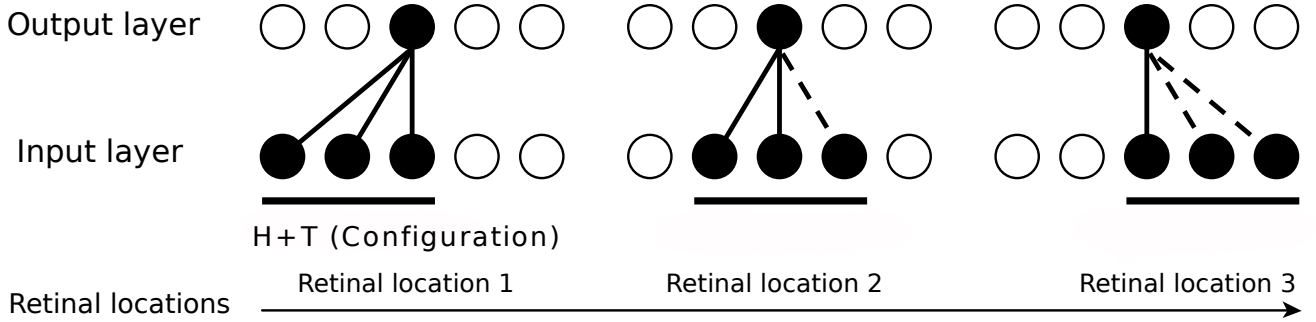


Figure 3.1: CT learning hypothesis. The figure shows the basic principle of how CT learning operates in a one layer network. In this case H+T configuration refers to the image of a target object presented in a fixed position with respect to the hand. In this example, the same hand-centred configuration is viewed in different retinal locations. Figure adapted from (Stringer et al., 2006)

across the retina, will generate the activation of spatially overlapping input neurons. For example, at the retinal location 2, some of the active input neurons will be the same as the ones that were activated in the adjacent retinal location 1. However, in retinal location 2 there will be activation of new input cells that were previously inactive. The activation of the overlapping cells that had their synaptic weight strengthened will drive the activation of the same output cell and make it more likely to win the competition again. This will in turn strengthen the synaptic weights between the new input cell that was previously inactive in the retinal position 1 and the same output cell (represented by the dashed line). Hence, the same output neuron will learn to respond to two different views or retinal locations of the same hand-object configuration, given that they share a high degree of spatial overlap. As it can be seen in retinal location 3, this process can be continued for subsequent shifts until the same output cell (or subset of cells) learn to represent the same

hand-centred configuration across a continuum of retinal locations.

In CT learning we use the standard Hebb rule (Hebb, 1949) to update the strengths of the feed-forward synaptic connections between the layers. This mechanism allows the network to bind together different views that belong to the same stimulus. The Hebb rule is given by

$$\delta w_{ij} = \alpha y_i y_j \quad (3.1)$$

where δw_{ij} is the change in the synaptic weight w_{ij} , α is the learning rate, y_i is the firing-rate of the post-synaptic neuron i , and y_j is the firing-rate of the pre-synaptic neuron j . In order to restrict and limit the growth of each neuron’s synaptic weight vector, its length is renormalised at the end of each timestep during training. This is important in competitive learning (Hertz et al., 1991) and is implemented by rescaling the synaptic weight vectors making sure that for each output cell i we have

$$\sqrt{\sum_j (w_{ij})^2} = 1. \quad (3.2)$$

where the sum is over all input cells j .

We hypothesise that as we explore a visual scene extensively, we often find different views of an object in a fixed position with respect to the hand. Adjacent retinal locations produced by gradual shifts of a particular hand-centred configuration across the retina will share a subset of visual features with a high degree of spatial overlap. A form of CT

learning may be able to associate the different retinal views that correspond to a particular spatial configuration of the hand and object, with the same subset of output neurons. After enough training, this would enable individual output cells to learn to respond to a particular spatial configuration of the hand and a target object across different retinal locations. Furthermore, gradual changes of the same hand-centred configuration will share a higher degree of overlap than the different views belonging to different hand-centred configurations. This would result in a population of output cells that respond to particular hand-centred configurations regardless of retinal locations. Such neurons will effectively display localized visual receptive fields that respond selectively to the position of a target in a hand-centred frame of reference.

It is important to emphasize that CT learning relies on spatial similarity of the images, therefore it is not assumed that all of the different views of the same hand-centred configuration have to necessarily occur in temporal proximity. In other words, this phenomenon could arise after extensive experience and visual exploration of scenes, in which we would have seen a hand-object configuration across a continuum of fixation points. The spatial overlap between different views of a particular hand-centred configuration is what allows CT learning to bind these different views into the same output cells.

3.2 Method

The simulations consisted of presenting different hand-object configurations during training. The spatial configurations presented during training are composed of a hand and a circular object near the hand. Each of the spatial configurations was shifted smoothly across the retina during training. The synaptic weights are strengthened using the Hebb rule, followed by weight vector normalisation after the presentation of each of the images of the sequence.

After training, the network is tested with images consisting of the hand and a single object presented in a fixed spatial configuration across different retinal locations. The network is tested on all of the trained hand-object spatial configurations across the different retinal views shown during training.

We begin by presenting a simple demonstration of how CT learning could provide a mechanism for developing representations of the location of visual targets with respect to the hand. We will then present the results as we increase the number of target locations around the hand. Lastly, we show the changes in the network's performance using self-organising maps and more diluted connectivity.

3.2.1 Model architecture and parameters

The experiments presented in this chapter were conducted using the VisNet architecture that has been described in Chapter 2, any modifications to the parameters will be specified

in each subsection. The network dimension used in these experiments are provided in table 3.1. The lateral inhibition parameters are reported in table 3.2. The sigmoid parameters used to control the global inhibition within each layer of the model are presented in table 3.3. The spatial sigmoid threshold is set to the 95th percentile point, which means that approximately 5% of the output neurons will remain active and hence learn to respond to the current input pattern (i.e. hand-object configuration).

	Dimensions	Number of Connections	Radius
Layer 4	32x32	100	12
Layer 3	32x32	100	9
Layer 2	32x32	100	6
Layer 1	32x32	100	6
Retina	128x128x16	-	-

Table 3.1: Network dimensions Network dimensions showing the size number of connections per neuron and the radius in the preceding layer from which 67% of the connections are received

Layer	1	2	3	4
Radius, σ	1.38	2.7	4.0	6.0
Contrast, δ	1.5	1.5	1.6	1.4

Table 3.2: Lateral inhibition parameters

Layer	1	2	3	4
Percentile	99	98	88	95
Slope β	190	40	75	26

Table 3.3: Sigmoid parameters The sigmoid parameters used to control the global inhibition within each layer of the model.

3.2.2 Stimuli

The training stimuli for this experiment consisted of computer-generated images showing a hand and one circular object in a particular spatial configuration near the hand on a grey background. The configurations were designed, arranged and aligned using MATLAB. Figure 3.2 shows an example of some of the stimuli that were utilized in our baseline simulation.

3.2.3 Training

For the baseline simulation, the three image sequences presented to VisNet during training are shown in the separate rows of Fig. 3.2. Each row displays a single sequence consisting of a set of computer-generated images showing a hand and a circular visual object in a particular spatial configuration. The visual object is fixed in one of three possible positions with respect to the hand (Up, Left and Right). The images belonging to a particular sequence arise from smooth shifts in the position of the hand and visual target on the 128×128 ‘retina’. During each of the three image sequences, the fixed spatial configuration

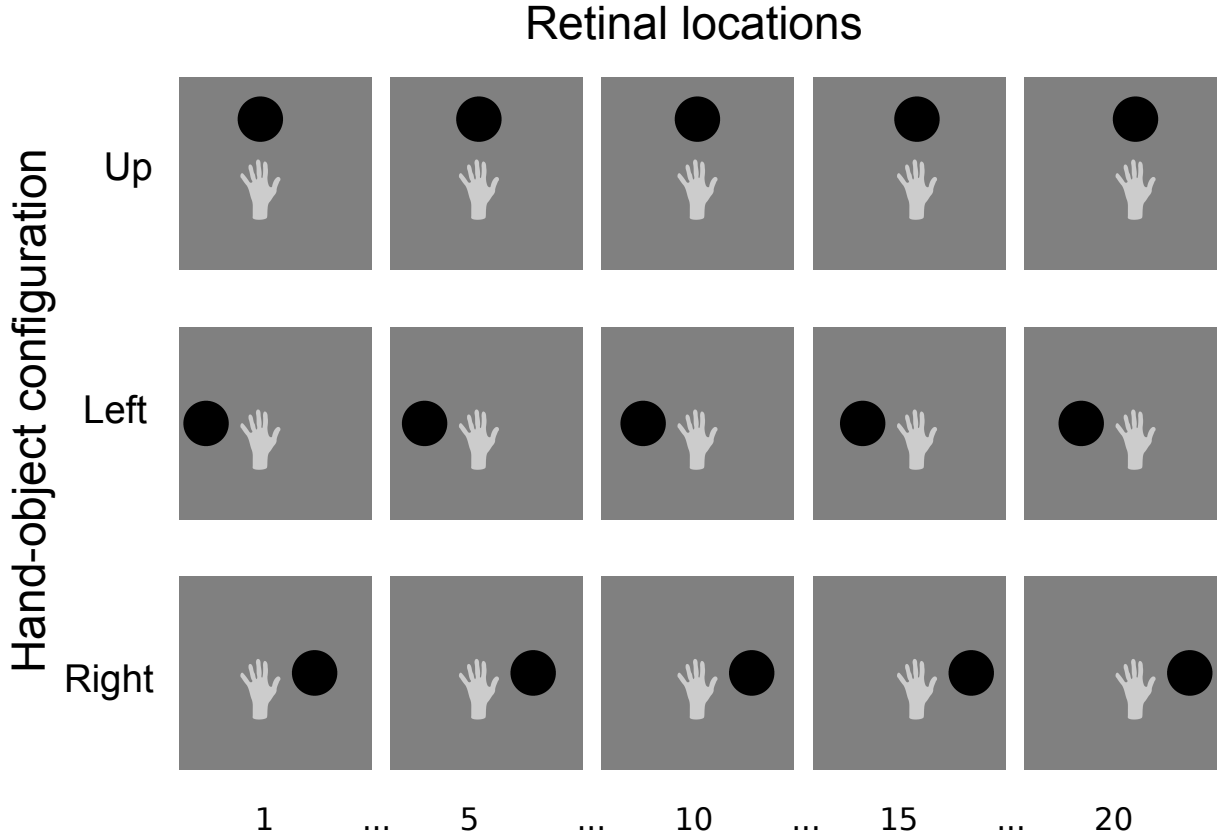


Figure 3.2: Image sequences. The three image sequences presented to VisNet during training are shown in the separate rows. All of the images in each sequence consist of a hand and a circular visual object. For each image sequence, the visual object is fixed in one of three possible positions with respect to the hand. Top row: visual object is shown in the ‘Up’ location with respect to the hand. Middle row: visual object is shown in the ‘Left’ location with respect to the hand. Bottom row: visual object is shown in the ‘Right’ location with respect to the hand. Each of the three image sequences could arise from a series of eye movements that result in small shifts in the position of the hand and visual object on the retina (e.g. drifts or microsaccades). Each of the three image sequences is composed by fixed spatial configuration of the hand and object translated 20 times by 1 pixel. Here we ordered the images according to the gradual shifts produced towards the right

of the hand and object is translated one pixel at a time towards the right.

During the presentation of every image the activation of individual neurons and their firing rates are calculated and subsequently the synaptic weights are updated. The presentation of all three image sequences (i.e. Up, Left and Right) across all 20 retinal locations constitutes 1 epoch of training. The network is trained one layer at a time starting with layer 1 and finishing with layer 4. In the simulations described here, the numbers of training epochs for layers 1-4 were 50, 50, 50 and 50, respectively.

CT learning relies on spatial similarity of the images, this mechanism allows for interleaving different images in the sequences. Therefore another way in which these sequences could arise is that after extensive experience and visual exploration of scenes, we would have encountered fixed hand-object configurations across a continuum of fixation points.

In order to investigate the effects of gradually increasing the density of hand-centred object locations, throughout this chapter we will be presenting simulations in which we train the network with different numbers of hand-centred object locations. In all of the testing cases, the centres of the circular objects will be evenly distributed along a semicircle with respect to the centre of the hand (See Figure 3.6 and Section 3.3.2).

3.2.4 Learning

For these simulations we used a standard Hebb rule (Hebb, 1949) with weight vector normalisation to adjust the strengths of the feed-forward synaptic connections between the

layers during training (see Equations 3.1 and 3.2). As outlined in our learning hypothesis 3.1.1, the gradual shifts in the position of a hand-centred configuration across the retina will generate the activation of spatially overlapping VisNet input filters. This in turn leads to the activation of the same subset of output cells. By updating the synaptic weights using a standard associative rule, this subset of postsynaptic cells will learn to respond to a set of input patterns that correspond to different views of the same hand-centred configuration.

3.2.5 Testing

During testing, the synaptic weights are unchanged. In order to test whether VisNet has been able to develop output neurons that respond selectively to particular hand-object configurations with translation invariance, image sequences of the hand and a circular object are presented in all of the trained retinal locations. The outputs of the network are computed during the presentation of each spatial configuration in each retinal location.

3.3 Results

3.3.1 Baseline simulation with three hand-centred target locations

The first step was to provide a simple demonstration that CT learning could enable a self-organising mechanism for developing cells in the output layer of VisNet that are tuned

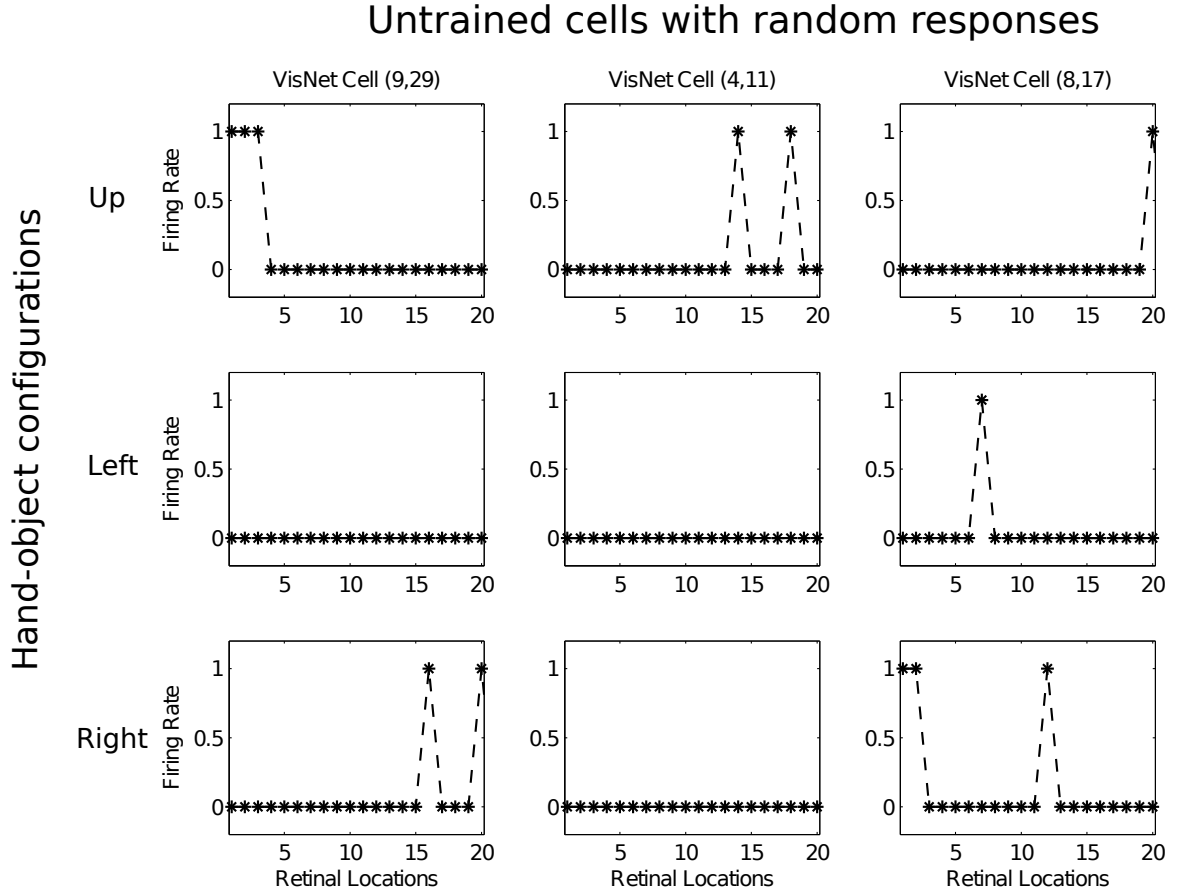


Figure 3.3: Response profiles of three neurons in the top (fourth) layer of VisNet before training. Each of the three columns shows the firing responses of a particular cell. Each row shows the responses of the cells to one of the three hand-object configurations over all 20 retinal locations shown along the abscissae. Top row: visual object is shown in the ‘Up’ location with respect to the hand. Middle row: visual object is shown in the ‘Left’ location with respect to the hand. Bottom row: visual object is shown in the ‘Right’ location with respect to the hand. It can be seen that each of the three cells initially responds randomly to each of the hand-object configurations over the different retinal locations.

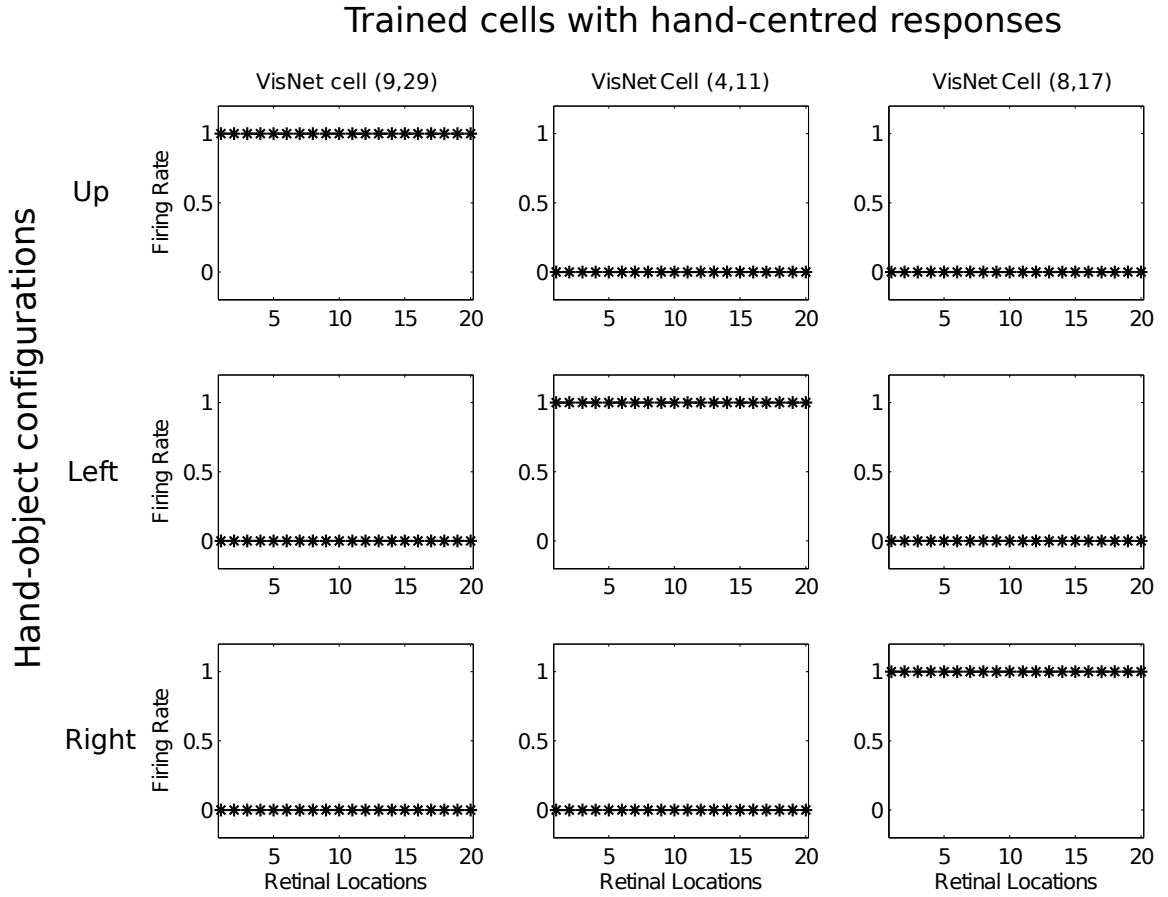


Figure 3.4: Response profiles of same three neurons in the top (fourth) layer of VisNet after training. Response profiles of the same three neurons from Fig. 3.3 after training on the images shown in Fig. 3.2. Conventions as in Fig. 3.3. It can be seen that each of the three cells responds selectively to just one of the hand-object configurations, and responds to that configuration over all the trained retinal positions shown along the abscissae. Moreover, each of the three hand-object configurations is represented by one of the cells.

to particular positions of a target object with respect to the hand, irrespective of retinal location. Therefore, in our first set of simulations we started with only three visual target locations with respect to the hand. In order for CT learning to operate effectively, our simulations incorporated successive small (i.e. one pixel) shifts of the retina across the visual scene for each hand-object configuration.

The response profiles of three neurons in the output layer of VisNet before training are shown in Fig. 3.3. Each of the three columns shows the firing responses of one particular cell. The three rows show the responses of the cells to the three hand-object configurations across all the trained retinal locations. The top row shows the cell responses when the visual object is in the ‘Up’ location with respect to the hand. The middle row shows the cell responses when the visual object is in the ‘Left’ location with respect to the hand. The bottom row shows the cell responses when the visual object is in the ‘Right’ location with respect to the hand. Fig. 3.3 shows that, before training, all three cells respond randomly or not at all to the different hand-object configurations.

Fig. 3.4 shows the response profiles of the same three neurons in the output layer of VisNet after training. It can be seen that, after training, each of the three cells has learned to respond to just one of the hand-object configurations, and responds to that configuration over all the trained retinal locations. The cell in the left column has learned to respond when the visual object is in the ‘Up’ location with respect to the hand. The cell in the middle column has learned to respond when the visual object is in the ‘Left’ location with

respect to the hand. The cell in the right column has learned to respond when the visual object is in the ‘Right’ location with respect to the hand. Furthermore, each of the three hand-object configurations is represented by one of the cells.

To provide a more global measure of network performance, we analysed the information carried by the output (fourth) layer neurons in VisNet about which of the three hand-object configurations is presented to the retina. Intuitively, if an output cell has learned to respond perfectly to just one hand-object configuration over all twenty retinal locations, then it will convey maximal information about which hand-object configuration is currently presented. Therefore, we applied the single and multiple cell information measures described in Section 2.2.2 to the entire population of 1024 neurons in the output layer before and after training.

The number of spatial configurations of the hand and target object is 3. In this case, the maximum single cell information measure is given by

$$\text{Max. single cell info.} = \log_2(3) = 1.58 \quad (3.3)$$

This gives a maximum single cell information measure of 1.58 bits. This level of information is achieved when the cell responds selectively to just one of the three spatial configurations, and responds to that spatial configuration over all five retinal positions.

Fig. 3.5 shows the information measures for the output (fourth) layer neurons before and after training. On the upper plot, the single cell information conveyed by individual output cells is shown in rank order. Before training, no cells conveyed the maximal single

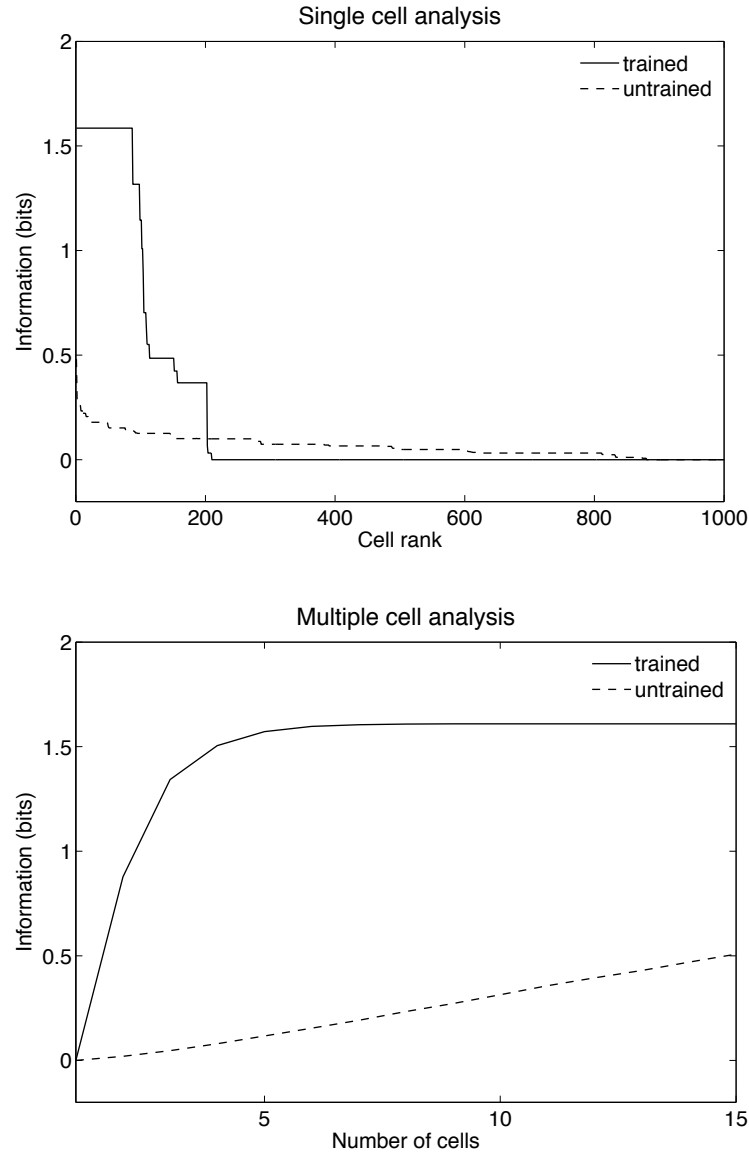


Figure 3.5: Information analysis. Analysis of the information about where the target object is with respect to the hand conveyed by the output (fourth) layer neurons before and after training. The upper plot shows the amount of single cell information carried by individual output cells in rank order. It can be seen that training the network has produced a large increase in the information carried by the output cells. In particular, after training, it was found that 87 cells reached the maximum amount of single cell information of 1.58 bits. In the untrained condition no cells reached maximal information. These cells responded perfectly to just one of the three hand-object configurations, and responded to that configuration across all 20 retinal locations. The lower plot shows the multiple cell information measures calculated across 15 cells with maximal single cell information. It can be seen that, after training, the multiple cell information asymptotes to the maximal value of 1.58 bits. This confirms that all three hand-object configurations are represented by the output cells.

cell information of 1.58 bits. However, after training, 87 cells had reached this level of single cell information. These cells responded to just one of the three hand-object configurations, and responded to their preferred configuration over all twenty retinal positions.

The lower plot shows the multiple cell information measures, which were calculated using 15 cells with maximal single cell information. Multiple cell information values are calculated for the subset of cells which, according to the single cell analysis, have the most information about which stimulus (i.e. hand-object configuration) is shown. After training, the multiple cell information is substantially increased and asymptotes to the maximal information value of 1.58 bits. Therefore, the multiple cell information results show that all three spatial configurations of the hand and object are represented by the output cells. The network simulations successfully developed output cells with neuronal responses which are tuned to a particular location of a visual object in a preferred hand-centred target location. Furthermore, all three hand-object configurations were represented by different subsets of output cells.

In this study however, the target objects in the three hand-centred locations did not have any overlap between them and were almost orthogonal to each other. In the next subsection we will explore the performance of our network when we increase the number of hand-centred locations shown during training.

3.3.2 Capacity study: increasing the number of hand-centred target locations

A key issue is how the network will perform when the number of hand-centred training locations is increased. Can CT learning enable the network to learn to represent increasing numbers of target locations until the whole space around the hand is represented by a continuum of cells? In this case, we would expect to find cells in the output layer with a localized (Graziano and Gross, 1998; Graziano et al., 2004), approximately Gaussian, receptive field centred on a preferred hand-centred location (Bremner and Andersen, 2012; Chang and Snyder, 2010). In this section, we explored the performance of the model for six test cases in which the number of training locations was gradually increased from three to eight in a semicircle around the central hand, as shown in Figure 3.6. The centres of the circular objects were evenly distributed along a semicircle with respect to the centre of the hand (diameter=36 pixels).

Model Parameters

In the simulations described here, the numbers of training epochs for all layers was 50, 50, 50 and 50. The parameters were the same as the ones described in Section 3.2.1.

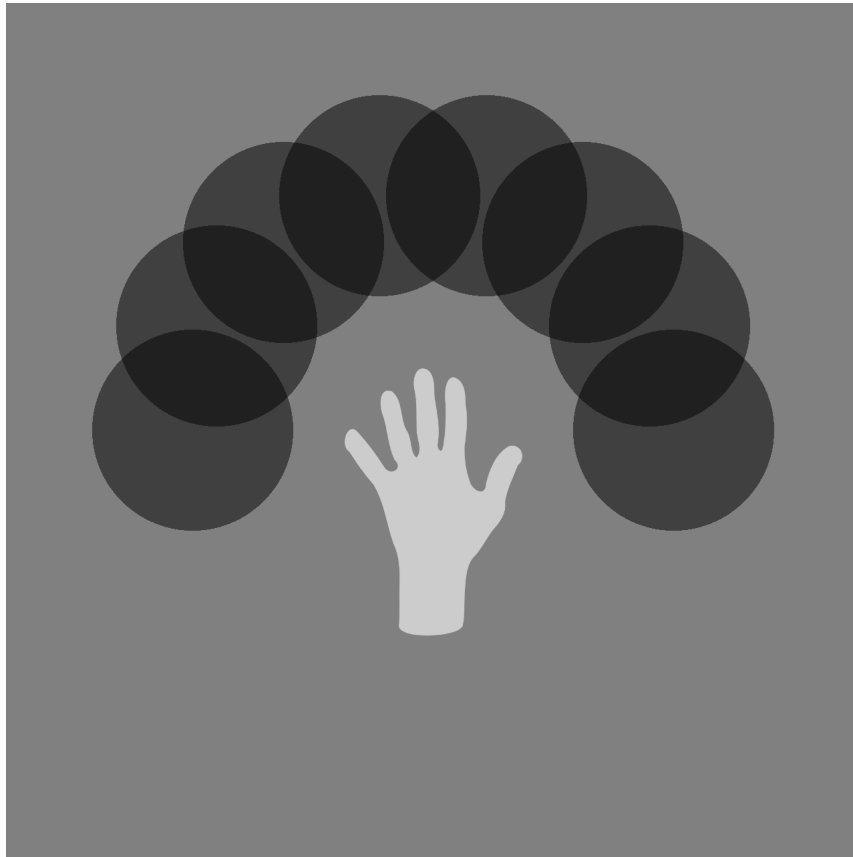


Figure 3.6: Image of the hand with 8 target locations. With 8 target locations around the hand, the targets begin to show some spatial overlap during training.

Results

Fig. 3.7 shows single and multiple cell information analyses for all the test cases. The single cell information analysis (upper plot) shows that in all cases we are able to find neurons conveying the maximal single cell information. However, the multiple cell information plot (lower plot) show that as the model is trained with more training locations, the network is unable to represent all the configurations selectively.

When the model is trained on a large enough number approximating a continuum of

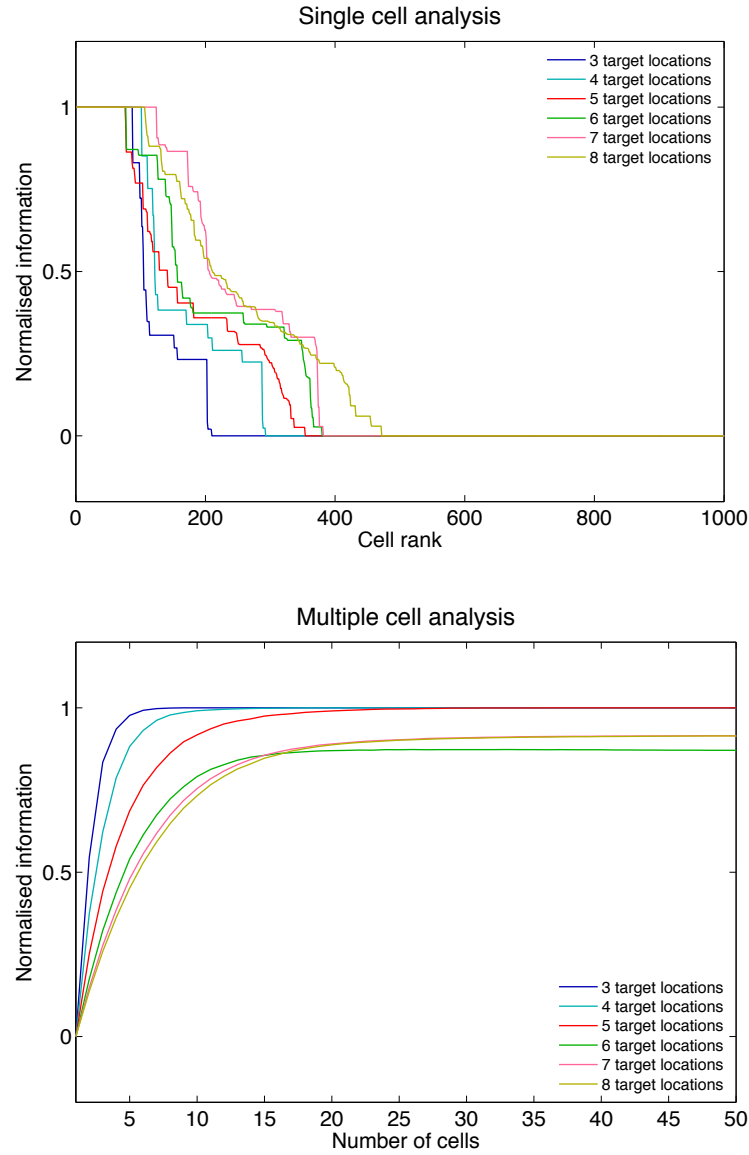


Figure 3.7: Information analysis for increasing number of target locations. Analysis of the information conveyed by the output (fourth) layer neurons about the location of the visual target with respect to the hand. Values on the y-axis were normalised by the maximal information possible, $\log_2(\text{number of target locations})$, for each simulation. The single cell information analysis (upper plot) shows that in all cases there was a subset of neurons conveying the (normalised) maximal single cell information. However, the multiple cell information (lower plot) shows that as we start training the network on a larger number of target locations the multiple cell information does not asymptote to the (normalised) maximal value. This means that not all of the hand-centred configurations are represented selectively by the network. In order to make all the cases comparable we sampled the information from 50 cells.

hand-centred target locations, the continuity of input images between any two hand-centred target locations may allow CT learning to associate images corresponding to different hand-centred target locations onto the same output cell. Such output cells will no longer appear to respond to one particular hand-centred target location. We anticipated that this problem could be exacerbated for the simulations in this chapter, which rely on CT learning to associate successive images of each specific hand-object configuration across different retinal positions in order to form the hand-centred cell responses.

As the number of target locations was increased to 6, it can be observed from the lower plot in Fig. 3.7 that the network started to fail to reach maximal multiple cell information, which meant that some of the target locations were not represented selectively. Some cells learned instead to respond to adjacent hand-centred target locations. Thus, this is an early indication that CT learning may prove to be too associative when the model is trained on a continuum of target locations. We anticipate that CT learning would eventually represent a larger and continuous region of space around the hand as the number of target locations keeps increasing. This raises a problem for proposing CT learning as a mechanism for developing localised hand-centred receptive fields. However, in the discussion we will suggest a potential role for CT learning in the development of large receptive fields.

3.3.3 Performance of the model using a SOM architecture

Tromans et al. (2012) have already shown that the incorporation of a self-organising map

(SOM) architecture within each of the layers of VisNet can dramatically affect the qualitative nature of the output cell firing responses that develop through learning, particularly if the cell response properties are degraded by unwanted CT learning effects. Most pertinently, SOMs are designed to form continuous neural maps of continuous input spaces, such as a continuum of hand-centred target positions. In this case, the SOM architecture may encourage the output cells to learn to represent the continuum of hand-centred target locations across a spatial map in the output layer. Each output cell would learn to respond to just one localised region of the hand-centred target location space with an approximately Gaussian drop off in its response away from that cell's preferred location. The layer of output cells would form a continuous map of the input space of hand-centred target locations, with nearby output cells responding to nearby hand-centred target locations. In this way, the SOM architecture may help to ameliorate the unwanted effects of CT learning, which would cause individual output cells to learn to respond to unrealistically large regions of space around the hand, while the remaining output cells do not fire, when the density of trained hand-centred locations is further increased.

As we have previously discussed, a competitive network (Rumelhart and Zipser, 1985) is able to classify different input patterns according to their similarity in a self-organising manner. A simple modification to the competitive network architecture by introducing lateral excitatory interaction enables the network to develop topological maps where cells with similar firing properties are arranged in spatial proximity within the output layer. A SOM

(Kohonen, 1982; von der Malsburg, 1973) is a neural network model that introduces both short range excitatory connections and long-range inhibitory connections within a layer. This lateral connectivity uses a "Mexican-hat" profile, which has the effect of clustering neurons with similar firing properties in the output layer after training. This occurs due to the mutual excitation properties of neurons that are adjacent, while neurons that are further away inhibit each other, consequently generating separate output pools according to the type of neuronal functionality.

For these simulations, we introduced short range excitatory connections between neurons within the higher neuronal layers, thus turning each competitive layer into a SOM. In Section 3.3.2 we explored the performance of the model for six test cases in which the number of training locations was gradually increased from three to eight in a semicircle around the central hand. In this section we explored the performance of CT learning with the incorporation of a SOM architecture within each layer of VisNet. A key issue here was to explore whether the localised selectivity of the network can be improved when the number of training locations is increased further. The assumption is that the model would be able to represent a larger number of target locations with respect to the hand and to develop smooth hand-centred receptive fields in localised regions.

Model parameters

In this section the VisNet model implemented a series of four layers of SOMs. Table 3.4 shows the lateral inhibition and excitation parameters used for the SOM. The rest of the

parameters were the same as the ones described in Section 3.2.1.

Layer	1	2	3	4
Excitatory Radius, σ_E	0.7	0.55	0.4	0.6
Excitatory Contrast, δ_E	5.35	33.15	117.57	120.12
Inhibitory Radius, σ_I	1.38	2.7	4.0	6.0
Inhibitory Contrast, δ_I	1.5	1.5	1.6	1.4

Table 3.4: Lateral inhibition and excitation parameters for the SOM

Results

The performance of VisNet is evaluated using a SOM. As in the previous section the number of training locations is increased up until 8 highly overlapping hand-object configurations. The upper plot in Figure 3.8 shows the single cell information analysis comparing the performance of the model for different number of target locations. The lower plot in Figure 3.8 shows the multiple cell information analysis.

As the number of target locations was increased to 7, it can be observed that the network again fails to reach maximal multiple cell information, which meant that some of the target locations were not represented selectively. The performance of the network incorporating the lateral interaction of a SOM architecture is marginally better than the performance of the standard competitive network. However, the SOM architecture does not seem to be sufficient to overcome the unwanted effects of CT learning when the density of trained hand-centred locations is further increased.

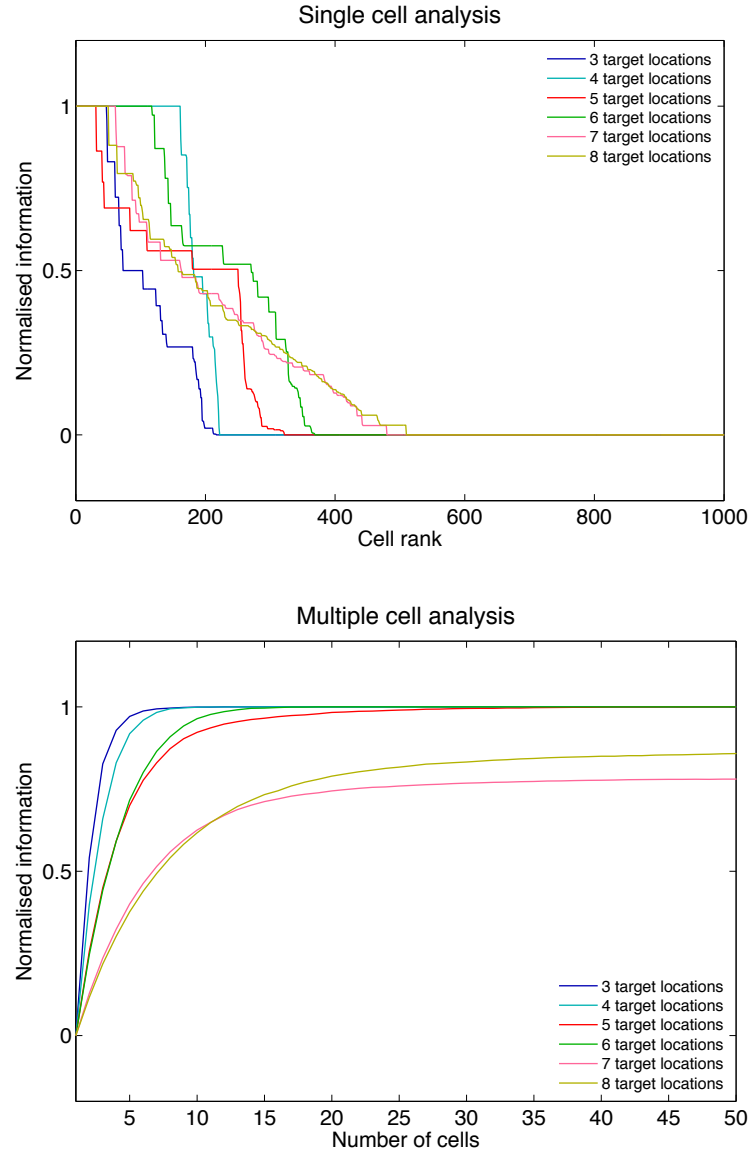


Figure 3.8: Information analysis for all densities using SOMs. Analysis of the information about where the target object is with respect to the hand conveyed by the output (fourth) layer neurons before and after training. The upper plot shows the amount of single cell information carried by individual output cells in rank order. The single cell information analysis shows that in all cases there were neurons conveying the (normalised) maximal single cell information. However, the multiple cell information (lower plot) shows that as we start training the network on a larger number of target locations the multiple cell information does not asymptote to the (normalised) maximal value. This means that not all of the hand-centred configurations are represented selectively by the network. In order to make all the cases comparable we sampled the information from 50 cells.

3.3.4 Effect of diluting connectivity in a SOM architecture and Competitive networks

In the previous subsections (Section 3.3.2 and 3.3.3, we have shown how the associativity of CT learning may encourage cells to learn to represent larger and continuous regions of hand-centred space as the number of target locations keep increasing. We have shown that introducing SOMs within each of the layers of VisNet has not been sufficient to overcome this limitation. How could we then stop CT learning from binding together different hand-object configurations as the number of training locations is increased?

It has been shown that diluted connectivity could be used to eliminate unwanted CT learning effects in head direction cells (Stringer and Rolls, 2006). Therefore, in this section we explored if reducing further the interlayer connectivity of our network could overcome the unwanted effects of CT learning. Diluted connectivity reduces the spatial overlap between the different transforms of a fixed hand-object configuration, this in turn might eliminate the unwanted CT effects. However, by reducing the spatial overlap we might also affect the CT learning ability of binding together the transforms of the same hand-centred configuration across different retinal locations.

The limited connectivity in VisNet, as opposed to a fully connected feed-forward network, is partly inspired by the sparse or diluted cortical connectivity in the brain. Having an architecture with a diluted connectivity in the feed-forward synaptic connections may also help to stabilize some of the feature analyzer properties of the neurons, by enabling

each neuron to learn to respond to a subset of the possible input feature combinations (Rolls, 2008). The advantages of using diluted connectivity in different types of architectures has been explored extensively (Rolls, 2008, 2012a; Rolls and Webb, 2012; Treves, 1991).

The VisNet architecture that we have been using already applies this principle of diluted connectivity. We therefore explored if a greater dilution of the connectivity could ameliorate the unwanted CT learning effects that may be binding the adjacent hand-centred configurations as the density of hand-centred training locations is increased. As we discussed previously, CT learning tends to hinder cells learning to represent a single localised region. Instead cells eventually learn to represent a much larger region of space around the hand as the number of target locations keeps increasing. In this subsection we present simulation results using a connectivity that was further reduced by a factor of two. These simulations were carried out with either competitive networks or SOMs implemented within each layer in VisNet.

Model parameters

For these simulations the interlayer connectivity was divided by 2, maintaining however the same radius of connectivity. Table 3.5 shows the network dimensions used in this subsection.

In the case of the simulations using four layers of SOMs, the lateral inhibition and excitation parameters were the same shown in Table 3.4. The rest of the parameters were

	Dimensions	Number of Connections	Radius
Layer 4	32x32	50	12
Layer 3	32x32	50	9
Layer 2	32x32	50	6
Layer 1	32x32	50	6
Retina	128x128x16	-	-

Table 3.5: Network dimensions with the feedforward connectivity halved

the same as the ones described in Section 3.2.1.

Results

The further dilution of the network connectivity was carried out by dividing the inter-layer feedforward connectivity by a factor of 2, maintaining however the same radius of connectivity.

Figure 3.9 shows the results of the simulations using reduced connectivity in the standard version of VisNet incorporating a series of competitive networks. The top plot shows the single cells information plotted for all output layer neurons while the bottom plot shows the multiple cell information. In Figure 3.10 we show the information analysis metrics when VisNet is modelled as four layers of SOMs with reduced feedforward connectivity. In both cases it can be seen that diluting the connectivity actually worsens the performance of the network in both the standard competitive architecture and a SOM architecture. In this case the diluted connectivity is now also impairing the ability of CT learning to bind together the features of the hand-centred configurations across the different retinal locations.

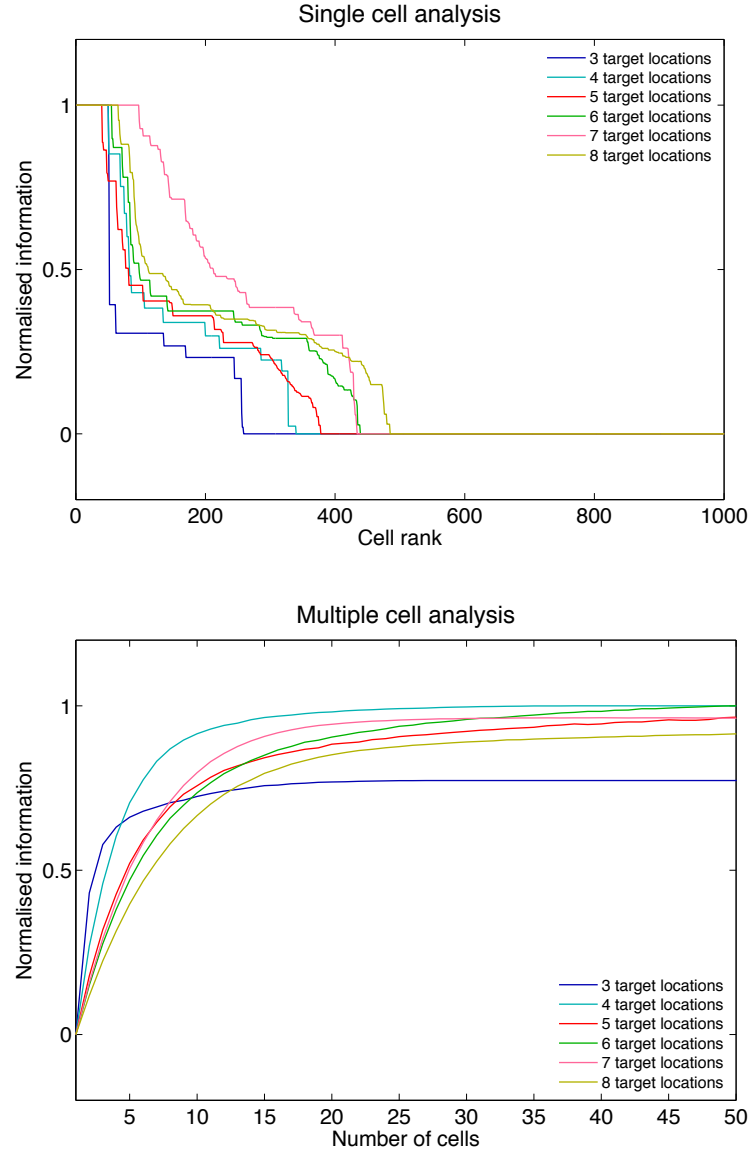


Figure 3.9: Information analysis for all densities using competitive network with further reduced connectivity. Analysis of the information conveyed by the output (fourth) layer neurons about the location of the visual target with respect to the hand. Values on the y-axis were normalised by the maximal information possible, $\log_2(\text{number of target locations})$, for each simulation. The single cell information analysis (upper plot) shows that in all cases there was a subset of neurons conveying the (normalised) maximal single cell information. However, the multiple cell information (lower plot) shows that as we start training the network on a continuum of target locations the multiple cell information does not asymptote to the (normalised) maximal value. This means that not all of the hand-centred configurations are represented selectively by the network.

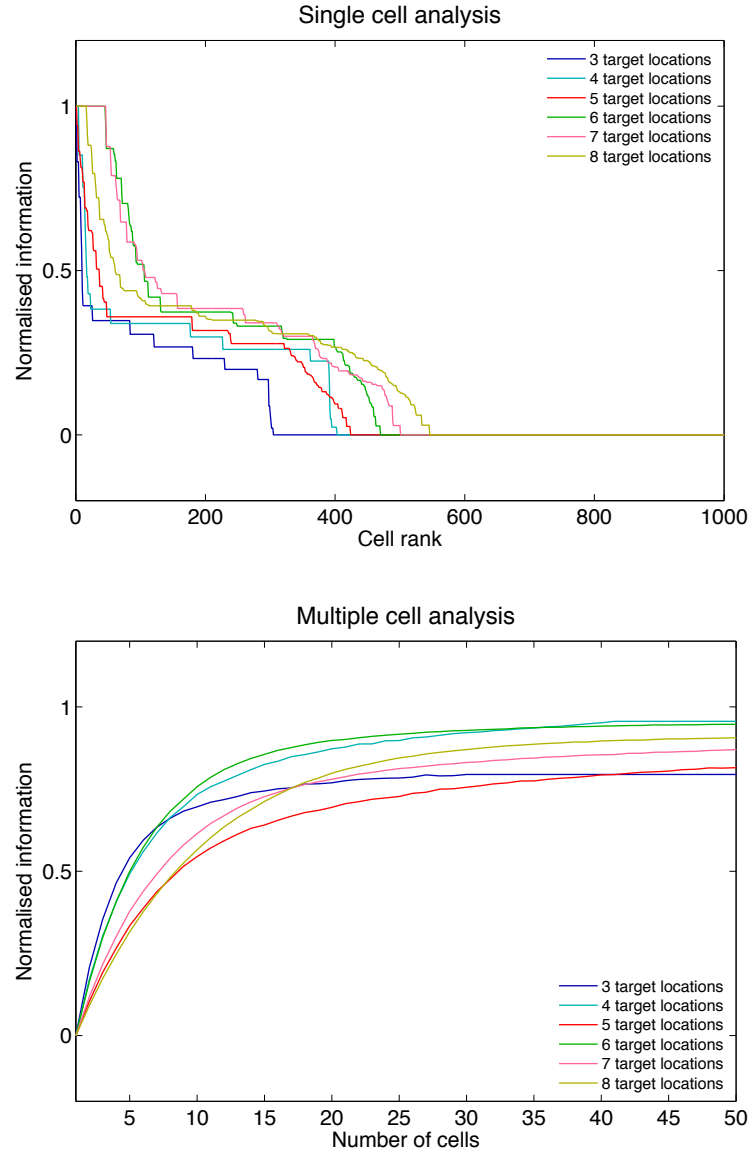


Figure 3.10: Information analysis for all densities using SOMs with further reduced connectivity. Analysis of the information conveyed by the output (fourth) layer neurons about the location of the visual target with respect to the hand. Values on the y-axis were normalised by the maximal information possible, $\log_2(\text{number of target locations})$, for each simulation. The single cell information analysis (upper plot) shows that in all cases there was a subset of neurons conveying the (normalised) maximal single cell information. However, the multiple cell information (lower plot) shows that in none of the test cases the network asymptotes to the (normalised) maximal value. This means that the hand-centred configurations are represented selectively by the network.

3.4 Discussion

The results presented in this study have explored how a self-organising neural network model, with a biologically plausible learning mechanism, could be capable of generating output cells that are tuned to specific target locations in a hand-centred frame of reference. The response profiles of cells within the network and the information measures from the base simulation initially suggested that CT learning could be a good candidate for the development of localised hand-centred representations.

However, as the density of hand-centred locations was increased (See section 3.3.2), it became clear that the highly associative aspect of this learning mechanism has unwanted deleterious effects. It was found that CT learning tended to produce output cells that responded to more than one target location with respect to the hand, and this was an early indication that CT learning may prove to be too associative when the model trained on a continuum of target locations.

As the density of locations is increased, the spatial overlap between different hand-centred configurations increases as well. The high associativity of Hebbian learning causes individual output cells to respond to multiple hand-centred target locations, rather than developing more localised responses. This is a fundamental problem because the network is also relying on CT learning to encourage output cells to learn to respond to particular hand-object configurations with translation invariance across different retinal locations.

In this chapter we attempted to improve the performance of the model by eliminating

this unwanted side effect of CT learning. We first introduced a SOM architecture within each of the layers of VisNet, which has been shown previously to dramatically affect the qualitative nature of the output cell firing responses and reducing unwanted CT learning effects (Tromans et al., 2012). However, the result showed only a marginal improvement.

Secondly, we reduced the connectivity of the feedforward synaptic connections to ameliorate the unwanted associative effects of CT learning. However, the results showed that the dilution of the connectivity actually worsened the performance of the network in both the standard competitive network architecture and a SOM architecture. In most cases this dilution was now impairing the ability of the cells to bind together the features of the hand-centred configurations across the different retinal locations. Altogether, these simulation results undermine CT learning as a potential mechanism for forming localised hand-centred representations.

However, there is a large variety of receptive fields sizes in the cells representing peripersonal space. While some of these cells have small and localised receptive fields, other cells have quite large receptive fields, in some cases even covering almost the entire body (Graziano and Gross, 1993). Moreover, in some atypical cases, these receptive fields could even extend to 100cm from the skin. Cells with body-centred or body-part centred receptive fields do not necessarily represent a localised region near the body, but could in some cases respond to a large region of space surrounding the entire hand or even large portions of the space surrounding the entire arm or the trunk. In this case, we could hypothesize

that CT learning could still be considered a candidate to account for the development of a continuous map of the space of hand-centred or body-part centered locations, in the case of cells with large receptive fields surrounding the body. Nevertheless, a CT learning mechanism would not be the best candidate for accounting for the development of more localised hand-centred or body-centred representations that have also been reported in the brain.

In the next chapter we show how VisNet could manage to develop more localised hand-centred representations in the output layer using a different learning mechanism called trace learning. Trace learning has been shown to be able to overcome the unwanted associativity problem of CT learning by binding together input representations that are temporally proximal instead of relying on spatial similarity. Furthermore, using a trace learning mechanism is not necessary for the network to be trained with the hand-object configurations across a continuum of fixation points. In other words, the training regime in a trace learning paradigm, would not require to show the network every possible smooth shift of the hand-object configuration (i.e. 1 pixel at the time). Instead, a trace learning mechanism would allow cells to learn to respond to particular hand-object configurations across larger retinal shifts that occur close together in time. Therefore, trace learning would allow a more flexible range of eye-movements during training, including larger saccades.

The supposition that the future resembles the past, is not founded on arguments of any kind, but is derived entirely from habit.

—David Hume, *Enquiries Concerning The Human Understanding and Concerning The Principles of Morals* (1737)

4

The development of hand-centred visual representations by trace learning

This chapter explores the ability of VisNet to develop hand-centred representations in the output layer using a different learning mechanism called trace learning. Trace learning is able to break the associativity of CT learning by matching together input representations that are temporally proximal instead of relying on spatial similarity. The results show that the trace rule encourages neurons at the end of the visual pathway to learn to respond to specific locations of a visual target object with respect to the hand, utilizing a more flexible range of eye-movements. The performance of the network is evaluated in different conditions, for example by increasing the density of target locations as well modifying the size of the retinal shifts. The performance of VisNet using layers of competitive networks is compared to VisNet’s performance using self-organising maps and more diluted

connectivity.

4.1 Introduction

As previously outlined, movement planning in the context of visually guided reaching seem to involve that the location of the target in retinotopic coordinates is transformed into a frame of reference that is appropriate for the motor command (i.e. a hand-centred or body-centred reference frame). Different parts of the posterior parietal cortex have been identified to code information in different reference frames. A gradient from eye-centred to body-centred neurons (the effector) is generally observed as the recording sites are moved from PRR towards cortical area 5 and premotor areas. Cells coding for targets in a hand-centred reference frame have been reported predominantly in Area 5d (Bremner and Andersen, 2012), as well as in other parts of the PPC and premotor areas (Chang and Snyder, 2010; Graziano et al., 1997).

The simulations presented in the previous chapter with Continuous Transformation (CT) learning, explored one way in which neurons in the brain might learn to represent the location of a visual target with respect to the hand. We assumed that the spatial continuity of the same hand-object spatial configuration shifting smoothly across the retina would allow the network to respond to a particular configuration throughout different locations and fixations. The response profiles of cells within the network and information measures suggest that CT learning is able to successfully represent only a small subset of

hand-centred representations. It is clear that the highly associative aspect of this learning mechanism has unwanted deleterious effects, especially as the density of hand-centred locations is increased. Considering that CT learning might not be the best mechanism to potentially explain the underlying the development of hand-centred responses, in this chapter we explore a different learning mechanism, using a *trace learning* rule (Földiák, 1991; Rolls, 1992). The trace learning rule incorporates a memory trace of recent cell firing activity, which has the effect of encouraging individual postsynaptic cells to learn to respond to a subset of input patterns that tend to occur close together in time. Therefore, rather than assuming a spatial continuity as utilised by CT learning, the trace rule relies on the assumption of temporal continuity.

We explore the hypothesis that trace learning in the feed-forward synaptic connections between successive neuronal layers in the network is able to encourage neurons at the end of the visual pathway to learn to respond to specific locations of a visual target object with respect to the hand, regardless of where the hand-object configuration appears on the retina, by exploiting natural eye movements including fixational eye-movements like drifts and microsaccades. Trace learning drives neurons in the later layers to learn to respond to input images that occur in temporal proximity. We assume that the eyes are continually performing rapid sequences of eye movements about any visual scene containing a fixed spatial configuration of a hand and object. Trace learning will then encourage individual cells to learn to respond to particular hand-object configurations across the

retinal shifts that occur naturally due to rapid sequences of eye movements (e.g. small drifts or microsaccades). The plausibility of this hypothesis is demonstrated in the computer simulations described below.

4.2 Hypothesis

Neurons have been found in multiple areas along the parieto-frontal network, that respond to the location of a visual target in a hand-centred frame of reference, irrespective of where the target is in the retinal frame of reference (Bremner and Andersen, 2012; Buneo and Andersen, 2006; Chang and Snyder, 2010; Graziano et al., 1997). How might these visually driven cells develop their interesting firing properties?

The central hypothesis of this chapter is that a form of trace learning rule may help neurons in these brain areas to learn to respond to the location of visual targets in a hand-centred frame of reference in the following manner. During early learning, the visual system is exposed to image sequences similar to those shown in Fig. 4.1. Each image sequence involves the target object shown in a fixed position with respect to the hand. The eyes are constantly performing movements around the visual scene. Even during fixation, a range of fixational eye movements are performed (e.g. drifts and microsaccades). This has the effect of creating image sequences in which each fixed spatial configuration of the hand and visual object is shifted across the retina. Consequentially, images of a target object in a particular position with respect to the hand, but occurring across different retinal

positions, will tend to occur close together in time. In this case, a trace learning rule may be able to associate the images within a particular temporal sequence, corresponding to one particular spatial configuration of the hand and object, with the same subset of output neurons. After enough training, individual output neurons will learn to respond to a particular relative spatial arrangement of the hand and the visual object across all possible retinal locations. This procedure can be repeated for all possible positions of the visual object with respect to the hand. Different output cells should learn to respond to different image sequences corresponding to different positions of the visual object with respect to the hand. We propose that this kind of learning may take place continually as the eyes are moving around the visual environment, even when the subject is not involved in a reaching task.

In the simulations presented below, we demonstrate how trace learning can lead to the development of individual output cells which respond to specific positions of the object with respect to the hand regardless of the retinal locations. We thereby show that trace learning is a potential mechanism for the development of the hand-centred cell firing responses observed in a subset of the PPC and premotor areas. Part of this work has been published in Galeazzi et al. (2013) and the potential integration of these hand-centred representations as part of a full motor network architecture have been discussed in Galeazzi and Stringer (2013).

4.3 Method

4.3.1 Model architecture and parameters

The simulations presented in this chapter utilized the VisNet model architecture described in Chapter 2, any modifications to the parameters will be specified in each subsection. The network dimensions used in these experiments are provided in table 4.1. The lateral inhibition parameters are reported in table 4.2. The sigmoid parameters used to control the global inhibition within each layer of the model are presented in table 4.3. These sets of initial parameters are the same as the ones that were used in Chapter 3.

	Dimensions	Number of Connections	Radius
Layer 4	32x32	100	12
Layer 3	32x32	100	9
Layer 2	32x32	100	6
Layer 1	32x32	100	6
Retina	128x128x16	-	-

Table 4.1: Network dimensions Network dimensions showing the number of connections per neuron and the radius in the preceding layer from which 67% of the connections are received

Layer	1	2	3	4
Radius, σ	1.38	2.7	4.0	6.0
Contrast, δ	1.5	1.5	1.6	1.4

Table 4.2: Lateral inhibition parameters

Layer	1	2	3	4
Percentile	99	98	88	95
Slope β	190	40	75	26

Table 4.3: Sigmoid parameters: The sigmoid parameters used to control the global inhibition within each layer of the model.

4.3.2 Trace learning

For these simulations we used a trace learning rule (Földiák, 1991; Rolls, 1992) to adjust the strengths of the feed-forward synaptic connections between the layers during training. The trace rule incorporates a trace $\bar{y}^\tau$ of recent neuronal activity into the postsynaptic term. The trace term reflects the recent activity of the postsynaptic cell. The effect of this is to encourage the postsynaptic cell to learn to respond to input patterns that tend to occur close together in time.

The parameter η may be set in the interval $[0, 1]$. For our simulations the trace learning η is set to 0.8. If $\eta = 0$ then the equation becomes the standard Hebb rule

$$\Delta w_{ij} = \alpha y_i^\tau x_j^\tau. \quad (4.1)$$

However, the version of the trace rule used in this chapter only includes the trace of activity from the immediately preceding timestep, as used in other studies (Rolls and Milward, 2000; Rolls and Stringer, 2001) for improving the performance of the standard trace rule and

enhancing the invariant representation. Thus, the trace rule now takes the following form

$$\Delta w_{ij} = \alpha \bar{y}_i^{\tau-1} x_j^\tau. \quad (4.2)$$

Neuronal mechanisms that might support trace learning in the brain have been previously discussed (Rolls, 1992; Wallis and Rolls, 1997). In particular, Evans and Stringer (2012) have shown how trace learning may arise naturally in a spiking neural network with long synaptic time constants. To restrict and limit the growth of each neuron’s synaptic weight vector, $\mathbf{w}_i$ for the i th neuron, its length is normalised at the end of each timestep during training as is usual in competitive learning (Hertz et al., 1991).

4.3.3 Stimuli

The stimuli of the hand and a circular object are similar to the ones used in the previous chapter. The training stimuli for this experiment consisted of computer-generated images showing a hand and one circular object in a particular spatial configuration near the hand on a grey background. The configurations were designed, arranged and aligned using MATLAB. Fig. 4.1 shows an example of some of the stimuli that were utilized in our baseline simulation.

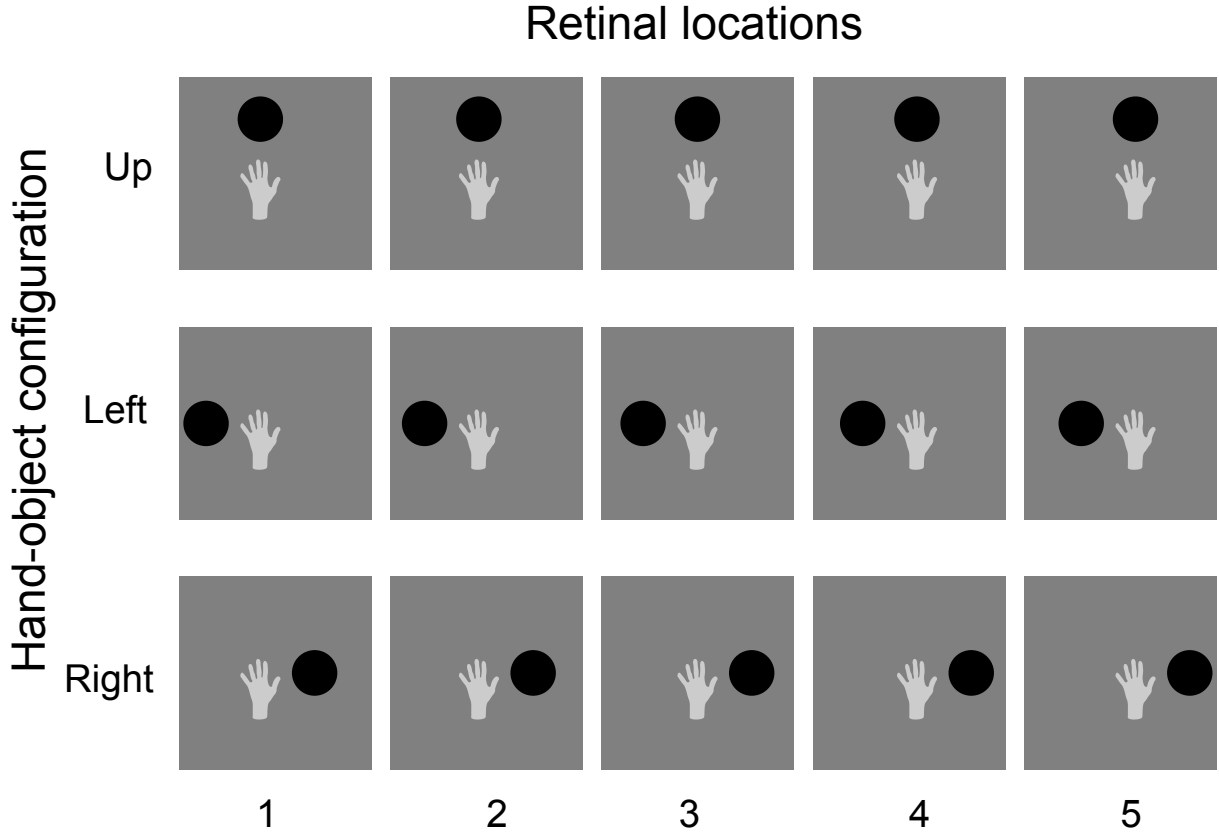


Figure 4.1: Image sequences. The three image sequences presented to VisNet during training are shown in the separate rows. All of the images in each sequence consist of a hand and a circular visual object. For each image sequence, the visual object is fixed in one of three possible positions with respect to the hand. Top row: visual object is shown in the ‘Up’ location with respect to the hand. Middle row: visual object is shown in the ‘Left’ location with respect to the hand. Bottom row: visual object is shown in the ‘Right’ location with respect to the hand. Each of the three image sequences arises from a series of eye movements that result in small shifts in the position of the hand and visual object on the retina (e.g. drifts or microsaccades). The relative positions of the hand and object are unchanged by these microsaccades. During each of the three image sequences, the fixed spatial configuration of the hand and object is translated five pixels at a time towards the right across columns 1 to 5.

4.3.4 Training

The three image sequences presented to VisNet during training are shown in the separate rows of Fig. 4.1. Each row displays a single sequence consisting of a set of five computer-generated images showing a hand and a circular visual object in a particular spatial configuration. The visual object is fixed in one of three possible positions with respect to the hand (Up, Left and Right). The images belonging to a particular sequence arise from a series of eye movements (e.g. saccades, drifts or microsaccades) and the resulting small shifts in the position of the hand and visual target on the 128×128 ‘retina’. During each of the three image sequences, the fixed spatial configuration of the hand and object is translated five pixels at a time towards the right across columns 1 to 5.

During the presentation of every image the activation of individual neurons and their firing rates are calculated and subsequently the synaptic weights are updated using a trace learning rule (Equation 4.2). The presentation of all three image sequences (i.e. Up, Left and Right) across all five retinal locations constitutes 1 epoch of training. The network is trained one layer at a time starting with layer 1 and finishing with layer 4. In the simulations described here, the numbers of training epochs for layers 1-4 were 50, 50, 50 and 50, respectively.

As we did in the previous chapter, we also investigate the effects of gradually increasing the density of hand-centred object locations. Throughout this chapter we will be presenting simulations in which we train the network with different numbers of hand-centred object

location. In all of the testing cases, the centres of the circular objects will be evenly distributed along a semicircle with respect to the centre of the hand (See Figure 4.7 and Section 4.4.4).

4.3.5 Testing

During testing, the synaptic weights are unchanged. In order to test whether VisNet has been able to develop translation invariance, image sequences of the hand and a circular object are presented in all of the trained retinal locations. The outputs of the network are computed during the presentation of each spatial configuration in each retinal location.

4.4 Results

4.4.1 Baseline simulation with three hand-centred target locations

The purpose of this baseline simulation was to demonstrate how trace learning can produce cell responses in the output layer of VisNet that are tuned to particular positions of a target object with respect to the hand, irrespective of retinal location. We studied the responses of the output (fourth) layer cells in VisNet before and after the network was trained on the image sequences shown in Fig. 4.1 as described above.

The response profiles of three neurons in the output layer of VisNet before training are

shown in Fig. 4.2. Each of the three columns shows the firing responses of one particular cell. The three rows show the responses of the cells to the three hand-object configurations across five retinal locations. The top row shows the cell responses when the visual object is in the ‘Up’ location with respect to the hand. The middle row shows the cell responses when the visual object is in the ‘Left’ location with respect to the hand. The bottom row shows the cell responses when the visual object is in the ‘Right’ location with respect to the hand. Fig. 4.2 shows that, before training, all three cells respond rarely or not at all to the different hand-object configurations.

Fig. 4.3 shows the response profiles of the same three neurons in the output layer of VisNet after training. It can be seen that, after training, each of the three cells has learned to respond to just one of the hand-object configurations, and responds to that configuration over all five retinal locations. The cell in the left column has learned to respond when the visual object is in the ‘Up’ location with respect to the hand. The cell in the middle column has learned to respond when the visual object is in the ‘Left’ location with respect to the hand. The cell in the right column has learned to respond when the visual object is in the ‘Right’ location with respect to the hand. Furthermore, each of the three hand-object configurations is represented by one of the cells.

To provide a more global measure of network performance, we analysed the information carried by the output (fourth) layer neurons in VisNet about which of the three hand-object configurations is presented to the retina. Intuitively, if an output cell has learned to re-

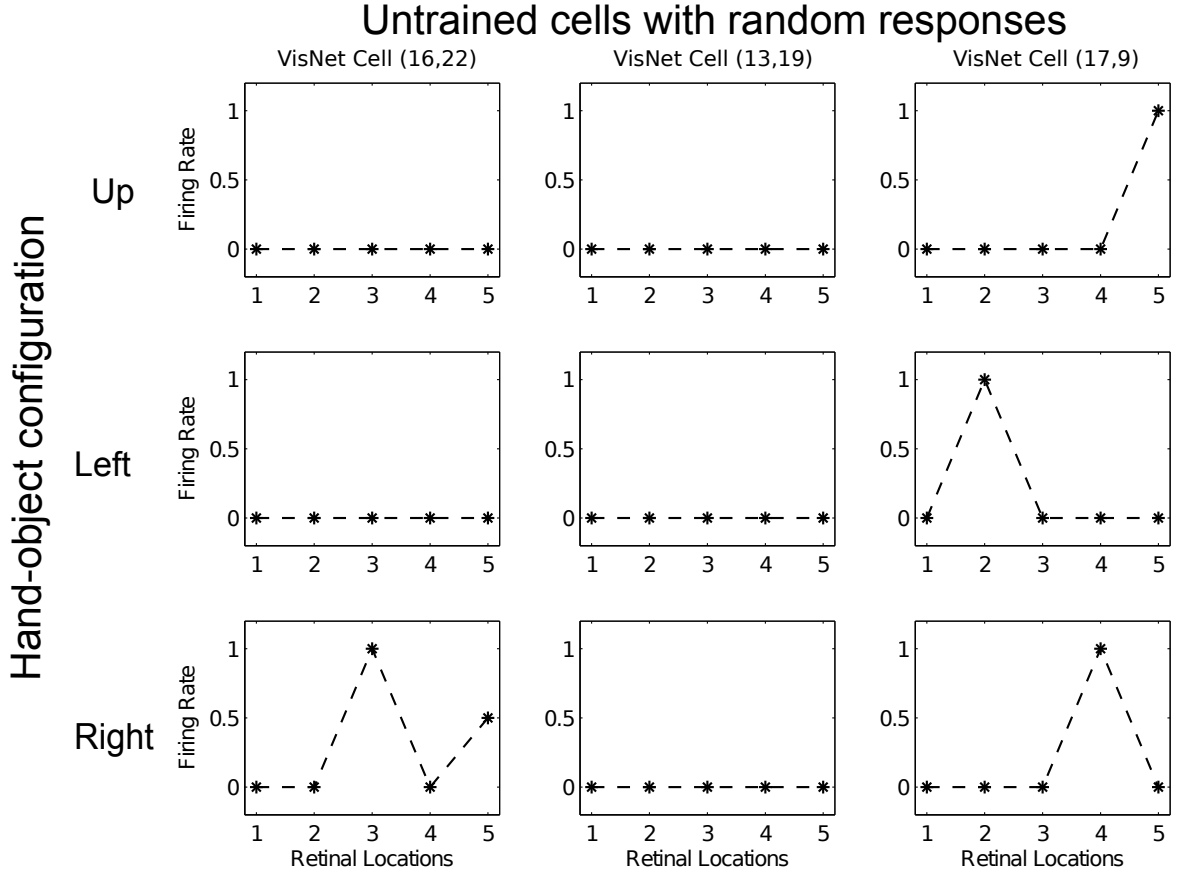


Figure 4.2: Response profiles of three neurons in the top (fourth) layer of VisNet before training. Each of the three columns shows the firing responses of a particular cell. Each row shows the responses of the cells to one of the three hand-object configurations over all five retinal locations shown along the abscissae. Top row: visual object is shown in the ‘Up’ location with respect to the hand. Middle row: visual object is shown in the ‘Left’ location with respect to the hand. Bottom row: visual object is shown in the ‘Right’ location with respect to the hand. It can be seen that each of the three cells initially responds rarely or not at all to each of the hand-object configurations over the different retinal locations.

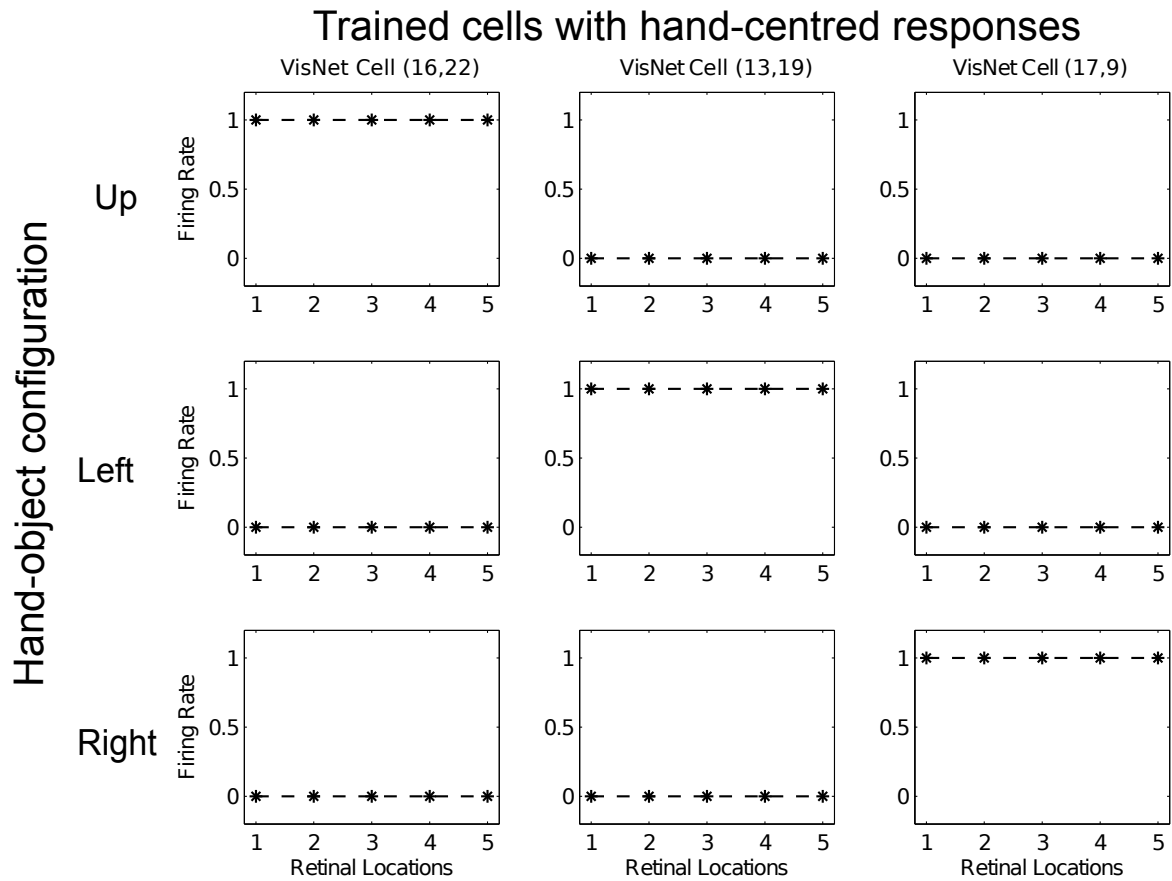


Figure 4.3: Response profiles of same three neurons in the top (fourth) layer of VisNet after training. Response profiles of the same three neurons from Fig. 4.2 after training on the images shown in Fig. 4.1. Conventions as in Fig. 4.2. It can be seen that each of the three cells responds selectively to just one of the hand-object configurations, and responds to that configuration over all five retinal positions shown along the abscissae. Moreover, each of the three hand-object configurations is represented by one of the cells.

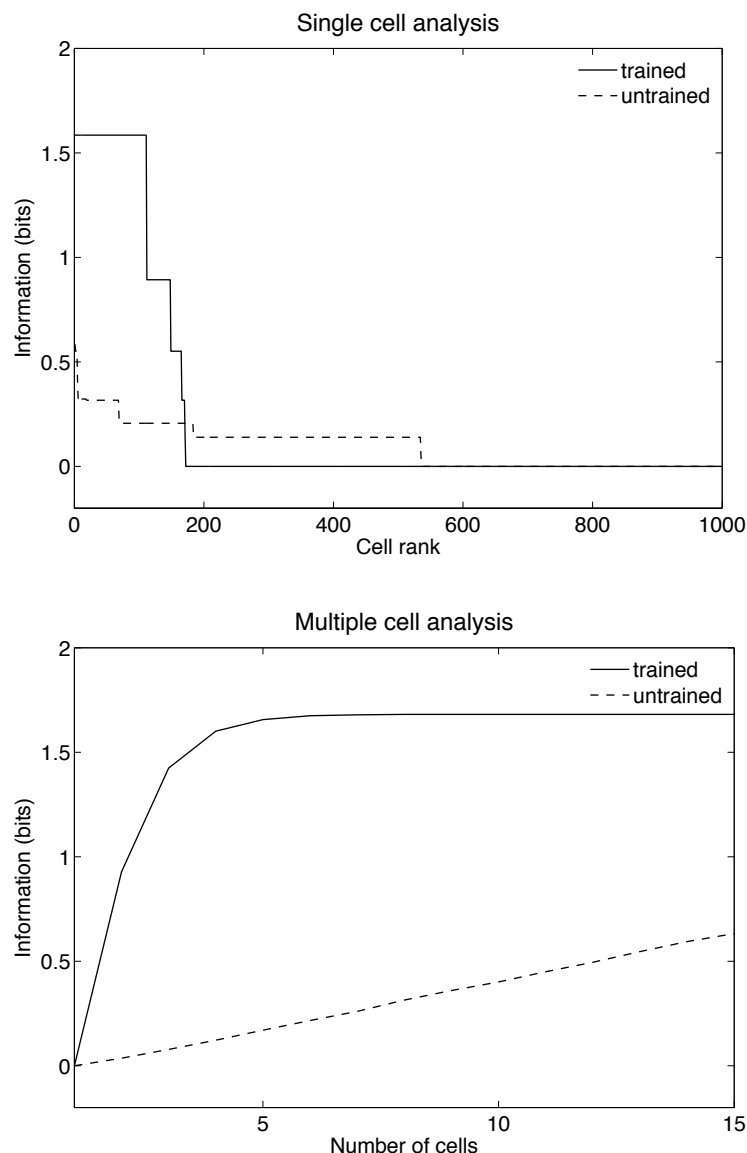


Figure 4.4: Information analysis. Analysis of the information about where the target object is with respect to the hand conveyed by the output (fourth) layer neurons before and after training. The upper plot shows the amount of single cell information carried by individual output cells in rank order. It can be seen that training the network has produced a large increase in the information carried by the output cells. In particular, after training, it was found that 111 cells reached the maximum amount of single cell information of 1.58 bits. In the untrained condition no cells reached maximal information. These cells responded perfectly to just one of the three hand-object configurations, and responded to that configuration across all five retinal locations. The lower plot shows the multiple cell information measures calculated across 15 cells with maximal single cell information. It can be seen that, after training, the multiple cell information asymptotes to the maximal value of 1.58 bits. This confirms that all three hand-object configurations are represented by the output cells.

spond perfectly to just one hand-object configuration over all five retinal locations, then it will convey maximal information about which hand-object configuration is currently presented. We, therefore, applied the single and multiple cell information measures described in Section 2.2.2 to the entire population of 1024 neurons in the output layer before and after training.

Fig. 4.4 shows the information measures for the output (fourth) layer neurons before and after training. On the left is shown the single cell information conveyed by individual output cells in rank order. Before training, no cells conveyed the maximal single cell information of 1.58 bits. However, after training, 111 cells had reached this level of single cell information. These cells responded to just one of the three hand-object configurations, and responded to their preferred configuration over all five retinal positions. The right plot shows the multiple cell information measures, which were calculated using 15 cells with maximal single cell information. After training, the multiple cell information is substantially increased and asymptotes to the maximal value of 1.58 bits. The multiple cell information results show that all three spatial configurations of the hand and object are represented by the output cells.

The simulation results described above confirm that trace learning can indeed produce learned neuronal responses which are tuned to a particular location of a visual object in a hand-centred frame of reference, as found in some neurons in the parietal cortex and premotor areas. The key observation is that, after training, the cells respond to specific

hand-centred locations regardless of retinal location. The trace learning rule has achieved this by encouraging output cells to learn to respond to images that tend to occur close together in time while the eyes are performing rapid (micro)saccades around the visual scene. Images of a particular configuration of hand and object presented across different retinal positions will tend to occur close together in time. In this case, a trace learning rule can associate all of the images of that spatial configuration with the same subset of output neurons.

4.4.2 Performance of model with different numbers of training epochs

Using the same framework from the previous simulation, we also examined the performance of the network as the number of training epochs was reduced. Fig. 4.5 shows single and multiple cell information analyses for six degrees of training: untrained, 1 epoch, 2 epochs, 5 epochs, 10 epochs, and 50 epochs. The single cell information analysis shows that already after the second epoch of training 57 output cells have achieved the maximum information content. Furthermore, the multiple cell information plot confirms that all of the three spatial configurations are represented by cells which are responding exclusively to one of the spatial configurations and not to any other. These results show that learning of all the spatial configurations occurs quite rapidly (e.g. already after 2 epochs) and remains stable as the number of epochs is increased. This means that only a few views of the

spatial configuration of the hand and the target need to be seen to start to develop an invariant response across the retina, making this learning mechanism efficient, flexible and ecologically plausible.

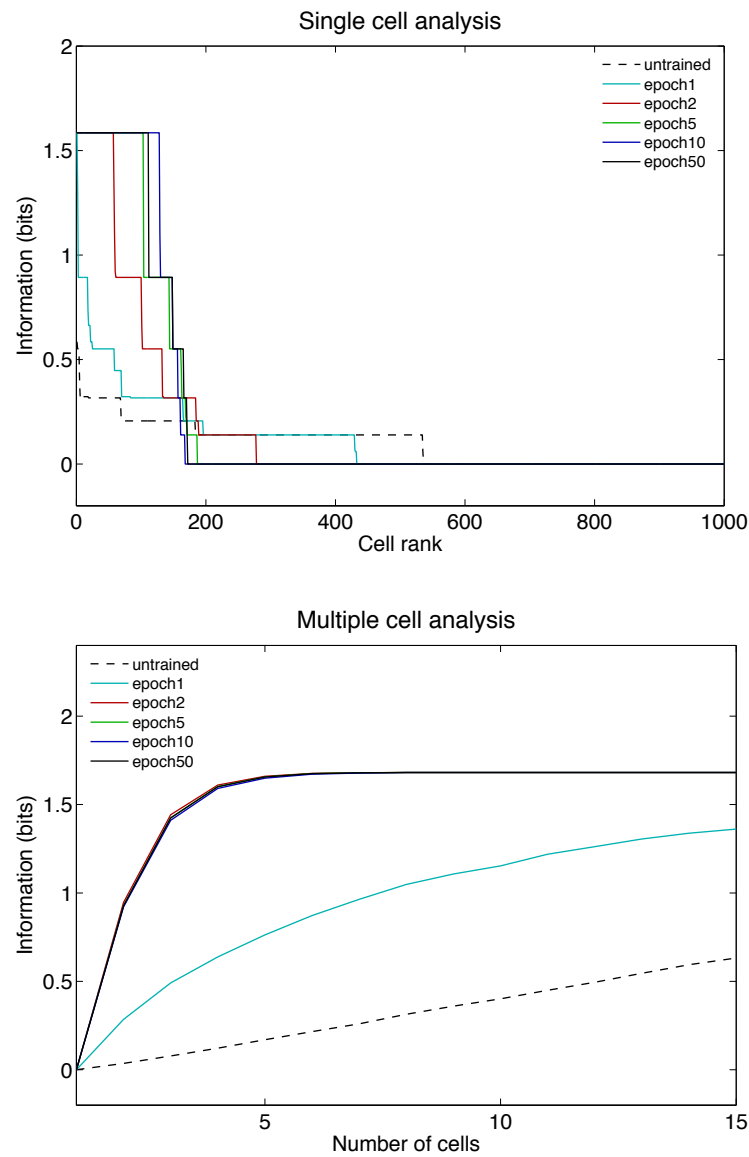


Figure 4.5: Information analysis for different training epochs. Analysis of the information about where the target object is with respect to the hand conveyed by the output (fourth) layer neurons for six degrees of training: untrained, 1 epoch, 2 epochs, 5 epochs, 10 epochs, and 50 epochs. The top plot shows the amount of single cell information carried by individual output cells in rank order. After the second epoch it was found that 57 cells reached the maximum amount of single cell information of 1.58 bits. The multiple cell information shown below asymptotes to the maximal value of 1.58 bits after the second epoch. This confirms that all three hand-object configurations are represented by the output cells.

4.4.3 Performance of the model with larger retinal shifts

In the simulations described above we used relatively small retinal shifts (i.e. five pixels). However, natural eye movements around any visual scene include larger saccades, which produce greater retinal shifts than the ones we have simulated. Therefore, it is important to test whether the model could learn to respond to a particular hand-object configuration across larger retinal shifts. We hypothesized that since trace learning relies only on the temporal proximity of the input patterns, this learning mechanism should be able to learn to respond to particular hand-object configurations across larger retinal shifts.

In this experiment we explore the performance of the model in the case of larger retinal shifts. However, to do this we needed to increase the size and hence resolution of the retina. In our previous simulations we used a 128×128 ‘retina’. In order to effectively simulate larger retinal shifts, we doubled the size of the model ‘retina’ (i.e. 256×256) and adjusted other network dimensions accordingly. The new dimension values of the network are given in Table 4.4. All of the other model parameters were the same as in the previous simulations. These dimension changes allowed us to produce larger eye movements that still ensured the hand and object appeared within the visual field.

We presented the network with three image sequences similar to the ones presented initially in Fig. 4.1. However, during each of the three image sequences, the fixed spatial configuration of the hand and object is translated 35 pixels at a time across the retina instead of 5 pixels.

	Dimensions	Number of Connections	Radius
Layer 4	32x32	200	48
Layer 3	32x32	200	36
Layer 2	32x32	200	24
Layer 1	64x64	200	24
Retina	256x256x16	-	-

Table 4.4: Scaled up retina: Network dimensions.

Network dimensions showing the number of connections per neuron and the radius in the preceding layer from which 67% are received.

Fig. 4.6 shows the information measures for the output (fourth) layer neurons before and after training. The single cell information analysis on the top shows that, after training, 119 neurons conveyed the maximal single cell information of 1.58 bits. After training, the multiple cell information shown on the bottom is substantially increased and asymptotes to the maximal value of 1.58 bits. This confirms that the three spatial configurations are each represented by cells which are responding exclusively to one of the spatial configurations and not to any other.

In the simulations described here, we can confirm that trace learning can produce learned neuronal responses which are tuned to a particular location of a visual object in a hand-centred frame of reference, even when relatively large eye movements are performed. Images of a particular configuration of hand and object, even if they occur across widely spaced retinal positions, can be associated by trace learning if they occur close together in time.

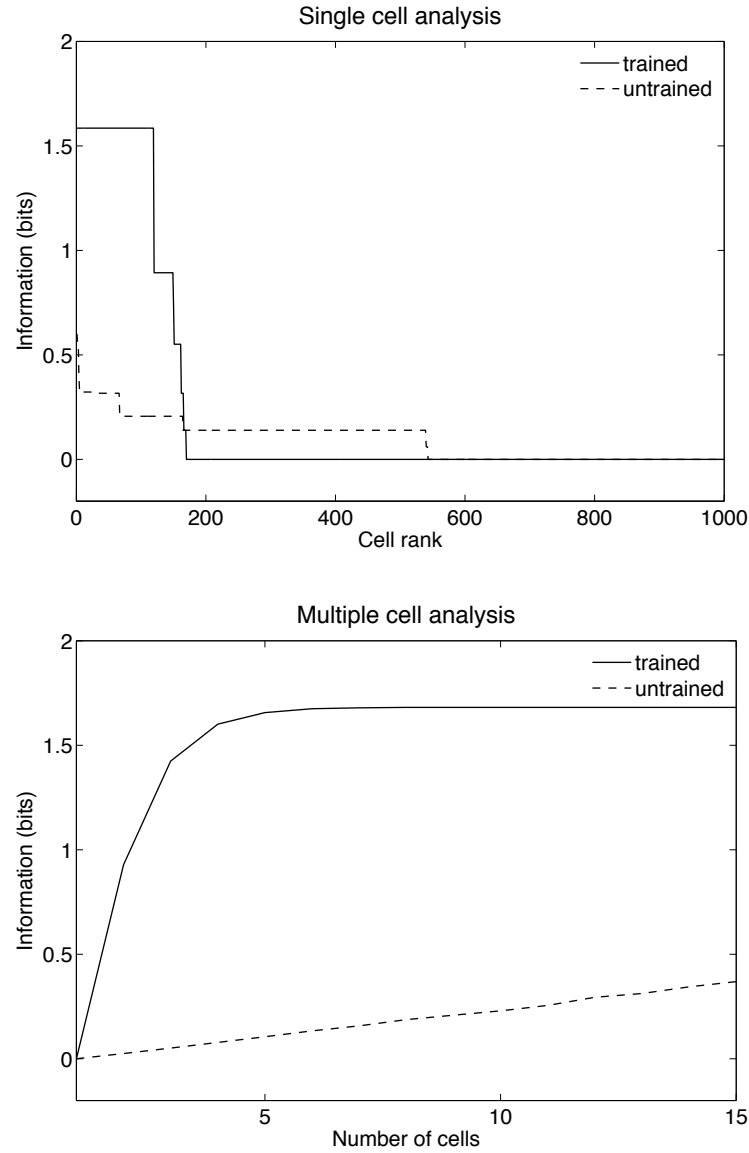


Figure 4.6: Information analysis for simulations with larger saccades over a scaled-up retina. Analysis of the information about where the target object is with respect to the hand conveyed by the output (fourth) layer neurons before and after training. The upper plot shows the amount of single cell information carried by individual output cells in rank order. After training, it was found that 119 cells reached the maximum amount of single cell information of 1.58 bits. These cells responded perfectly to just one of the three hand-object configurations, and responded to that configuration across all five retinal locations. The lower plot shows the multiple cell information measures calculated across 15 cells with maximal single cell information. It can be seen that, after training, the multiple cell information asymptotes to the maximal value of 1.58 bits. This confirms that all three hand-object configurations are represented by the output cells.

4.4.4 Capacity study: increasing the number of hand-centred target locations

A key issue is how the network will perform when the number of training locations is increased. In theory, the model should be able to represent a continuum of target locations with respect to the hand. In this section, we explored the performance of the model for eight test cases in which the number of training locations was gradually increased from three to ten in a semicircle around the central hand. The centres of all the objects were evenly distributed along the semicircle (diameter=36 pixels). The case of ten training locations is shown in Fig. 4.7. It can be seen that the ten locations effectively form a continuum of locations around the hand.

Fig. 4.8 shows single and multiple cell information analyses for all the test cases. The single cell information analysis shows that in all cases more than 100 neurons conveyed the maximal single cell information while the multiple cell information plot confirms that in all cases, the different configurations are each represented by cells which are responding exclusively to one of the spatial configurations and not to any other. In the simulations described here, the numbers of training epochs for all layers were 50, 50, 50 and 50, respectively.

Table 4.5 summarises for each of the eight test cases the number of cells that are perfectly selective to each one of the target locations with invariance over all five retinal locations. The results are the average of five simulations conducted with identical model



Figure 4.7: Image of the hand with 10 target locations. With 10 target locations around the hand, the targets are highly overlapping. In this case, the targets begin to form an effective continuum of hand-centred locations.

parameters, but with different random synaptic weight initializations and different random synaptic connectivities.

When the number of target locations was increased to 10, it can be seen from Fig. 4.7 that the density of locations was starting to approach an effective continuum. However, as the density of overlapping target locations is increased, it can be observed that some target locations were represented by only a few neurons. For example, in simulations with

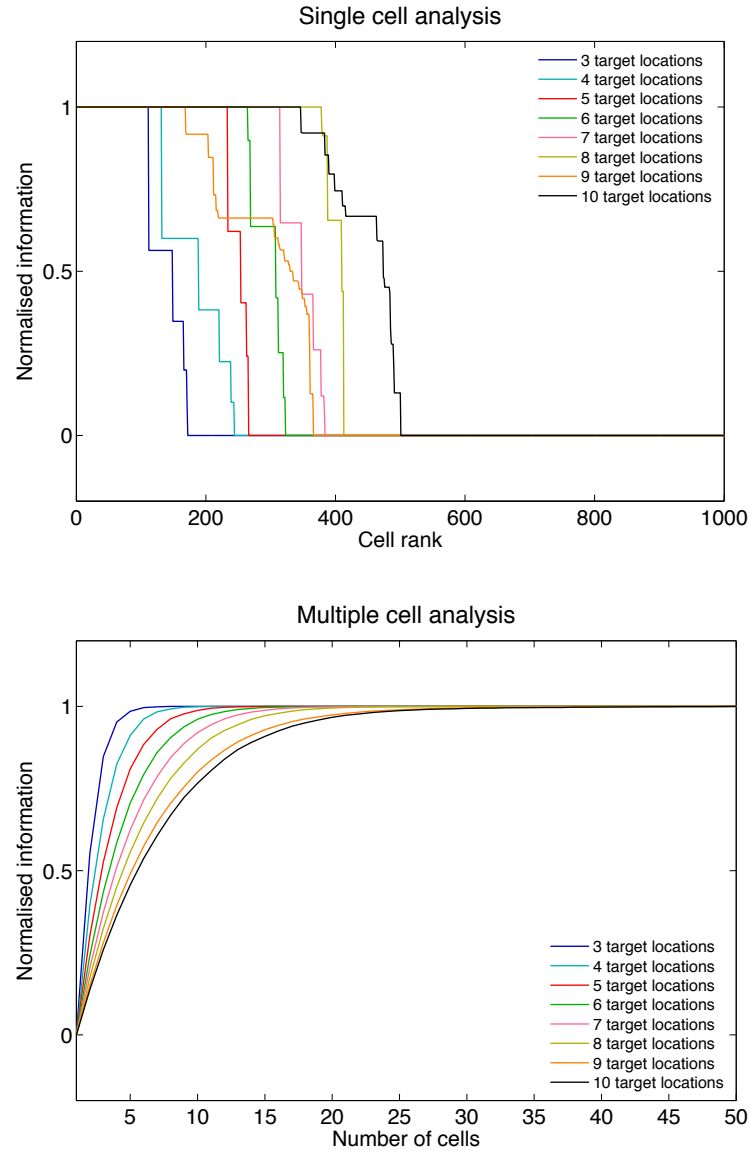


Figure 4.8: Information analysis for simulations with increased number of target locations. Analysis of the information conveyed by the output (fourth) layer neurons about the location of the visual target with respect to the hand. Values on the y-axis were normalised by the maximal information possible, $\log_2(\text{number of target locations})$, for each simulation. The single cell information analysis (top plot) shows that in all cases more than 10% of the neurons conveyed the maximal single cell information. The multiple cell information measures (bottom plot) were calculated using five cells for each stimulus that had the most single cell information about that stimulus. For simulations with three target locations this results in a population of 15 cells, while for ten target locations the multiple cell information analysis used 50 cells. It can be seen that in all cases the multiple cell information asymptotes to the (normalised) maximal value, confirming that all of the hand-object configurations are represented by the output cells.

10 target locations, target location 4 was only represented by an average of 2.8 neurons (see Table 4.5).

In the simulations with 10 target locations, there were some cells that had learned to respond to two adjacent hand-centred target locations, in addition to the cells that responded to only one hand-centred target location across all tested retinal locations. However, as the number of target locations was increased further, we found that the number of cells responding invariantly to a specific hand-centred location was also reduced, and in the test case of 20 target locations, VisNet starts to fail to represent all of the hand-centred locations in all retinal locations. At this point, individual cells responded to a cluster of adjacent hand-centred locations. Therefore, we expect individual cells to learn to represent a localised region of hand-centred space as the number of target locations goes to infinity. As we did in the previous chapter, in the next couple of subsections, we will explore the differences in the performance of the model using a SOM and more diluted connectivity.

Total number of target locations		Specific target location										
		1	2	3	4	5	6	7	8	9	10	Total
3	Mean	46.0	39.4	36.6	-	-	-	-	-	-	-	122
	SD	3.5	7.9	7.5	-	-	-	-	-	-	-	
4	Mean	44.0	25.6	35.0	30.2	-	-	-	-	-	-	134.8
	SD	8.6	15.7	9.4	13.8	-	-	-	-	-	-	
5	Mean	45.0	43.4	44.4	44.4	41.6	-	-	-	-	-	218.8
	SD	6.3	5.9	6.7	7.8	11.2	-	-	-	-	-	
6	Mean	42.2	32.8	31.0	27.0	40.2	40.4	-	-	-	-	213.6
	SD	10.4	19.2	24.7	18.0	21.5	14.7	-	-	-	-	
7	Mean	44.4	35.4	34.0	28.0	38.2	48.2	47.2	-	-	-	275.4
	SD	10.0	11.3	11.7	19.8	21.6	3.3	4.5	-	-	-	
8	Mean	39.8	42.6	36.2	10.6	10.0	35.4	40.8	47.2	-	-	262.6
	SD	21.3	13.5	15.7	17.8	17.9	22.3	16.1	5.5	-	-	
9	Mean	37.2	31.0	27.6	19.2	10.8	17.0	42.6	32.6	38.0	-	256
	SD	21.1	19.6	15.9	13.1	10.4	18.4	3.5	17.3	19.2	-	
10	Mean	39.8	29.4	36.6	2.8	15.6	6.6	29.0	37.4	46.4	42.6	286.2
	SD	7.4	15.0	15.4	3.0	21.7	9.2	23.7	21.7	4.3	6.9	

Table 4.5: Distribution of responses as the number of spatial locations is increased. This table shows the average number, over five simulations, of perfectly selective neurons responding to each specific spatial configuration of the hand and object across all five retinal locations. Each row corresponds to averaged results and standard deviations from simulations with a fixed total number of target locations. While each column refers to the number of perfect cells found for each specific target location.

4.4.5 Performance of the model using a SOM architecture

As we have mentioned in Section 3.3.3 of the previous chapter, a simple modification to the competitive networks introducing short range excitatory connections and long-range inhibitory connections, has the effect of clustering neurons with similar firing properties in the output layer. This occurs due to the mutual excitation properties of neurons that are adjacent, while neurons that are further away become inhibited, therefore generating separate pools according to the type of neuronal functionality.

In the previous chapter we assumed that a SOM architecture would encourage the output cells to learn to represent the continuum of hand-centred target locations across a spatial map in the output layer with a higher degree of selectivity. We showed that the effects of introducing a SOM in our CT learning simulations only showed a marginal improvement on the unwanted effects of CT learning, which caused output cells to learn to respond to unrealistically large regions of space around the hand, when the density of trained hand-centred locations was further increased.

For the simulations presented in this subsection, we again introduced short range excitatory connections between neurons within the higher neuronal layers, thus turning each competitive layer in VisNet into a SOM. In Section 4.4.4 we explored the performance of the model for eight test cases in which the number of training locations was gradually increased from three to ten in a semicircle around the central hand. Here we will explore how the network will perform when the number of training locations is increased even

further. In theory, the model should be able to represent a continuum of target locations with respect to the hand and to develop smooth hand-centred receptive fields in localised regions. However, we expect that a SOM architecture would produce even more selective output neurons with more localised receptive fields. Therefore, the performance of VisNet using a competitive network is compared with the performance using a SOM as the density of training locations is increased to 20 highly overlapping hand-object configurations.

Model parameters

In this section the VisNet model was modelled a series of four layers of SOMs. Table 4.6 shows the lateral inhibition and excitation parameters used for the SOM. The rest of the parameters were the same as the ones described in Section 3.2.1.

Layer	1	2	3	4
Excitatory Radius, σ_E	0.7	0.55	0.4	0.6
Excitatory Contrast, δ_E	5.35	33.15	117.57	120.12
Inhibitory Radius, σ_I	1.38	2.7	4.0	6.0
Inhibitory Contrast, δ_I	1.5	1.5	1.6	1.4

Table 4.6: Sigmoid parameters The sigmoid parameters used to control the global inhibition within each layer of the model.

Results

Figure 4.9 shows the single information analysis comparing the performance of the model in different test cases with up to 20 overlapping hand-centred locations. In these analyses

it can be seen that the performance of the SOM and the competitive networks are highly correlated and are not particularly different until the network is trained with the highest density of training locations. In the test case of 20 trained target locations, the SOM architecture is able to develop 302 neurons containing the maximum amount of information while the competitive networks develop only 88. More importantly, the multiple cell information plots in Figure 4.10 show that as the number of target locations is increased to 15 or more overlapping locations, VisNet's competitive network start to fail as the cells in the output layer do not reach the maximal amount of information in the multiple cell analysis. This means that the network is unable to represent all of the hand-centred configurations in every tested retinal location. However, it can be observed that a SOM network architecture is able to generate more output neurons responding selectively for specific hand-object configurations in all retinal locations. Although, in the test case of 20 target locations, the SOM is still unable to reach the maximal information in the multiple cell analysis (4.32 bits) because it fails to represent all the configurations in all retinal locations.

These findings imply that the SOM architecture is able to produce more selective output neurons, which have learned to represent smaller regions of the hand-centred space. However, as the number of hand-centred training locations approaches a continuum, the individual output neurons will eventually learn to represent localised regions of hand-centred space comprised of multiple adjacent hand-centred training locations.

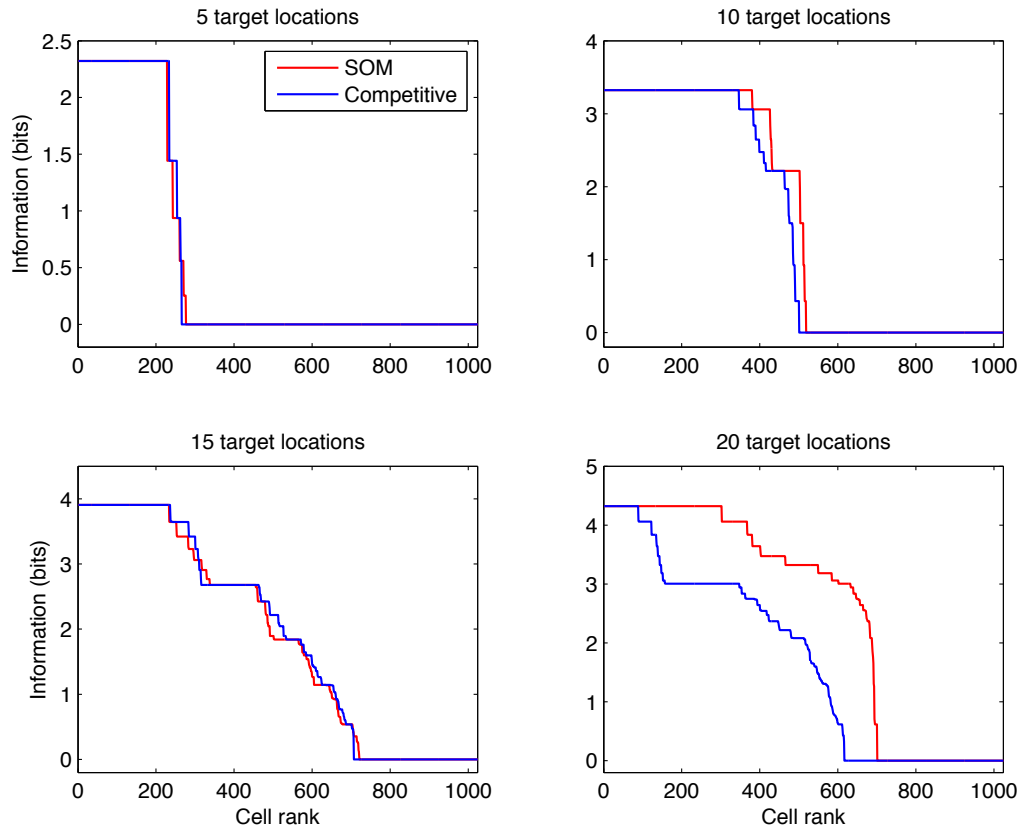


Figure 4.9: Information analysis comparing competitive networks with SOMs.

Single cells analysis results are shown for simulations with 5 target locations, 10 target locations, 15 target locations and 20 target locations. The single cell information analysis shows that in all cases more than 10% of the neurons conveyed the maximal single cell information in both the competitive network and SOM. However, as the number of training location reaches 20, more cells are found to convey the maximal information in the SOM network than the competitive network.

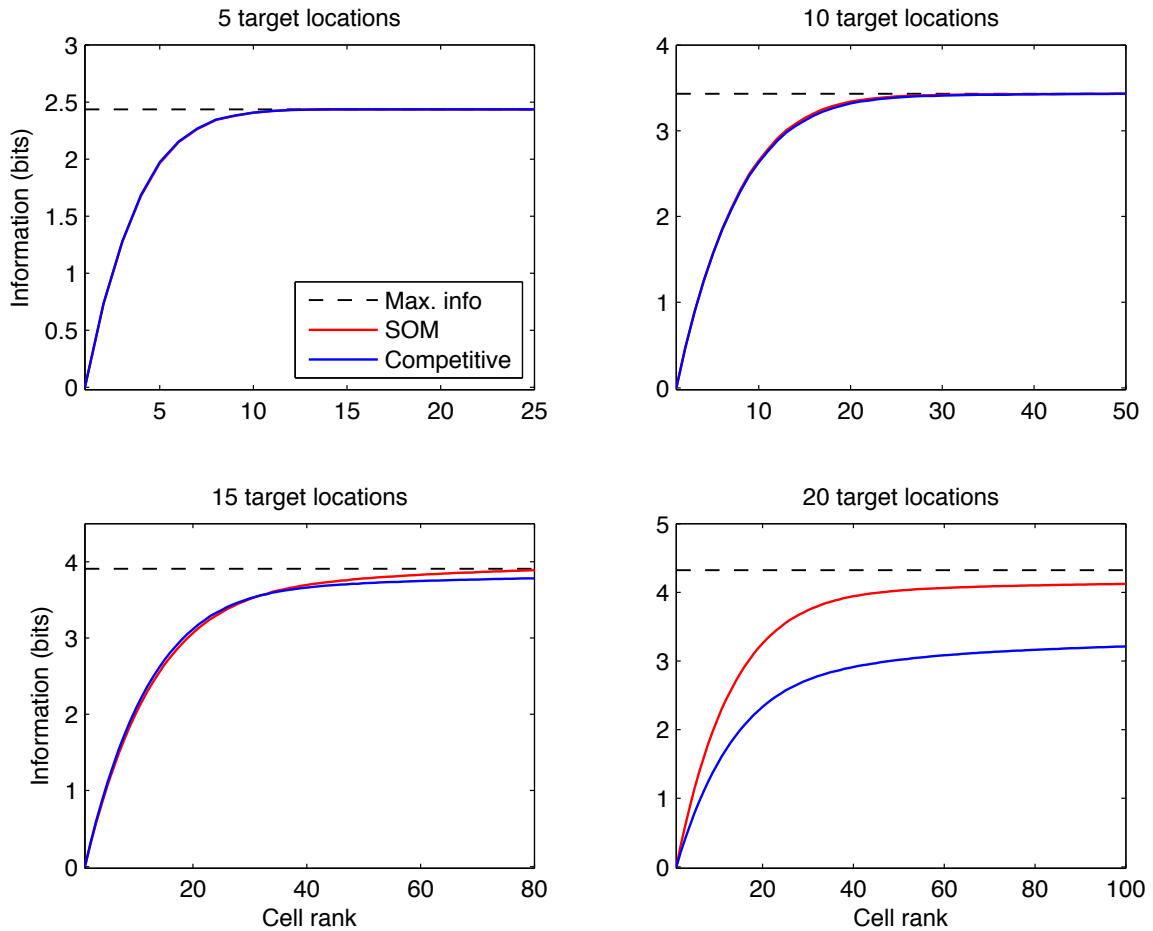


Figure 4.10: Information analysis comparing competitive networks with SOMs.

The multiple cell information shows that the competitive network (blue line) is able to reach the maximal value (dashed line) with 5 and 10 trained target locations, However, in the cases of 15 and 20 target locations, the competitive networks fails to reach maximal information. The SOM (red line) is able to improve the performance of the network by reaching maximal information in the case of 15 target locations, however it fails to reach maximal information at 20 target locations.

Effect of diluting connectivity in a SOM and a competitive network

The limited connectivity in VisNet, as opposed to a fully connected feed-forward network, is partly inspired by the sparse or diluted cortical connectivity in the brain. Having an architecture with a diluted connectivity in the feed-forward synaptic connections may also help to stabilize some of the ‘feature-analyzer properties of the neurons, by enabling each neuron to learn to respond to a subset of the possible input feature combinations (Rolls, 2008). The advantages of using diluted connectivity in different types of architecture has been explored extensively (Rolls, 2008, 2012a; Rolls and Webb, 2012; Treves, 1991). The architecture that we have been using already applies this principle of diluted connectivity.

In the previous subsection we hypothesized that a more diluted connectivity could ameliorate some of the unwanted CT effects, however we showed that diluting the connectivity actually worsened the performance of the network in both the standard competitive architecture and a SOM architecture. In that case the diluted connectivity was impairing the ability of CT learning to bind together the features of the hand-centred configurations across the different retinal locations.

Here we hypothesize that further dilution of the network connectivity will not necessarily affect the performance of the network under a trace learning paradigm as it did for the CT learning simulations, especially because the network is already capable of developing localised receptive fields. The further dilution of the network connectivity was carried out

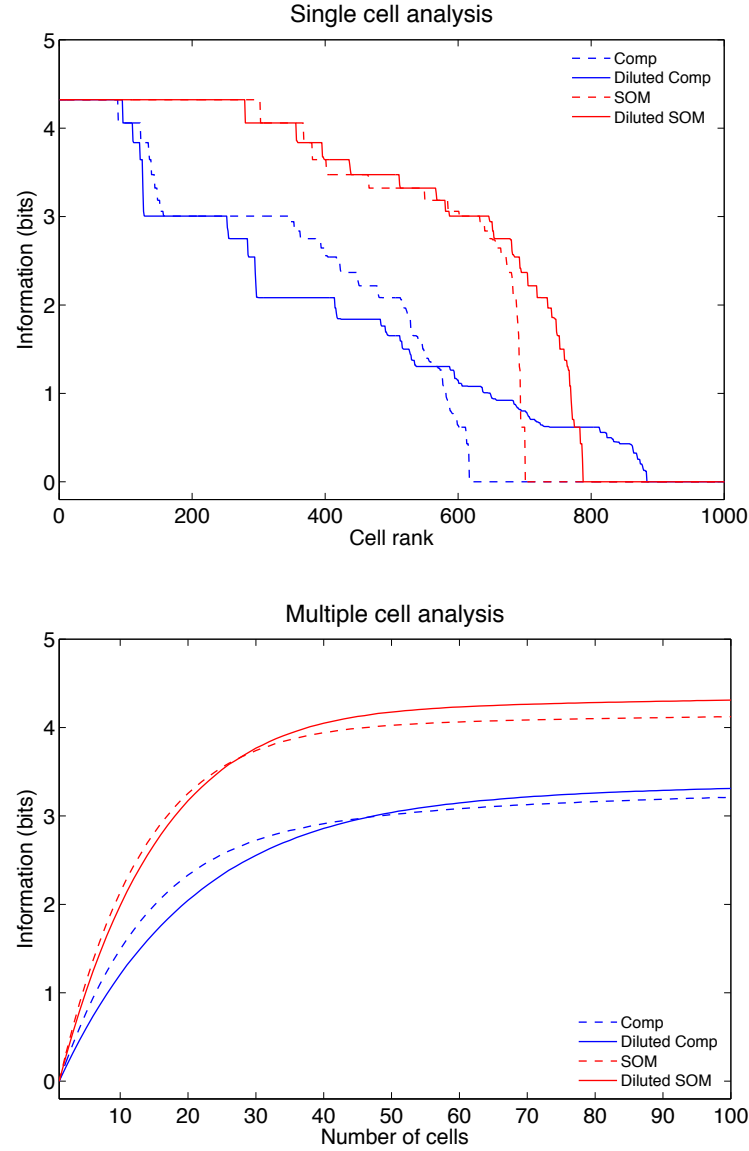


Figure 4.11: Information analysis with diluted connectivity. Analysis of the information about where the target object is with respect to the hand conveyed by the output (fourth) layer neurons before and after training. The upper plot shows the amount of single cell information carried by individual output cells in rank order. After training, it was found that using a SOM with diluted connectivity performs the best, allowing 279 cells to reach the maximum amount of single cell information of 4.32 bits. The lower plot shows the multiple cell information measures calculated across 100 cells with maximal single cell information. After training, the multiple cell information asymptotes to the maximal value of 4.32 bits only when a SOM is used with diluted connectivity, being the only case where all hand-object configurations are represented by the output cells.

by dividing the intralayer connectivity by a factor of 2, maintaining however the same radius of connectivity. The results of the simulations are shown in Figure 4.11 which compares the information analysis metrics of simulations with either competitive networks or SOMs implemented within each VisNet layer, and with either standard diluted connectivity or connectivity further reduced by a factor of two. The single cell information measures reveal as expected that using a SOM (red) or a further diluted version of the SOM (red dashed line) develops more hand-centred cells containing the maximum amount of information than the competitive (blue) and the diluted competitive network (blue dashed line). The multiple information measures show that only the SOM network with further diluted connectivity is capable of reaching the maximal amount of multiple cell information. In this simulation, the effect of further diluting the connectivity marginally improved the performance for the SOM architecture, allowing it to fully represent the whole density of hand-centred configurations in all tested retinal locations. However, the effect is not particularly large.

4.5 Discussion

In this Chapter we have presented a second self-organising hypothesis using a learning mechanism called *trace learning* (Földiák, 1991; Rolls, 1992). Here we incorporated a memory trace of recent neuronal activity which encourages neurons to learn to respond to input patterns that tend to occur close together in time. This learning mechanism

enables the binding of successive retinal images corresponding to the same hand-object configuration, thus forming localised hand-centred output representations.

The results presented in this chapter have shown how a self-organising neural network model with a biologically plausible learning mechanism is capable of generating output cells that are tuned to specific target locations in a hand-centred frame of reference. The network is able to learn these representations in just a few epochs, and the use of temporal continuity instead of a spatial continuity allows the network flexibility with respect to the nature of the eye-movements used to explore the visual world. Moreover, the simulations reported in this chapter showed that the trace learning mechanism is robust enough to represent several hand-object configurations even when they are overlapping. Thus, the network is able to develop localised hand-centred receptive fields even as the number of hand-object configurations increases during training.

Overall, the response profiles of cells within the network and information measures under the different conditions validate trace learning as the potential mechanism underlying the development of localised hand-centred representations. The hypothesis presented in this chapter provides a novel computational account for how neurons responding in a hand-centred frame of reference might develop in a biologically plausible way by unsupervised visually-guided learning.

In the following chapters the trace learning hypothesis is advanced by gradually increasing the realism of the visual training conditions. In the next chapter, the network will

be trained using more natural visual scenes consisting of multiple targets presented at the same time around the hand instead of just an isolated object in a particular location with respect to the hand.

”The hand is the cutting edge of the mind”

—Jacob Bronowski, *The Ascent of Man* (1973),
115-6.

5

The development of visual hand-centred representations under more realistic training conditions

This chapter explores VisNet’s ability to develop hand-centred representations under more realistic training regimes. Contrary to the previous simulations, the visual scenes now consist of multiple targets presented at the same time instead of just an isolated object in a particular location with respect to the hand. Three sets of simulations are presented in this section. First, VisNet is trained with idealised images consisting of pairs of circular objects presented simultaneously near the hand. The network is then tested with just one circular object near the hand in order to verify if the output cells are capable of building hand-centred representations with a single localised receptive field. I explain the underlying principles of the *statistical decoupling* that allows the output cells of the network to develop

single localised receptive fields even when the network is always trained with more than one object close to the hand. The experimental results show how the performance of the network improves as the number of object pairings is increased, which in turn increases the statistical decoupling between any pair of object locations around the hand. In the second set of simulations, I applied these same principles to train VisNet with more realistic images of a hand and with a variety of natural objects, as well as adding a subset of simulations exploring if the hand-centred receptive fields would develop any shape selectivity. Lastly, I explored these principles using natural visual scenes collected by participants. This is an important step in order to show how single, localised, hand-centred receptive fields could emerge in more ecological settings during visual training.

5.1 Introduction

The simulations presented in previous chapters involved showing only a hand and single circular object at any one time during training. Much of the past research done with VisNet has also consisted of presenting objects in isolation to the network (e.g. Rolls and Milward, 2000). However, in the real world we rarely encounter one object at the time, in fact, our visual system is mostly confronted with a complex environment consisting of multiple objects. Previous research with VisNet (Stringer and Rolls, 2000) revealed how challenging it is for the network to build invariant representations when it is trained on cluttered backgrounds. After training on natural backgrounds and testing on plain

backgrounds, the cells seemed to be responding mostly to the features of the background and not the actual stimuli (Rolls, 2012b; Stringer and Rolls, 2000). In these early studies, however, the background was represented as static images composed of foliage. More recently, a novel approach has been taken, in which the visual scene consists of multiple changing objects instead of a static background.

Following this new approach (Stringer and Rolls, 2008; Stringer et al., 2007) it was shown that neurons in the output layer are able to learn to represent individual objects even when a network is trained in a more complex visual setting with more than one object presented simultaneously. In a first study, a one-layer competitive network was trained on all possible pairings of objects (Stringer et al., 2007). It was reported that the network was able to form independent representations of each individual object even when they were never seen in isolation. This works when there is a sufficiently large number of objects. Presenting the network with every possible pair of objects during training leads to a statistical decoupling between any two of the objects. That is, any two objects will be seen together only rarely. This prevents individual neurons in the output layer from learning to respond to the pairs of objects actually seen during training. Instead, the neurons are forced to learn to respond to the individual objects themselves.

However, in order to allow this statistical decoupling to occur, the pool of parent stimuli had to be more than five and every possible pair had to be presented, otherwise the network was unable to represent separately each individual stimulus and instead the

network would learn to encode the object pairs themselves. Stringer and Rolls (2008) explored this principle of statistical decoupling further using VisNet. The study showed that VisNet is capable of forming separate object representations in the output layer if there is a sufficient level of 'statistical decoupling' between a large number of objects that are presented as random pairs during training. Additionally it was shown that VisNet can develop transform invariant responses of the individual objects by using a trace learning rule that would allow the network to bind together the different views of a particular object that occur close together in time.

The first part of this chapter explores whether VisNet is able to build separate representations for individual hand-centred locations, even if the network is trained with pairs of objects appearing simultaneously near the hand. In other words, the network is never trained with an individual object presented independently in each of the hand-centred locations, as in previous chapters (i.e. one hand-centred location at a time). Thus, we explore the hypothesis that this can be solved by the ability of competitive neural networks to form representations of individual visual stimuli which are seen together in different combinations at different times (Stringer and Rolls, 2008; Stringer et al., 2007). We explore whether increasing the number of pairings of hand-centred object locations which increases the statistical decoupling between object locations drives the development of cells with single, localised hand-centred receptive fields. Furthermore, the network was trained with particular spatial configurations of the hand and pairs of objects presented in

different retinal locations. By adding a temporal trace of the previous activity of each cell to the synaptic learning rule, we enabled the network to form hand-centred representations that are invariant to retinal location.

Previous research using pairs of objects did not include the presence of the hand in the images. This is not trivial, especially considering that deleterious CT effects can arise due to the visual presence of the hand in all of the images during training, enabling CT learning to bind together different hand-centred locations by the same cell. This could have seriously degraded the hand-centred location specificity of neurons. Nevertheless, we show that in fact after training, the statistical decoupling enables output cells to form separate representations of different hand-centred locations even if they are always exposed to visual scenes containing pairs of visual objects around the hand.

5.2 Methods

The experiments presented in this chapter were conducted using VisNet. Any changes to the parameters will be specified in each subsection. In contrast with the simulations in the previous chapters, the spatial configurations presented during training are now composed by a hand and pairs of non-overlapping circular objects presented simultaneously to VisNet in a semicircle around the hand (for more details see Section 5.2.1). Each configuration is then presented in different parts of VisNet’s retina. The translation of the hand-object configuration across the retina is meant to arise from a series of eye movements exploring

the visual scene. A temporal trace of recent neuronal activity is included in the synaptic learning during the visual exploration of a particular configuration of the hand and two circular objects, according to the trace learning rule described in Equation 4.2. The trace rule is used to associate together different retinal images corresponding to the hand-object configurations seen in different retinal locations.

After training, the network is tested with images consisting of the hand and a single object presented in a fixed spatial configuration across different retinal locations. The network is tested on all such hand-object spatial configurations across all retinal locations. These results were analysed to determine whether individual output cells had developed single, localised, hand-centred receptive fields, rather than responding to multiple (e.g. 2) hand centred locations as seen during training. Four different experiments are simulated in which the effect of statistical decoupling is assessed as the number of object pairings is increased from 4 to 7.

5.2.1 Stimuli

The training stimuli for this experiment consisted of computer-generated images showing a hand and two circular objects in a particular spatial configuration near the hand on a grey background. The centres of all the objects were evenly distributed along a semicircle with respect to the centre of the hand. The configurations were designed, arranged and aligned using MATLAB.

The circles were presented in all possible pairings of object locations. The number of possible pairings of objects taken from a pool of n objects is given by

$${}_nC_r = \frac{n!}{r!(n-r)!} = \binom{n}{r} \quad (5.1)$$

Where r is set to 2.

Figure 5.1 shows an example when $n = 4$. Here there are $({}_4C_2)$ six training images, with each composed of the hand and a pair of hand-centred target locations. If for example, we are training the network with a set of $n = 7$ hand-centred target locations, there would be a total of $({}_7C_2)$ 21 pairings, each containing two different target locations with respect to the hand simultaneously. In Table 5.1 we can see the total number of object pairings for each of the four test cases $n = 4, 5, 6$ & 7.

In order to present the configurations of the hand and objects in different retinal positions, each of the paired spatial configurations is translated by five pixels at a time across VisNet's retina. In Table 5.1 we can see the total number of retinal locations trained and the total number of stimuli presented to the network in each of the four test cases ($n = 4, 5, 6, 7$). Evidently, as the total number of target locations n is increased, both the number of possible pairings ${}_nC_2$ and the total number of stimuli is increased accordingly.

The testing stimuli consisted of images showing the hand and just one circular object with respect to the hand. The objects were presented during testing at the same hand-centred locations used during training.

n	${}_nC_2$	Retinal locations	Total No. of stimuli
4	6	5	30
5	10	5	50
6	15	5	75
7	21	5	105

Table 5.1: Training stimuli consisting of object pairs. The table shows the increase in the total number of object pairings as the number of training objects n is increased

5.2.2 Model parameters

In this experiment we used a scaled up version of the model ‘retina’ (i.e. 256×256) and adjusted other network dimensions. These dimension changes were to increase the resolution of the training images and to enable that all pairs of object locations could appear on the retina in a non-overlapping semi-circle around the hand. The network dimensions and the values of the parameters are the same as used in Section 4.4.3.

5.2.3 Training

The training procedure for these experiments consisted of presenting VisNet with pairs of non-overlapping circular objects displayed in a semicircle around the hand. During training, the circular objects were presented in pairs and never in isolation (see Figure 5.1).

Each of the images belonging to a particular paired configuration is trained across five different retinal locations. The image sequences are meant to arise from a series of eye

movements (e.g. saccades, drifts, microsaccades) and the resulting shifts that this would generate in the position of the hand and pair of objects on the 256×256 ‘retina’. During each of the image sequences, the fixed spatial configuration of the hand and pair of circular objects is translated five pixels at a time.

During the presentation of every image the activation of individual neurons within a layer and their firing rates are calculated. Subsequently, the synaptic weights between layers are updated according to the trace rule as in the previous chapter (see Equation 4.2). This process is repeated for all layers of the network. As per the standard VisNet simulation protocol, the network is trained one layer at the time, starting with layer 1 and finishing in layer 4.

One training epoch consisted of the presentation of every possible object pairing, each across the five different retinal locations. Figure 5.1 shows an example of the six training images for $n = 4$ objects composed of the hand with a pair of hand-centred target locations. In these simulations the network was trained over fifty epochs per layer. The learning rates used were 0.1, 0.1, 0.1 and 0.1 in each layer.

5.2.4 Testing

Throughout the testing phase, the synaptic weights are unchanged. In order to test whether VisNet has been able to develop translation invariance, image sequences of the hand and only one circular object are presented in five different retinal locations.

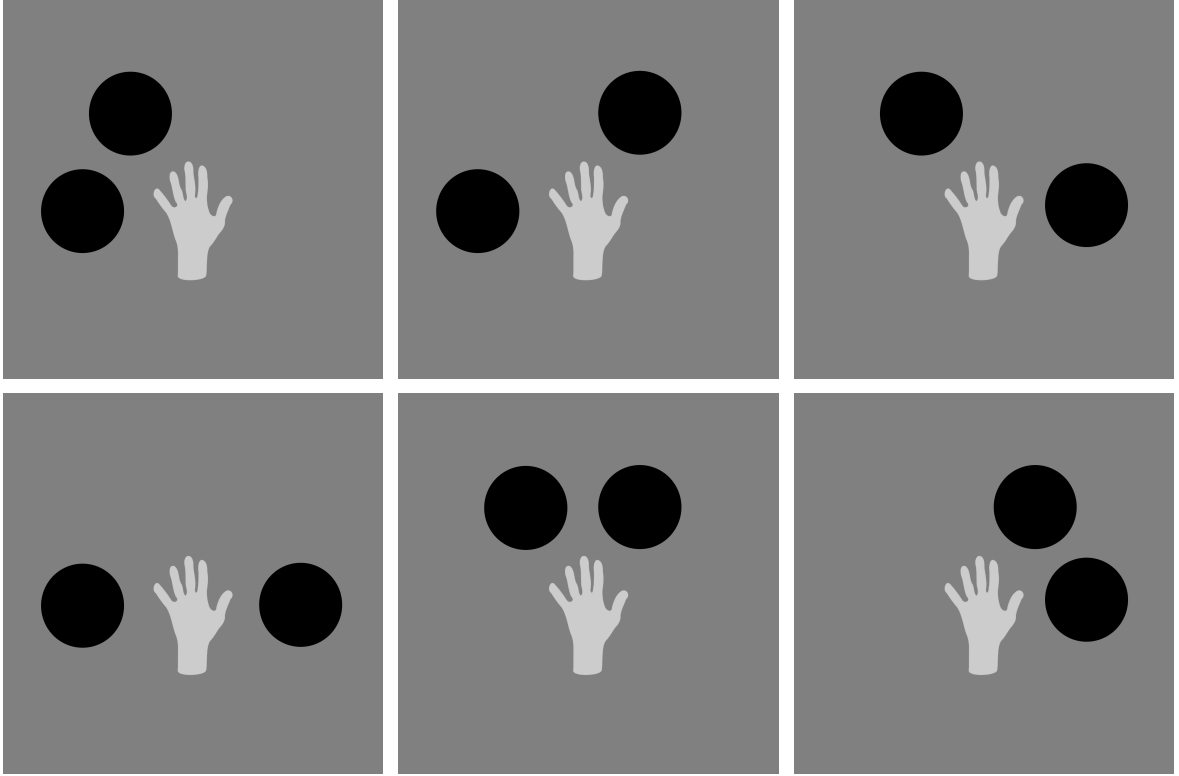


Figure 5.1: Paired target locations. Each of the six figures represents one of the possible pairings when $n = 4$. For each of the six images an image sequence was generated arising from a series eye movements that result in shifts in the position of the hand and visual objects on the retina. The relative positions of the hand and the pair of objects are unchanged during these eye movements. The fixed spatial configuration of the hand and objects is translated across five different retinal locations.

The outputs of the network are computed during the presentation of each spatial configuration in each retinal location. Figure 5.2 shows an example of four testing images when $n = 4$, each of these images is presented in five different retinal locations.

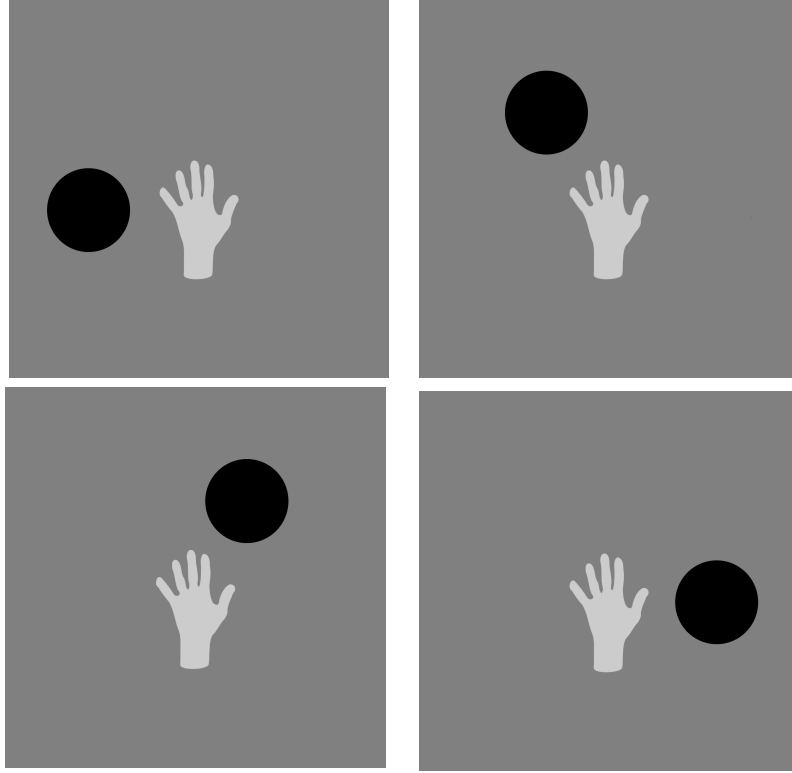


Figure 5.2: Testing stimuli. The four figures represent the hand-object configurations that were tested when $n = 4$. Each of the four fixed spatial configurations of the hand and object was presented at the five different retinal locations in which the network was previously trained with the object pairings.

5.3 Results

After the network was trained on all possible object pairings, the network was tested to determine whether individual cells in the output layer had learned to encode single hand-centred locations across the different retinal locations. Information analysis was conducted for each of the test cases. As in previous chapters we show the single-cell and multiple-cell analyses. The information we are measuring corresponds to whether the cell responds selectively to a single position of the object with respect to the hand across the tested

retinal locations.

Fig. 5.3 and 5.4 shows the information measures for the output (fourth) layer neurons after training with pairs of hand-object configurations. Each of the coloured solid lines correspond to the 4 test cases ($n = 4, 5, 6, 7$).

The single cell information analysis shows that, after training, neurons conveying the maximal single cell information could be found in all of the test cases. This means that there were output cells that were tuned to individual positions of a target object with respect to the hand, irrespective of retinal location. However, in order to make sure that all n spatial configurations of the hand and object location are represented by different neurons, we need to conduct the multiple cell analysis.

The multiple cell information analysis shows that, after training, the multiple cell information does not reach the maximal value for the test cases where $n \leq 5$. This means that even though in the single cell analysis we find cells that are able to represent individual hand-object configurations, the network has not been able to develop separate output cells for every one of the n spatial configurations. However, in the cases where $n \geq 6$ multiple cell information asymptotes to the maximal value, which means that all of the n spatial hand-object configurations are successfully represented by separate cells in the output layer.

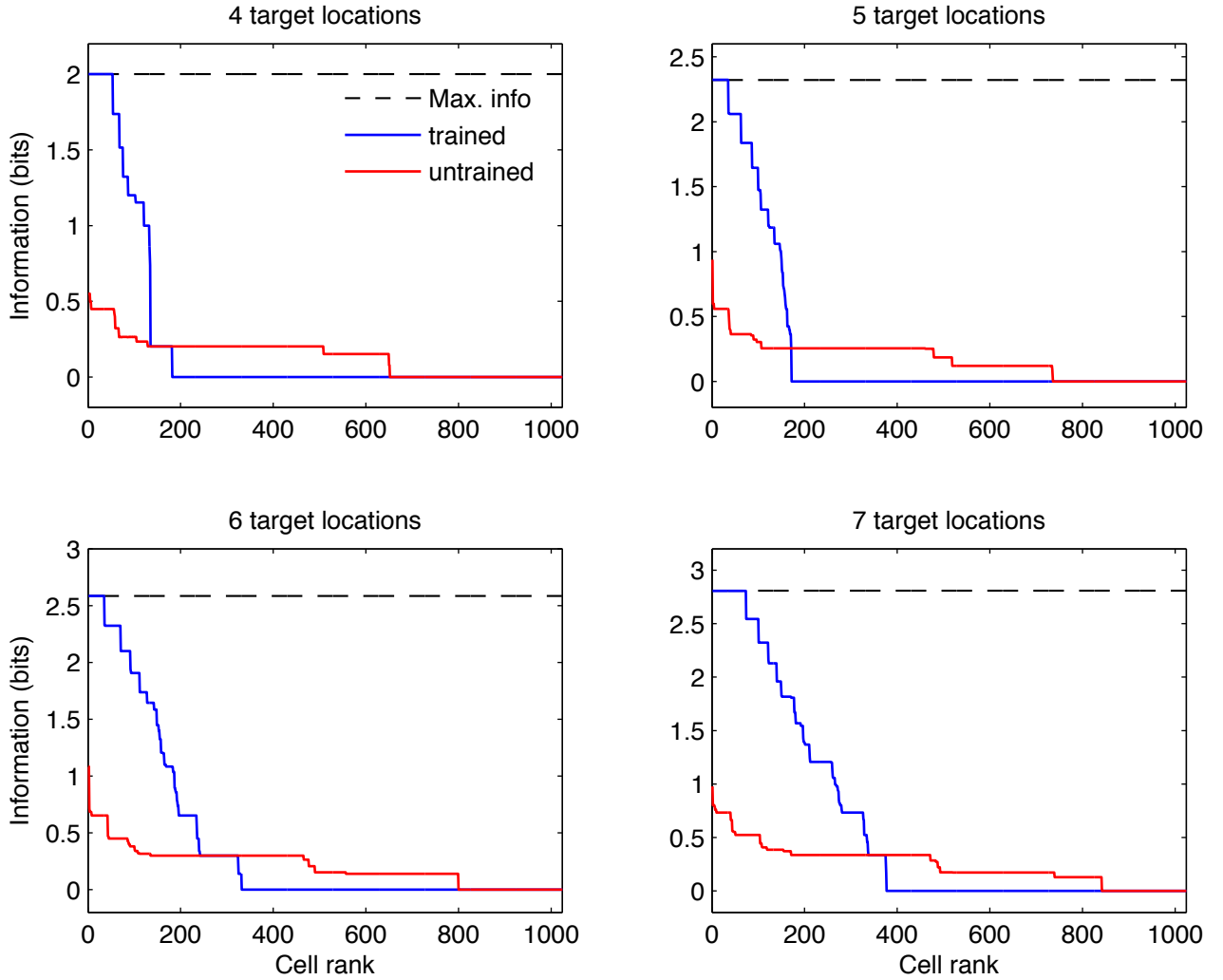


Figure 5.3: Single cell information analysis. Analysis of the information conveyed by the output (fourth) layer neurons about the location of the visual target with respect to the hand. The single cell information analysis (top plot) shows that in all cases more than 10% of the neurons conveyed the maximal single cell information.

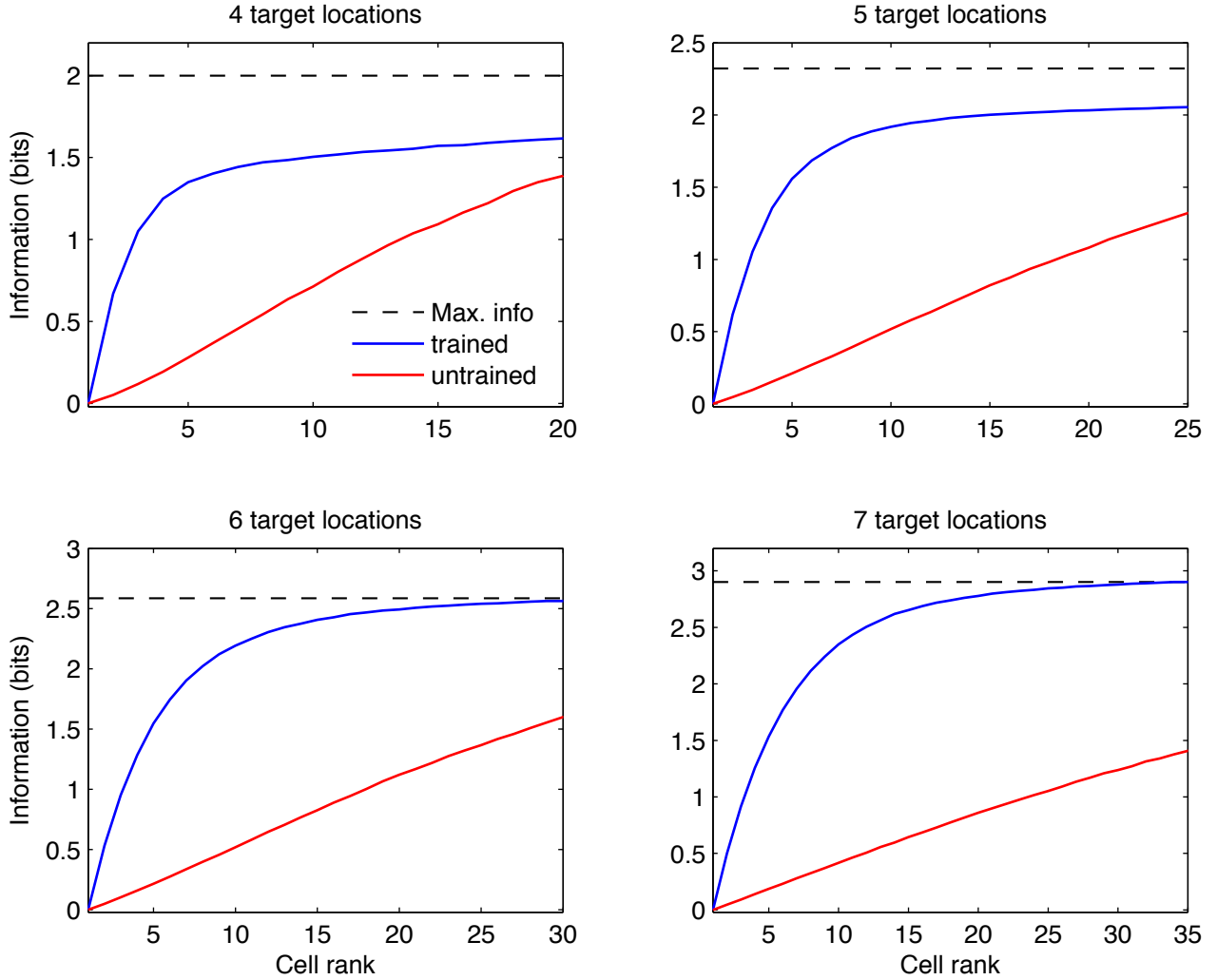


Figure 5.4: Multiple cell information analysis. The plot shows the multiple cell information measures calculated across 5 cells with maximal single cell information per each of the target locations (e.g. 4 target locations = 20 cells sampled). The multiple cell information measures show that only when the number of target locations is equal to 6 or more, does the multiple cell information asymptote to the corresponding maximal value after training, meaning that only in these cases all of the individual hand-object configurations are represented by the output cells. In the cases of less than 6 target locations, the networks fail to reach maximal information.

5.4 Discussion

The results show that VisNet is able to develop neurons in the output layer that respond exclusively to single, localised hand-centred locations regardless of the retinal location, even when the network was always presented with objects seen together in multiple hand-centred locations during training. Trained output cells with the maximal single cell information were responding exclusively to single, localised hand-centred locations across all of the retinal locations and not to any other spatial configuration. This shows that trace learning allows the network to develop localised hand-centred receptive fields through statistical decoupling between different hand-centred locations even when learning occurs in the presence of multiple objects around the hand.

An individual hand-object configuration comprises a particular subset of visual features that are highly correlated with each other across different visual scenes. A neuron that starts to represent a particular pair of hand-object configurations, will subsequently start to get decoupled due to the statistics of the training environment, where the features of an individual hand-object configuration occur together more frequently than with the features of any of the other possible hand-object configurations. Thus, the neuron learns to represent the features of just one hand-object configuration while other cells represent different hand-centred locations. This is a remarkable feature of VisNet, which is able to build representations of individual hand-centred locations while at the same time building transform invariant representations over different retinal locations.

The multiple cell information asymptotes to the maximal value only for the test cases with $n \geq 6$. This finding suggests, that the effects of the statistical decoupling improve as the total number of target locations increases. This result is consistent with previous simulations exploring the effects of statistical decoupling on the formation of separate object representations (Stringer et al., 2007).

Regardless of the progress achieved by training the network with two objects at a time, this training procedure still makes use of very abstract visual stimuli. The stimuli are still comprised of a pair of circular objects and a featureless hand. Meanwhile, the real world-environment is composed of a variety of natural objects with rich features. These objects are seen in different settings and backgrounds. Nonetheless, evidence from single cell recordings in primates suggests that cells that are still able to develop single, localised hand-centred receptive fields (see Chapter 1 on hand-centred responses or Inseparable T+H). The following section extends the application of the principle of statistical decoupling to more realistic visual training conditions, including images of a realistic hand and natural objects, which are present against more natural textured backgrounds during training.

5.5 Training the network with more realistic natural objects presented around the hand with more natural textured backgrounds

Introduction

In real-world visual scenes the various objects that we encounter throughout our sensory-motor experiences have different shapes and sizes. Nevertheless, cells in the dorsal stream seem to be able to generalize and form delineated visual receptive fields in different reference frames. Hand-centred cells for example, would seem to fire maximally to an object, independently of its attributes, as long as it appears in a particular position with respect to the hand. These hand-centred responses are also independent or 'invariant' to the direction of the gaze as well as the position of the arm with respect to the body.

In the previous section we took a step towards improving the ecological plausibility of our training regime by presenting the network with pairs of objects simultaneously. However, the objects were idealised circles. In this second section of the chapter, we explore whether VisNet would still be able to develop output cells with single, localised, hand-centred receptive fields when the network is exposed to more realistic images comprised of natural objects presented around a more realistic hand with a range of natural textured backgrounds.

More specifically, this was carried out by incorporating images of a realistic hand and a variety of everyday objects presented in pairs simultaneously around the hand. Instead of displaying the images on a grey background, the images of the objects and the hand could now appear on more natural textured backgrounds during training. We applied here the same principles of statistical decoupling that allowed the network to break the associations of the pairs of objects into separate representations of the individual hand-centred object locations. Therefore, even when the network is trained with multiple objects presented simultaneously in different backgrounds, it learns to represent the spatial configuration of the hand and a particular object location.

This is possible due to the fact that the network learns to bind the features that are seen more frequently together than any other combination of features in the environment. Thus, the network does not have any knowledge of which features belong to a particular object, it self-organizes by learning to respond to the features that co-occur the most. Therefore our hypothesis was that after extensive training with a variety of natural objects, the network would develop visual receptive fields that would encode a combination of features of the hand and a variety of features that appear in a certain region near the hand. This would allow the network to generalize and respond to the presence of a novel object, as long as it is presented within the visual hand-centred receptive field. Additionally, we used trace learning to produce a temporal binding of the different retinal locations of the hand-object configurations. This allows the cells in the output layer of VisNet to develop responses that

are tuned to a particular position of an object with respect to the hand, independently of where on the retina this configuration appears.

Using natural backgrounds in VisNet.

The purpose of these simulations is to increase even further the ecological plausibility of the training procedures by making them more similar to natural visual scenes. Therefore, in addition to the use of natural objects during training, we were also interested in presenting them against realistic textured backgrounds, instead of the plain backgrounds that we had been using before. Previous research with VisNet (Stringer and Rolls, 2000) has revealed the difficulty for the network to build invariant representations when it is trained on cluttered backgrounds. In the aforementioned study, the network successfully learned to recognize faces that were trained on a plain background and later tested on a cluttered natural background. However, when the network was trained with a natural background and then tested on a blank or plain background, the cells did not seem to be able to respond selectively to the faces anymore.

The two backgrounds selected for those simulations were static images composed of foliage. Each of these backgrounds remained static while the faces were presented in different parts of the background. During training, the cells were exposed not only to the features of the faces but in this case they were also learning to respond to features of the background. The features of the background were stable and salient enough that cells in the output layer ended up responding mostly to the background rather than the face

stimuli (Rolls, 2012b; Stringer and Rolls, 2000).

The authors of this study suggested that the failure of the network to learn about the faces in these cluttered backgrounds might be related to the fact that the network was always trained on the same two static backgrounds. Unlike these previous experiments, we make use of a variety of textured backgrounds, and even when the background remains static with respect to the hand and objects, the visual exploration of the network across the visual scene presents different views of the background features. This effectively presents the features of the background in different parts of the model's 'retina'. The variety of backgrounds in which the hand and objects are seen facilitates the extraction of the most relevant and concurrent features.

Thus, instead of having a background providing unwanted noise, the network actually benefits from the variety and richness of its training regime, which allows it to successfully extract the relevant features that will be concurrent across different settings and backgrounds in different retinal views. The network then is not learning to respond to particular features of the background, but to the salient features of the hand and the features of objects around the hand. As it is shown in the results of this study, VisNet is now able to respond to individual hand-centred configurations after being trained in natural backgrounds, even when these configurations are presented during training in a novel plain background.

5.6 Methods

5.6.1 Stimuli

The training stimuli for this experiment consisted of images showing a hand and two natural objects in different spatial configurations surrounding the hand. Unlike the previous simulations, the training images consisted of realistic images of the hand and natural objects presented on different textured backgrounds.

The images of the hand, objects and backgrounds were selected from open source pictures on the internet. The templates were designed, scaled and arranged using Adobe Photoshop software. The images were generated in RGB color and subsequently converted to monochrome using the MATLAB function `rgb2gray`. Figure 5.5 shows a sample of some of the training images of hand-object configurations that were generated for this study.

The backgrounds of the images were extended to 512×512 pixels for the preprocessing stage. The filtered outputs were then cropped back to the original size 256×256 . This step is important to avoid possible artefacts or edge effects from the filters in the initial layer of the network.

The centres of all the objects were distributed along a semicircle in six different possible locations around the hand. Similar to the previous simulations the images showed all possible pairings of the six object locations. Each image would contain two different natural objects drawn randomly from a pool of 42 realistic objects. The number of possible pairs of objects locations may be calculated by Equation 5.1, where $n = 6$ and $r = 2$, which

gives a total of 15 pairings of object locations.

For each such pair of object locations we randomly selected two different objects from the pool of 42 realistic objects to be presented in that pair of locations. However, each such pair of objects was presented in both possible arrangements: i.e. object 1 in location 1 and object 2 in location 2, and then object 1 in location 2 and object 2 in location 1. This led to a total of 30 hand-object configurations. Then each of the blocks of these 30 hand-object configurations was presented against one of the 21 different textured backgrounds. A total of 630 image sequences were generated, each sequence presented a pair of natural objects near the hand on a textured background. In order to present the configurations of the hand and objects in different retinal positions, each of the paired spatial configurations was translated by six pixels at a time across VisNet’s ‘retina’.

During training, the hand was always shown surrounded by a pair of objects and never with a single object in isolation. After training was completed, the stimuli used during testing consisted of images showing the hand and a novel object in a specific position relative to the hand. In the test image, the novel objects were shown in one of the same six hand-centred object locations that were used during training.

5.6.2 Model parameters

In this experiment we continued using an up-scaled version of the model ‘retina’ (i.e. 256×256). The values of the network dimensions used for these simulations are in Table

5.2. The rest of the parameters remained unchanged.

	Dimensions	Number of Connections	Radius
Layer 4	32x32	100	12
Layer 3	32x32	100	9
Layer 2	32x32	100	12
Layer 1	64x64	100	12
Retina	256x256x16	-	-

Table 5.2: Network dimensions Network dimensions showing the number of connections per neuron and the radius in the preceding layer.

5.6.3 Training

The training procedure for this experiment consisted of presenting VisNet with pairs of non-overlapping natural objects displayed around the hand on a textured background. During training, the objects were presented in pairs and never in isolation, see Figure 5.1.

In order to develop invariant responses across different retinal views, each of the images representing a particular configuration of a hand and objects is trained across five different retinal locations. Following the same protocol as in previous simulations, the image sequences are meant to arise from a series of eye movements and the resulting shifts in the position of the hand and visual objects on the 256×256 ‘retina’. During each of the image sequences, the fixed spatial configuration of the hand and pair of objects is translated six pixels at a time. During the visual exploration of a particular spatial configuration the natural background is always the same. A new background is only used when a new

configuration of a hand and objects is presented.

During the presentation of every frame of the image sequence the activation of individual neurons within the layer and their firing rates are calculated. Subsequently, the synaptic weights between layers are updated according to the trace rule used in the previous chapter, see Equation 4.2. This process is then repeated for every layer of the network. Similar to all the simulations presented heretofore, the network is trained one layer at the time, starting with layer 1 and finishing in layer 4.

One training epoch consisted of the presentation of every possible object pairing, each across the five different retinal locations. Figure 5.5 shows examples of six training images, each composed of a pair of hand with two natural objects presented simultaneously to the network. In these simulations the network was trained for fifty epochs per layer. The learning rates used were 0.1, 0.1, 0.1 and 0.1 in each layer.

5.6.4 Testing

Throughout the testing phase, the synaptic weights were not changed. In order to test whether VisNet has been able to develop translation invariant neurons with a single, localised, hand-centred receptive field, image sequences of the hand and only one circular object are presented in five different retinal locations. However because the goal is to test whether neurons respond to a specific hand-centred location irrespective of the object form, the test images used a simple textured object as shown in Figure 5.6

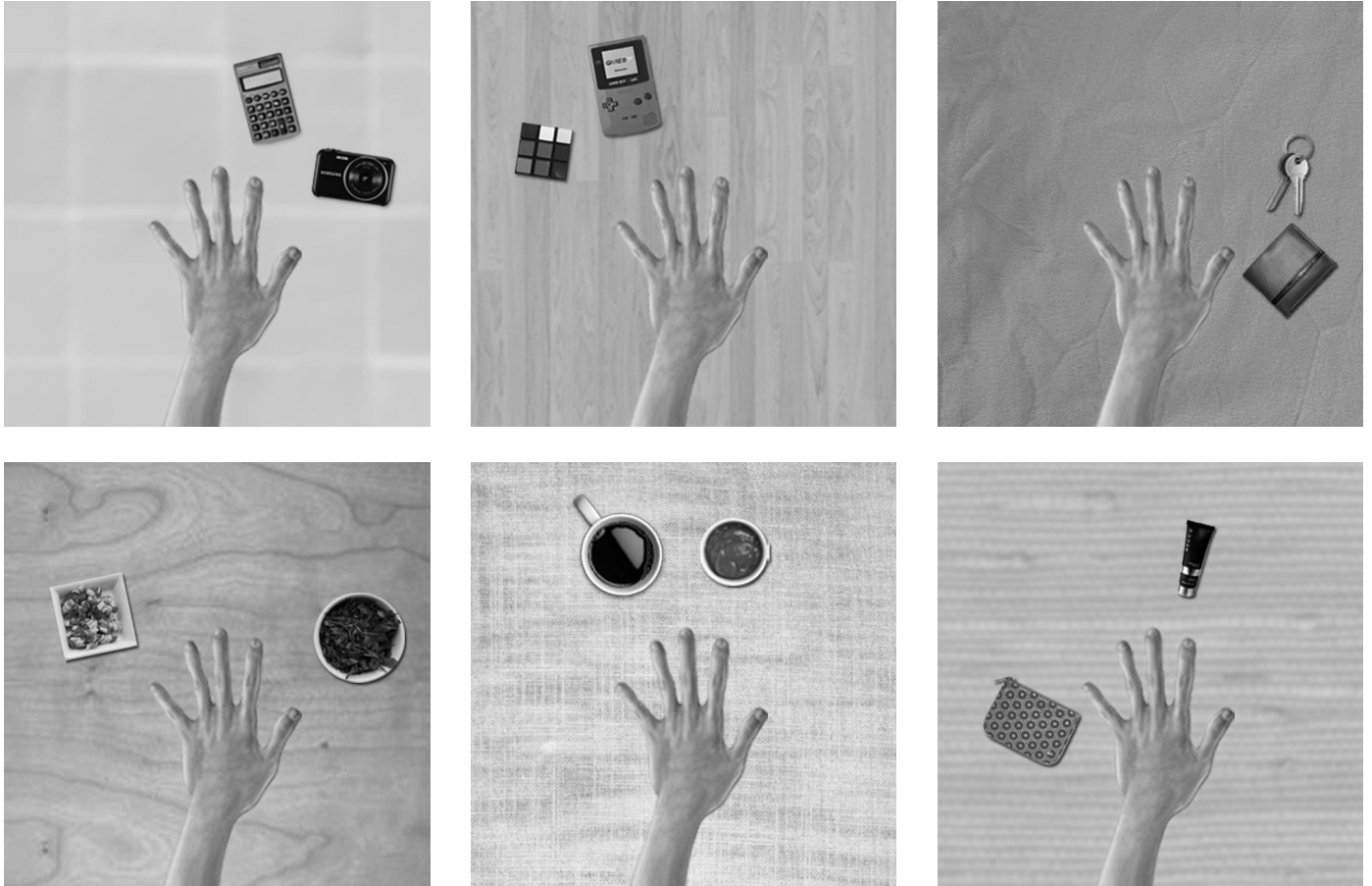


Figure 5.5: Realistic training images with natural objects and textured backgrounds. These are examples of some of the training images used in the simulations. A pair of objects would appear in different hand-centred locations simultaneously. The eyes would move exploring the visual scene producing different views of the same configuration across different retinal locations. The six figures represent a sample of the possible images of object pairings generated from the pool of natural objects and textured backgrounds. The relative positions of the hand and the pair of objects are unchanged during the eye movements.

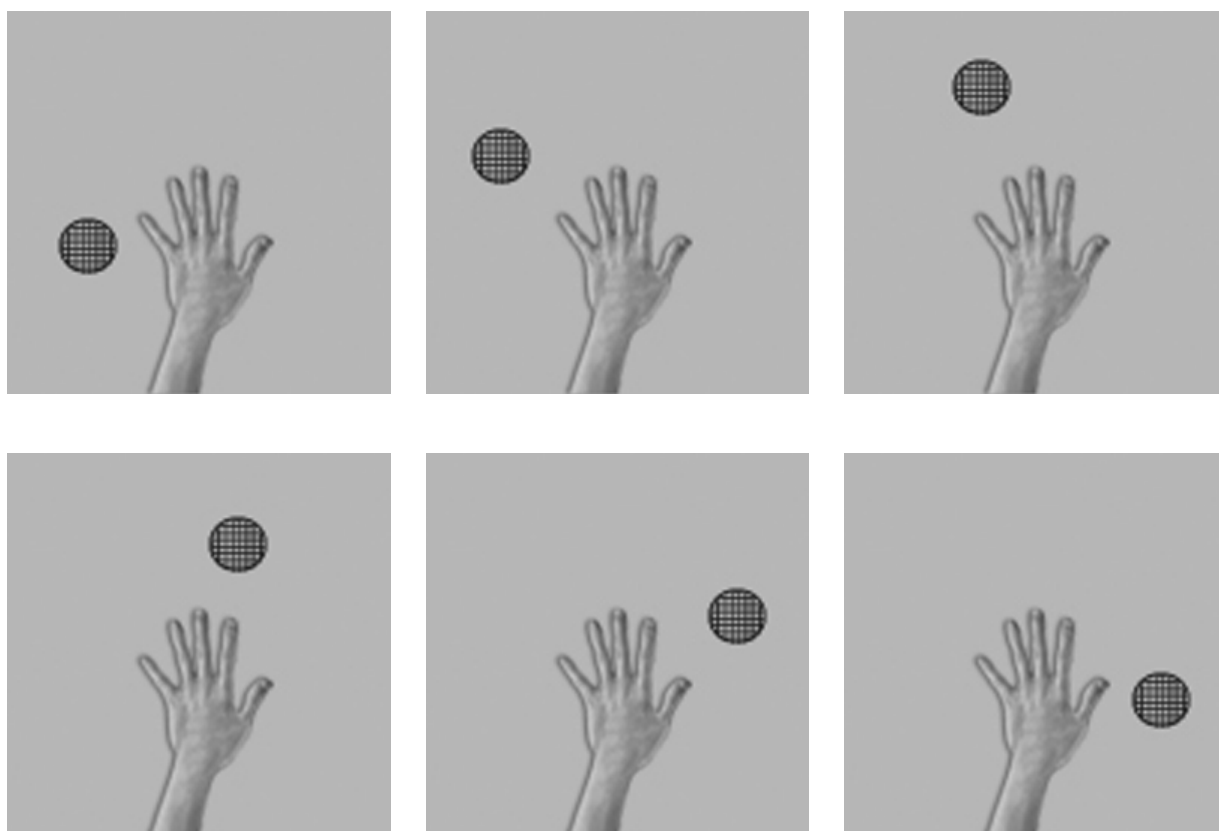


Figure 5.6: Testing with a plain background and one textured stimulus. These are examples of six testing images used in this simulation study. Unlike the training condition where two objects were presented simultaneously, in this case a single textured stimulus was presented in six different hand-centred locations. The hand-centred object locations were the same as those used during training. Each hand-object configuration would be tested in five different retinal locations.

The outputs of the network are computed during the presentation of each hand-object configuration in each retinal location. Figure 5.6 shows an example of the six testing images presented to the network. Each of these images were tested in five different retinal locations. Lastly, a recent addition to the inspection tools of VisNet enables the user to select an output cell after training and then trace back the connections through layers that have been strengthened by learning. This process can be repeated up until the point that we reach the bank of Gabor filters in the input layer. This permits us to identify which visual features of the input images the selected output cell is responding to most strongly.

5.7 Results

We studied the responses of the output (fourth) layer cells in VisNet before and after the network was trained on the images of hand-objects configurations. After the network was trained, the network was tested to determine whether cells in the output layer had developed single, localized hand-centred receptive fields and responded invariantly across the different retinal locations. Information analysis was then conducted on the responses of the cells to all of the test images.

In previous simulations in which VisNet was trained on all possible pairings of objects, it was reported that as the number of objects increased $n \geq 6$, the statistical decoupling between the objects started to force the network to learn to represent the objects individually (Stringer et al., 2007). However, in the simulations carried out in this chapter, the

image of the hand is always present together with images of objects. In this case, the most correlated features will correspond to a combination of features of the hand and features of the trained objects presented in a particular location with respect to the hand. Therefore, individual cells will learn to respond to a particular spatial configuration of the hand and a hand-centred object location.

Fig. 5.7 shows the response profiles of six neurons in the output layer of VisNet before training. Following the same conventions of previous chapters, each of the six columns of plots contains the firing responses of a particular output cell, these are labelled at the top of the column. Whereas the six rows of plots show the responses of the cells according to each of the six hand-object configurations presented during testing.

Each plot shows the responses of the given cell to the particular hand-object configurations over five retinal locations. The x axis in the individual plots represents the five retinal locations of the hand-object configuration on which the neuron was tested, while the y axis represents the corresponding firing rate of the output neuron.

The top row shows the cell responses when a single textured object is presented in the first of the testing locations with respect to the hand. This corresponds to the upper left image in Figure 5.6. The following rows show the cell responses when the visual object is presented in successive test locations with respect to the hand. The last row corresponds to the configuration displayed in the bottom right image of Figure 5.6.

In Figure 5.7 we can see that before training, all of the six cells respond rarely and ran-

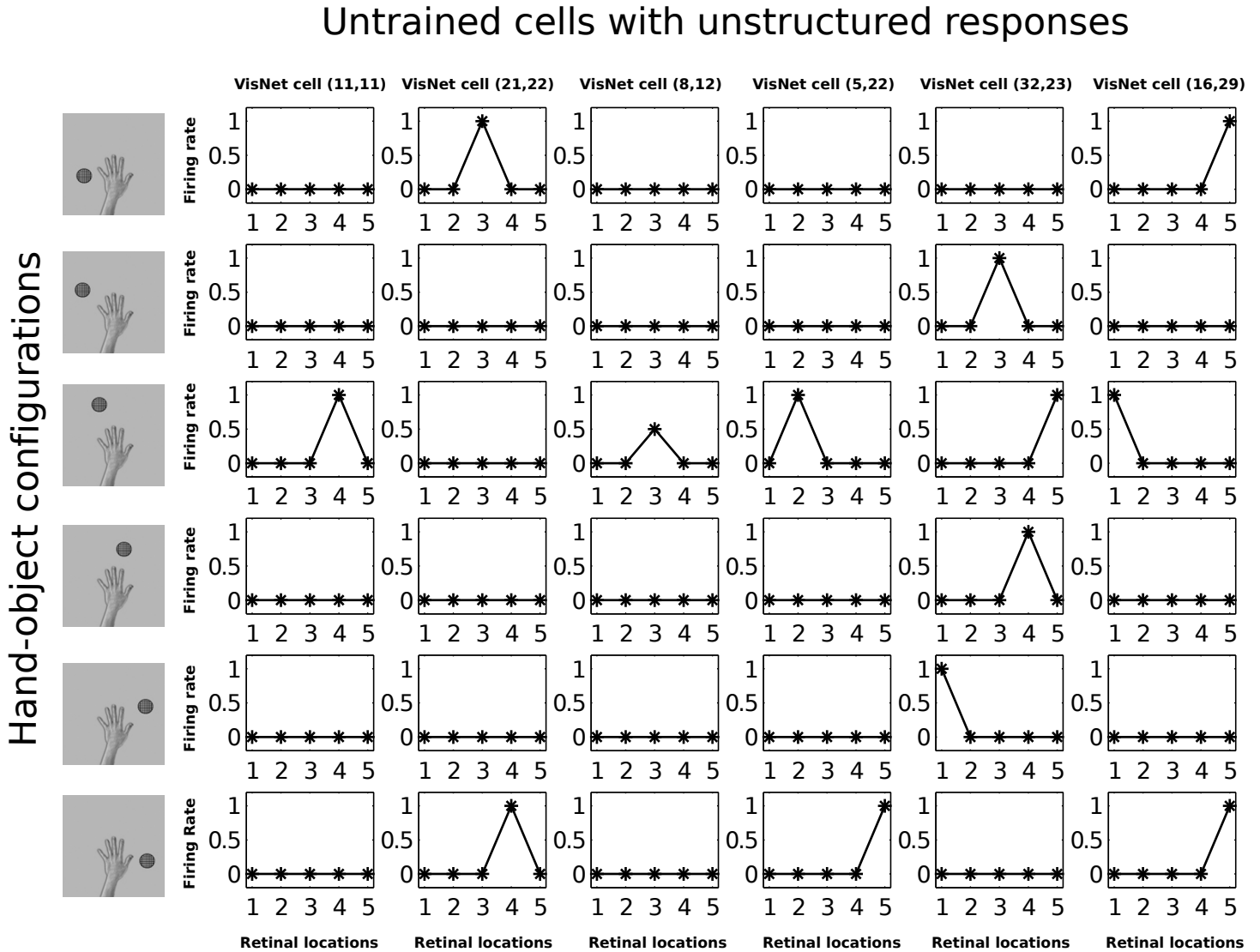


Figure 5.7: Response profiles of six neurons in the top (fourth) layer of VisNet before training. Each of the six columns shows the firing responses of a particular cell. Each row shows the responses of the cells to one of the six hand-object configurations over all five different retinal locations shown along the abscissae. It can be seen that each of the six cells initially responds randomly to each of the hand-object configurations over the different retinal locations.

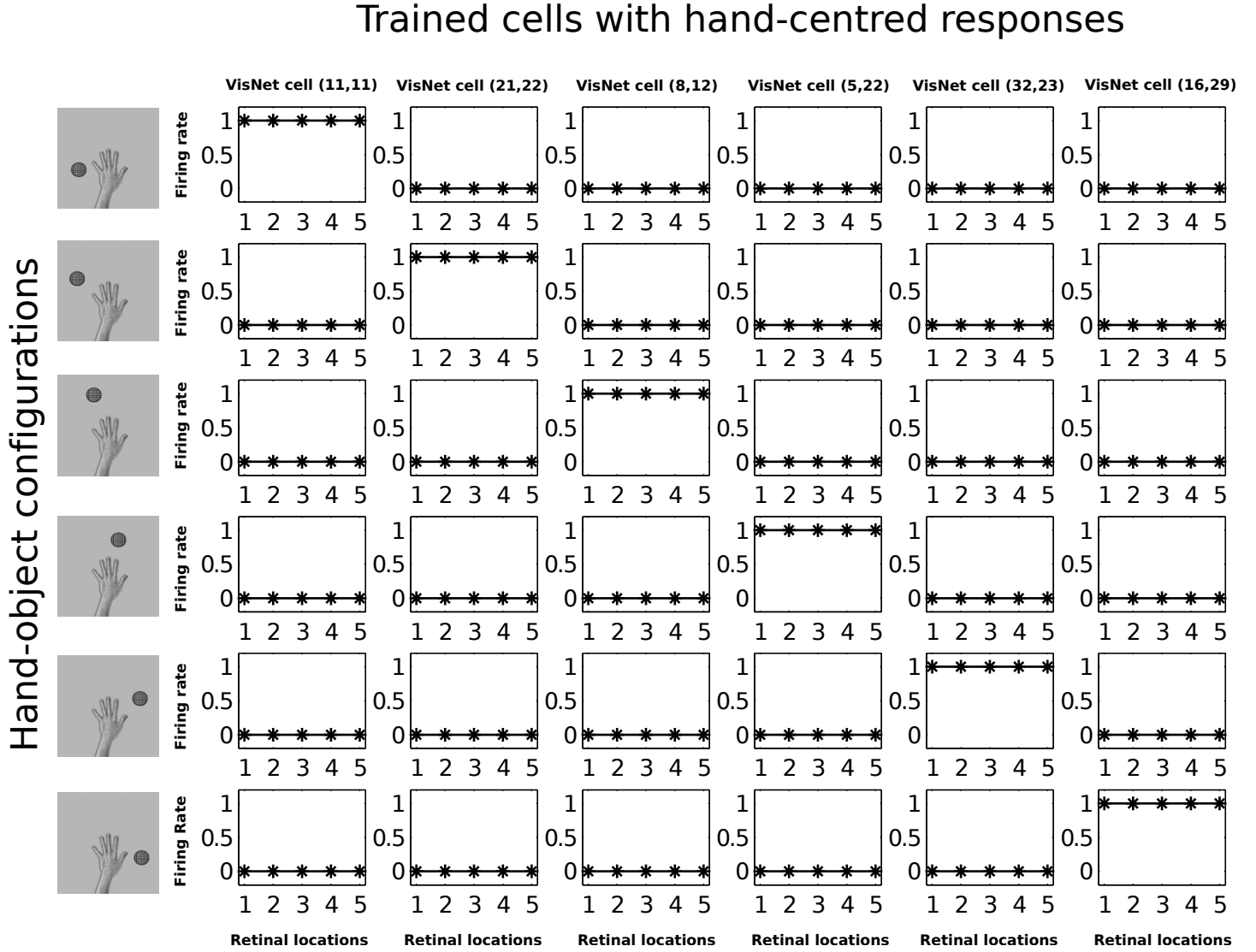


Figure 5.8: Response profiles of same six neurons in the top (fourth) layer of VisNet after training. Response profiles of the same six neurons from Fig. 5.7 after training on the images shown in Fig. 5.6. Conventions as in Fig. 5.7. It can be seen that each of the six cells responds selectively to just one of the hand-object configurations, and responds to that configuration over all five retinal positions shown along the abscissae. Moreover, each of the six hand-object configurations is represented by one of the cells.

domly to the different hand-object configurations. The responses do not have a particular ordered structure. In Figure 5.8 we can see the response profiles of the same six neurons in the output layer of VisNet after training. In this case it can be seen that, after training, each of the six cells has learned to respond to just one of the hand-object configurations, and responds to that configuration over all five tested retinal locations. Furthermore, we can see here already that each of the six hand-object configurations is represented by one of the cells.

In order to have an overview of how these configurations are represented in the output cell population, we present the information analysis measures. Fig. 5.9 shows the single and multiple information measures for the output (fourth) layer neurons before and after training with all of the hand-object configurations.

The single cell information analysis, (Fig. 5.9 top) shows that, after training, 115 neurons conveyed the maximal single cell information of 2.58 bits. This means that there were 115 output cells that responded to only a single position of the test object with respect to the hand, and those cells responded irrespective of retinal location. In the response profiles after training shown in Figure 5.8, I showed examples of neurons representing each of the six hand-object configurations. The multiple cell analysis would similarly confirm that all six hand-object configurations are represented across all retinal views by separate output neurons.

The multiple cell information analysis (Fig. 5.9 bottom) shows that, before training,

the multiple cell information does not reach the maximal value of 2.58 bits. However, after training we can see that multiple cell information asymptotes to the maximal value, which means that all of the possible individual hand-object configurations are successfully represented by separate cells in the output layer.

We traced the strengthened connections from each one of the output cells through successive layers to the input Gabor filter driving that cell. Figure 5.10 shows the Gabor input filter with strengthened connections to a trained output neuron that had learned to respond to one of the hand-centred locations.

On the left side of Fig. 5.10 we can see the Gabor filters that are most strongly driving the responses of the particular output cell. In this example, we show a cell that is representing a subset of Gabor filtered inputs corresponding to the hand, as well as a subset of inputs representing a visual location near the hand. Tracing back the synaptic connectivity in this way, allows us a visual inspection of the nature and extension of the hand-centred visual receptive field developed by the output cell after training. We can thus determine not only the ability of the cell to represent an individual region with respect to the hand, but also the input features that were extracted from the database of objects shown.

Altogether, the individual cell firing rate responses, the information analysis and the inspection of connectivity in this study demonstrate that VisNet is able to develop neurons with a single, localised, hand-centred visual receptive field even when trained on more

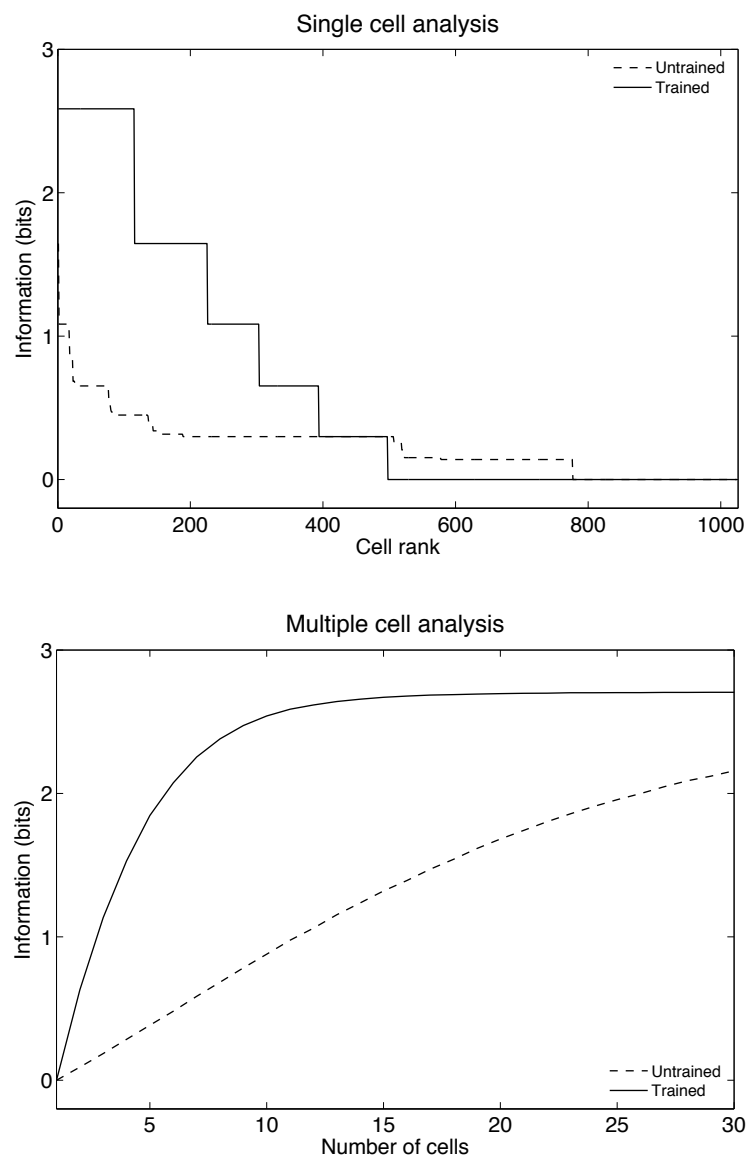


Figure 5.9: Information analysis. The upper plot shows the amount of single cell information carried by individual output cells in rank order. It can be seen that after training, it was found that 115 cells reached the maximum amount of single cell information of 2.58 bits. In the untrained condition no cells reached maximal information. These cells responded perfectly to just one of the six tested hand-object configurations, and responded to that configuration across all five different retinal locations. The lower plot shows the multiple cell information measures calculated across 30 cells with maximal single cell information. It can be seen that, after training, the multiple cell information asymptotes to the maximal value of 2.58 bits. This confirms that all six tested hand-object configurations are represented by the output cells.

realistic images with multiple natural objects shown with the hand against various textured backgrounds.

In particular, the principles of statistical decoupling continue to operate successfully under these more ecological training conditions. After extensive training, the output cells learn to respond to the features that are seen more frequently together throughout training. This is a basic property of competitive learning.

Since the network is trained on multiple natural objects with the hand against various textured backgrounds, the features that appear more frequently together are the hand (which is always present) and a subset of features that are associated with a particular object location. Consequently, individual output neurons learn to represent a particular configuration of the hand and one object location with separate neurons responding to different hand-centred object locations. However, the statistical decoupling between any two object locations is too weak to permit individual output cells to learn to respond to more than one hand-centred location. Additionally, the trace learning mechanism enables the network to encode these representations across different retinal locations. Since the network we have described so far is guided by visual training, these cells will respond to the same (inseparable T+H) configuration irrespective of the actual hand location with respect to the body and regardless of the gaze direction. These hand-centred cells will fire maximally as long as the spatial configuration of the hand and an object is the same.

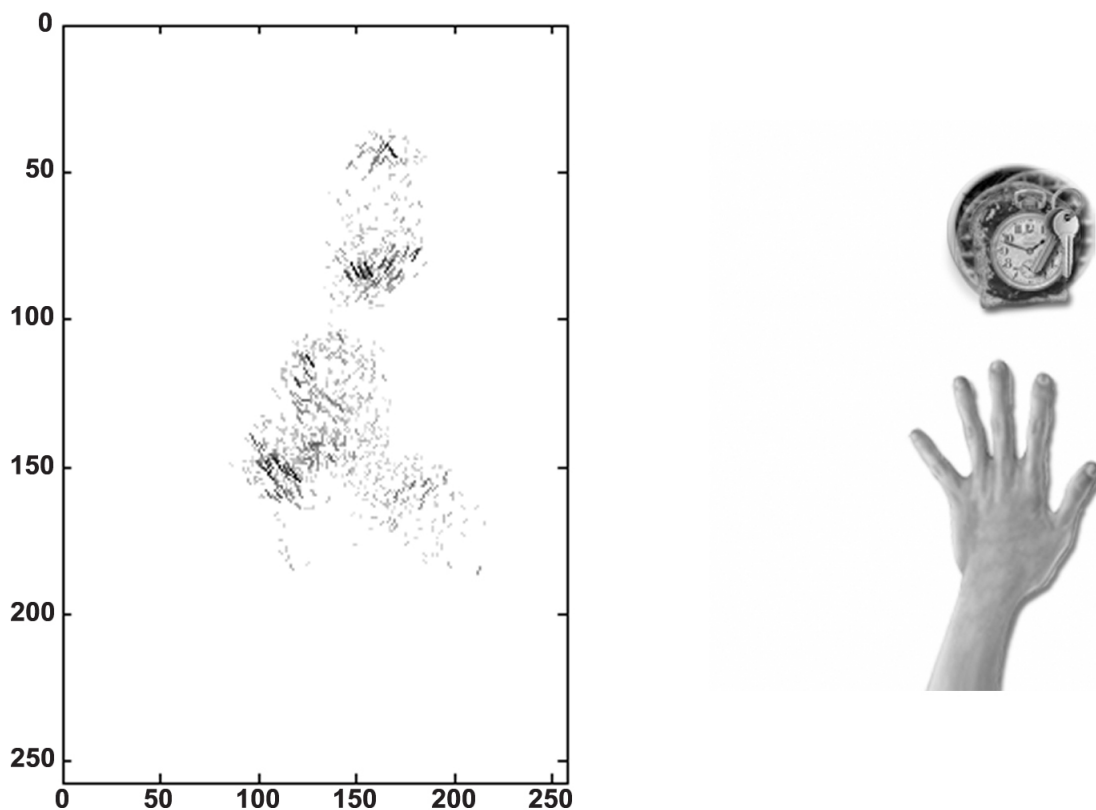


Figure 5.10: Tracing back the synaptic connections from a trained output cell to the input Gabor filters. The left side shows the input Gabor filters that an output cell has learned to respond to after training. This is an example of a neuron that represents a hand-object configuration with the object above the hand. In this image the Gabor filters with the strongest connectivity through the layer to the output cell are plotted, where each Gabor filter is weighted by the strengths of the feed-forward connections from that filter through the successive layers to the output neuron. It can be seen that this neuron receives the strongest inputs from a subset of Gabor filters that represent the location of the target on top of the hand. The right side shows the image of the hand and the overlapped images of all the training objects that appeared during training in this hand-centred location.

5.8 Discussion

The simulations presented in the previous chapters provided a first examination of the self-organizing principles that could underlie the visual development of hand-centred representations. However, the visual stimuli and training conditions were highly unrealistic, consisting of presenting the network with a single circular object at a time. Therefore, the main goal of this chapter was to explore whether VisNet was still able to develop output neurons with single, localised, hand-centred receptive fields when faced with more ecological and realistic visual training conditions.

The first section of this chapter explored when the training procedure involves the presence of more than one object at a time. Previous research (Stringer and Rolls, 2008; Stringer et al., 2007) showed that it was possible for competitive networks to build separate representations of individual objects even when the training environment was composed of multiple objects presented simultaneously and never in isolation. This is possible as long as the features that we want to bind are seen more frequently together than any other combination of features in the environment.

Stringer et al. (2007) presented all possible pairs of objects to the network during training. Consequently, it was found that as the number of objects increased so too did the statistical decoupling between any two objects. This, in turn, forced the output neurons to switch from representing the object pairs, on which the network was trained, to representing the individual objects. It was further found that this effect of statistical decoupling could be

combined with trace learning to produce translation invariant representations of individual objects.

In modelling the visually guided development of hand-centred visual receptive fields, the goal is not to generate a representation of the hand and separate representations of each hand-centred object location, but to generate representations of the hand and an object together in various spatial configurations. The network has to self-organize exploiting the statistics of the input patterns, binding together the features that are seen together more frequently. In the simulations presented in this chapter, the patterns of input features that are most frequently seen together each consist of the hand and one of the hand-centred object locations. The network therefore learns to represent these input patterns. However, any two hand-centred locations are rarely activated together, and so the output neurons do not combine these onto the same output cells. Thus, each output neuron learns to respond to the combined presence of the hand and a single hand-centred object location.

Therefore, with a constant presence of the hand during training the statistical decoupling produced cells that represent particular hand-object configurations with a single, localised hand-centred receptive field. Neurons that begin to represent a combination of a hand and a pair of object locations get decoupled into separate hand-object configurations due to the statistics of the training environment, where the features of an individual hand-object configuration occur together much more frequently than the features comprising the pair of objects with the hand.

The results of the first section of this chapter showed that VisNet is able to develop output neurons with single, localised, hand-centred receptive fields even when the training set presents the hand with pairs of hand-centred targets, and no configurations of the hand with a single object were presented in isolation. Therefore, this meant that the principles of statistical decoupling could be extrapolated to our training paradigm. This represented a significant advance towards having greater ecological plausibility in our training procedures. However, the stimuli we were using were still highly abstract, consisting only of a hand with circular objects on a plain untextured background.

In the real world we encounter multiple objects with rich and diverse features. Therefore, in the second part of the study we improved the realism of the training stimuli by using realistic images of the hand and images of natural objects surrounding the hand, and presented these configurations against a range of more natural textured backgrounds.

Previous research with VisNet (Stringer and Rolls, 2000) failed when the network was trained on natural backgrounds and then tested on either the same or plain backgrounds. In these experiments the stimuli were presented against the same static background during training. In this case the neurons simply learned to respond to the background features.

Considering that in natural visual experience we encounter objects in a variety of settings, we trained the network as well with a range of different textured backgrounds. This way, the most salient input features that would be seen more frequently across all the training sequences will be input features of the hand and the natural objects presented in

particular locations near the hand. The features of the changing backgrounds were not learned by the network. The results obtained in this section by testing the network on a plain background and tracing the synaptic connectivity from the output cells to the input Gabor filters, shows that the network is able to extract the individual hand-object configurations and represent them with location invariance. Thus, instead of providing unwanted noise, the network can actually benefit from the variety and complexity of its background environment, which allows the network to extract the key features relating to the hand and surrounding object locations that will be concurrent across different visual backgrounds.

5.9 Assessing the selectivity of neurons for object shape within their localised hand-centred receptive fields

For the aforementioned experiment, we were not interested in developing hand-centred cells that were selective to specific objects. On the contrary, we were primarily interested in the development of hand-centred receptive fields where the neuron would respond to the presence of almost any object as long as it is presented within the receptive field. These cells are thought to mostly provide information about the location of an object with respect to the hand, rather than representing the detailed features of the object. However, our simulations do not preclude the possibility that some shape selectivity could arise after training.

The dominant theories of the functional roles of the ventral and the dorsal streams

seem to agree on the involvement of the ventral stream predominantly in the perception of object form and object recognition (Goodale and Milner, 1992; Ungerleider and Haxby, 1994; Ungerleider and Mishkin, 1982). There is substantial evidence for the neuronal selectivity of different parts of the ventral stream to represent features of objects such as shape and colour (Kellenbach et al., 2001). There is less consensus, however, on the functional role of the dorsal stream. Lesion and imaging studies seem to suggest that the dorsal stream plays an important role in the localization of objects in space (the “where” pathway, Ungerleider and Haxby, 1994; Ungerleider and Mishkin, 1982), as well as visual planning and control of movements (i.e. vision for action Goodale and Milner, 1992).

However, it has been suggested that there are areas within the dorsal stream that play important roles not only in localisation of objects and action planning but also in perception and understanding of our actions (Rizzolatti and Matelli, 2003). There is evidence of activation in the dorsal stream for the retrieval of certain visual attributes of objects (Oliver and Thompson-Schill, 2003). Regions of the dorsal stream, like cells in the AIP for example, seem to actually differentiate and recognize specific features of objects, since these features are informative when we need to adjust the shape of our hand to prepare a grasp (Murata et al., 2000). Cells in VIP also seem to be selective for stimuli that could represent a threat (Graziano and Cooke, 2006). Other high level dorsal regions such as LIP have been reported to show selectivity to simple geometrical shapes (Serenio and Maunsell, 1998). This study, however, did not distinguish or analyse the frame of reference of the

recorded cells. It has been reported that the representation of space in LIP cells ranges from eye to head centred cells (O'dhanziel et al., 2005), including cells with eye-position gain fields (Andersen et al., 1990). Additionally, both LIP and AIP receive strong connections from inferior temporal areas TE and TEO (Webster et al., 1994). Therefore, LIP and AIP could also be receiving information about the visual attributes of the stimuli from areas within the ventral stream.

Nevertheless when we make a comparison at a single-cell neuron level between high-level ventral regions that are shape selective, such as the anterior inferotemporal cortex (AIT) and high level dorsal regions that have been also been reported as shape selective (e.g. LIP), it has been found that AIT neurons on average had higher shape selectivity than those of LIP (Lehky and Sereno, 2007). AIT neurons also had significantly more units that were highly selective to shape, whereas LIP had very few neurons that were highly selective to shape.

In this subsection we explored whether the output cells of our model developed hand-centred receptive fields that were selective to shape, as well as evaluating whether this selectivity is affected when the network is trained with more objects. We found that, similarly to the findings reported in high-level dorsal regions such as LIP, only a few neurons were selective to object shape, while the majority of cells would fire maximally to a variety of stimuli that appeared within the localised hand-centred visual receptive field. In the discussion we provide an hypothesis on the learning mechanisms that account for

this difference in the proportion of shape selective cells in the ventral and dorsal visual pathways.

5.9.1 Method

5.9.2 Stimuli

The training stimuli for this experiment consisted of images showing a hand and a natural object selected randomly from a pool of images of natural objects. The images of the objects were selected from open source pictures on the internet and we used the same image of the hand shown in the previous section. The templates were designed, scaled and arranged using Adobe Photoshop software. The images were generated in RGB color and subsequently converted to monochrome using the MATLAB function `rgb2gray`.

The testing images consisted of 100 images showing the hand and novel objects in the same hand-centred position in which the network was trained. The objects used during training and testing were not the same. In Figure 5.11 we show examples of the pool of objects used for training and testing.

5.9.3 Training and testing

For these experiments we were interested in testing whether the output cells that developed visual hand-centred receptive fields could also show shape selectivity. Therefore, we started by training the network with an image of the hand with a single natural object in a

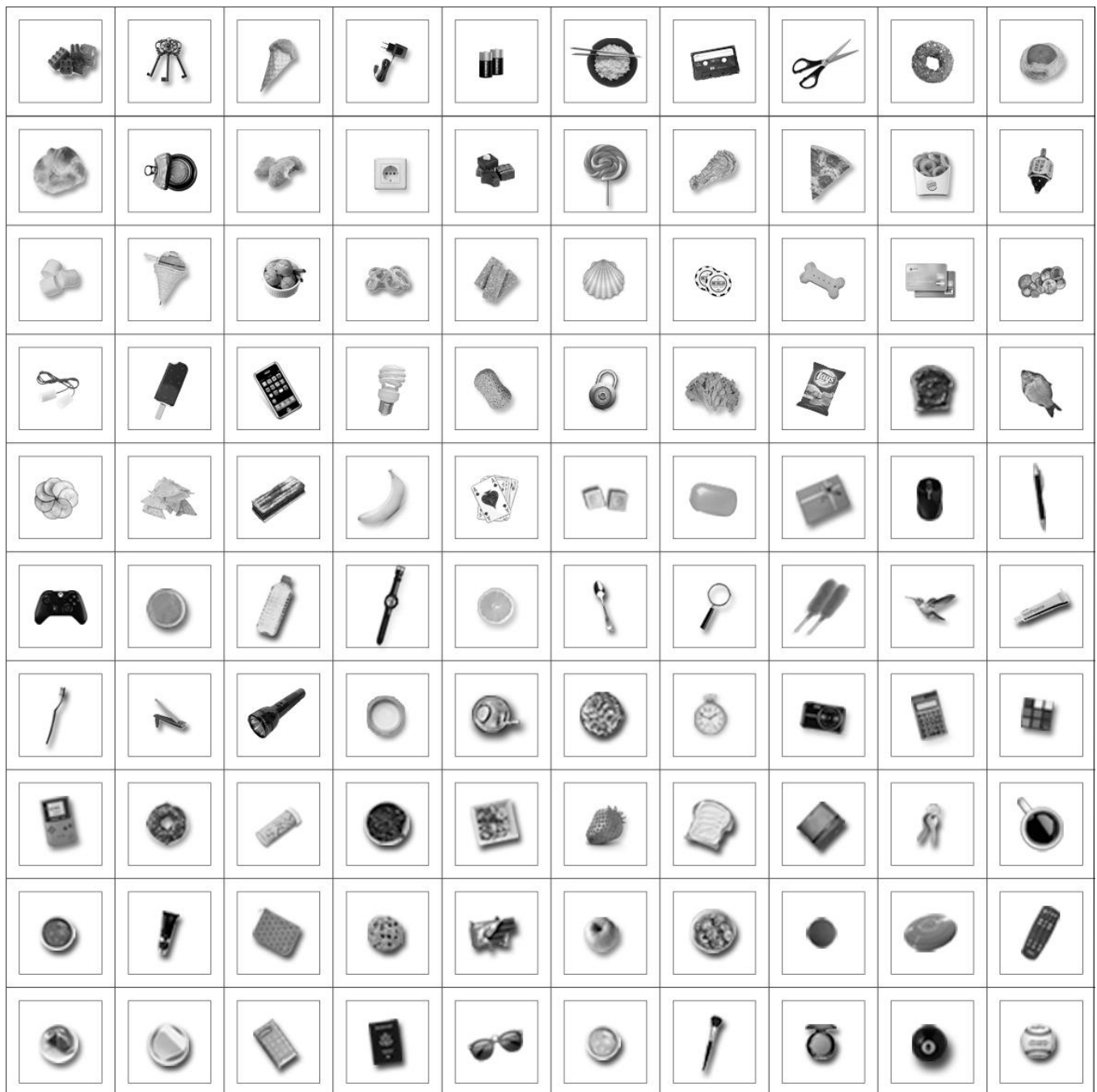


Figure 5.11: Pool of objects used for training and testing in the simulation study with natural objects. The figure shows various examples of natural objects that were presented in the same location with respect to the hand during training and testing. Different objects were used during training and testing. The objects were always presented in the same hand-centred location in order to explore how the shape selectivity of neurons representing that location depended on the number of objects presented there during training.

particular position with respect to the hand. We then tested the network with a pool of novel objects presented in the same hand-centred location. Then we systematically increased the number of objects that appeared in the same hand-centred location during training, and tested the network again with the same set of novel objects. The aim was to investigate how the shape selectivity of neurons that learned to respond to that hand-centred location was affected by the number of natural objects seen there during training.

During the presentation of the training images the activation of individual neurons within each of the layers and their firing rates are calculated. Subsequently, the synaptic weights between layers are updated according to the Hebb rule (see Equation 3.1). The network is trained one layer at the time, starting with layer 1 and finishing in layer 4.

One training epoch consisted of the presentation of all the training images, where each image contained a hand with a natural object always located in the same position with respect to the hand. In these simulations the network was trained for fifty epochs per layer. The learning rates used were 0.1, 0.1, 0.1 and 0.1 in layers 1 to 4.

After training was completed, the network was tested with a 100 images showing the hand and a novel object. The novel objects were always shown in the same hand-centred position that was used during training.

5.9.4 Results

By testing the network on images with a variety of novel objects in the same hand-centred location used during training, it was possible to assess whether the cells that had learned to respond to that hand-centred location would fire selectively to objects of a particular shape. A variety of experiments were performed with sampling different objects during training, the results presented here are taken from one of these experiments.

Figure 5.12 shows the average number of novel objects that the hand-centred cells in the network respond to after training as a function of the number of objects that the network has seen at that hand-centred location during training. The ordinate corresponds to the percentage of novel objects that the cells respond to while the abscissa corresponds to the number of objects seen during training. We can see from these simulations that the cells with hand-centred receptive fields start to lose their shape selectivity as they get trained with more and more objects in the same hand-centred location. Even when we still find a few shape selective cells, the proportion of highly selective cells is substantially reduced as the training is increased. This means that most of the cells will respond to the presence of an object in a region of space near the hand regardless of the form of the object.

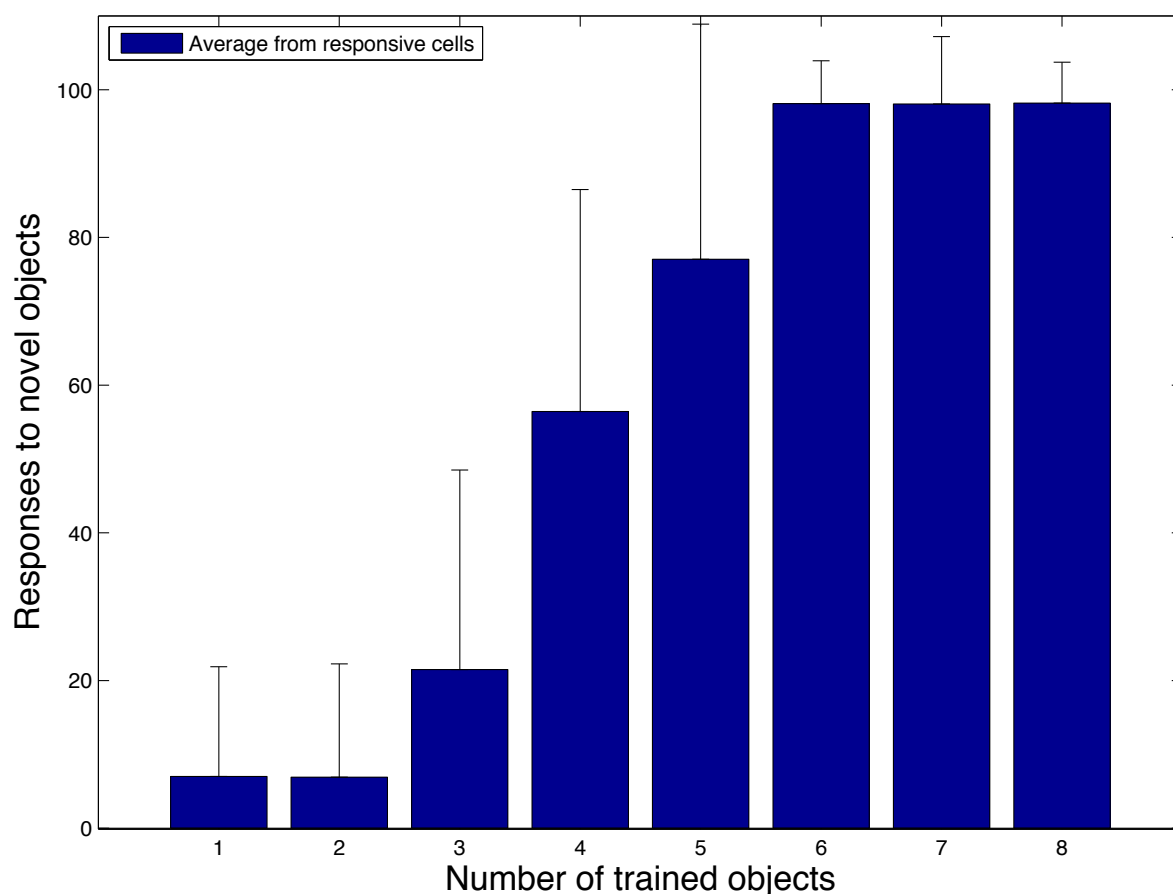


Figure 5.12: Reduction in the shape selectivity of neurons as the network is trained on increasing numbers of natural objects presented in the same hand-centred location The plot shows how, as the network is exposed to more objects during training, most cells start to lose their shape selectivity and respond to a larger percentage of novel objects.

5.9.5 Discussion

What learning mechanism leads to a reduction in the shape selectivity of neurons as the network is trained on increasing numbers of objects at the same hand-centred location?

When the first object is presented with the hand during training, a small subset of output

neurons will win the competition and respond. Then Hebbian associative learning in the feedforward connections within the network will increase the tuning of these cells to respond to that particular object in that hand-centred location. However, when another object is presented in the same hand-centred location, the two objects may share some features in common. The activation of these common features may then cause the same subset of output neurons to respond again because the relevant feedforward connections were strengthened during training with the first object. The effect of this will be to associate the features of the new object with the same output neurons. This process may be repeated with a number of successive different objects presented with the hand. All of the features of these objects will become associated with the same output neurons. Thus, the output neurons gradually lose their selectivity to the form of the objects, and merely respond to any object presented in that hand-centred location. This would produce receptive fields that represent the locations in which the objects appear with respect to the hand without being particularly selective about the differences between the features of these objects. Thus, as the results show, as the network is trained with more and more objects, the localised hand-centred receptive fields start to lose their shape selectivity and respond to a variety of novel objects as long as they appear within the hand-centred receptive field. This learning process is somewhat similar to continuous transformation (CT) learning, which drives the development of invariant neuronal responses by exploiting the similarities between visual stimuli.

In previous work VisNet has been used to model the development of object-selective responses in the ventral visual pathway. In these earlier simulations, the objects were presented either on their own or with randomised combinations of other objects during training. This ensured that all of the visual resources were available to encode the different features of the trained objects. This, in turn, enabled the network to orthogonalise the representations of the different objects, which were encoded by different output neurons. Hence, in these simulations, neurons learned to show a reasonable degree of object selectivity. However, in the simulations of the dorsal visual pathway presented here, in which objects are always presented with the hand, the shape selectivity of neurons drops off quite rapidly as the number of trained objects is increased. This is because the neurons had to learn to encode a combination of both the visual features of the hand (which is always present) as well as features of the object. This meant that neurons in the higher layers of the network had less resources to devote to representing the structural details of the objects, and orthogonalise their responses based on fine dissimilarities between objects. Furthermore, since the hand is common to all of the training images, this major common feature may cause neurons in higher layers to generalise their responses across images containing different objects. Consequently, the output neurons were not able to show the same level of shape selectivity.

We hypothesise that the above analysis of model behaviour helps to explain why neurons in different regions of the dorsal stream seem to show limited selectivity for object form.

In this case, the incoming signals required to represent the location of the hand within the visual scene reduces the available resources left to form a detailed representation of the multiple features of the object near the hand. Additionally, the overlapping signals of the hand would activate the same subset of output cells, causing the network to generalize even further to a large range of novel objects. Therefore, the network would start forming localised hand-centred receptive fields instead of object-specific representations. Moreover, it has been suggested that the sight of objects near the hand biases visual processing towards the lower-spatial resolution magnocellular visual pathway (Derrington and Lennie, 1984), associated with the dorsal visual stream. Experiments suggest a lower spatial resolution in the visual representation of near hand space (Gozli et al., 2012) giving rise to detrimental effects on feature binding and the formation of object-specific representations (Gozli et al., 2014). These various points may explain the difference in the proportion of cells that are highly selective to shape between the ventral and dorsal streams as well as accounting for the lower spatial resolution and poor feature binding and object recognition in near hand space.

However, although our results show some degree of generalization to a range of novel objects after being trained with a few objects, this does not preclude the possibility of some neurons maintaining selective responses to particular object features. As we increase the training, the number of highly selective cells is reduced. However, as the results have shown, we still find a few cells that maintain a level of shape selectivity. The amount of training

required to reduce the shape selectivity of neurons depends largely on the resolution of the model and should not be taken in absolute terms. For example, if we were to increase the resolution of the retina and intermediate layers of the model, we would increase the quality and resolution of the visual input and how it is represented in the intermediate layers. This could reduce significantly the level of overlap between the features of the different objects and allow output cells to maintain a higher degree of selectivity to specific shapes. In this case the network would have to be trained with a larger number of objects in order to decrease the selectivity of the neuronal responses.

Furthermore, some areas of the dorsal stream that seem to be more selective to object shape could be receiving better quality visual input signals that provide higher resolution representations of objects. Consistent with this hypothesis, connections have been reported between the ventral pathway area TE and parts of the intraparietal sulcus (IPS), which would include area AIP for example (Webster et al., 1994). Additionally, different objects and shapes produce different tactile feedback, as well as requiring different hand shapes during grasping. AIP receives top-down signals representing the motor command needed to grip an object from the ventral premotor cortex (area F5). This may also guide competitive learning in area AIP and thereby help to produce selectivity to object shape. Thus, these additional sources of feedback could provide a top-down signal that helps to separate the visual representations according to shape.

5.10 Training the network with natural scenes

In the previous sections of this chapter it has been shown that VisNet is still capable of developing single, localised, hand-centred receptive fields even as we started to increase the ecological validity of the visual training conditions by either increasing the number of target objects presented simultaneously with the hand or by using more realistic images of natural objects. However, the stimuli that I have used so far have still been somewhat contrived in that during training the objects were presented in no more than one or two predetermined hand-centred locations at a time.

The aim of the following section was to explore whether the network could still develop localised hand-centred receptive fields in an even more ecological training regime, in which the hand is shown against a large collection of different natural background scenes. In this case, the visual features representing background objects are present across the whole visual scene. Furthermore, the background scenes used were entirely natural with no careful control of what objects were present and where they were located.

The hypothesis I explored was essentially the same as investigated in the previous simulations described above. That is, I tested whether the statistical decoupling between objects in different hand-centred locations would allow individual neurons to start extracting combinations of the features of the hand together with features of objects seen in the same localised space near the hand. In simulations with the hand presented against natural scenes, the background objects are not seen in every possible pair of hand-centred loca-

tions during training, as was simulated above in section 5.1. However, I speculated that this would not be necessary to achieve the required level of statistical decoupling between different hand-centred locations. Even when not every possible pair of hand-centred object locations is seen during training, the features representing a particular spatial configuration of object location and hand will be seen together more frequently than features from different object locations. As long as the objects at a particular hand-centred location are paired with objects at two or three other hand-centred locations during training, then this should be enough statistical decoupling to drive the development of neurons that respond to only a single localised hand-centred location. This is consistent with recent modelling work using VisNet, where it has been shown that the network does not necessarily have to be trained in every possible pairing in order to start developing individual representations (Tromans, 2012). We anticipated, however, that since we are not carefully controlling the regions near the hand in which the objects appear, the receptive fields will develop with different sizes and shapes. In this section the images of hand-object configurations are shifted across different retinal locations during training while using trace learning to drive the development of location invariant hand-centred responses.

5.11 Methods

5.11.1 Visual stimuli

In order to generate our pool of natural visual scenes, I asked 4 volunteers to provide 10 to 12 pictures of natural visual scenes from their everyday life in which they would normally use their hands to manipulate objects. All of the subjects were naive and unaware of the purpose of the study. I provided several examples (e.g. using cutlery in a meal, grasping a cup) and provided three sample photos in order to give them a general idea of the nature of the scenes I was interested in collecting. I provided further instructions regarding the angle and distance at which the photos should have been taken. The pictures were meant to be taken from a first person point of view and the distance between the objects and the camera should be at arm's length. Additionally we asked them not to include the image of their own hand in the picture.

The training stimuli for this experiment consisted of images showing a picture of a real hand that was superimposed in all of the natural visual scenes collected by our participants. The templates were scaled and arranged using Adobe Photoshop software. The images were generated in RGB color and subsequently converted to monochrome using the MATLAB function `rgb2gray` and then resized to a 256×256 matrix. Figure 5.13 shows a sample of some of the training images that were generated for this study.

A total of 48 natural images were used. In order to present the configurations of the hand and objects in different retinal positions, each of the fixed spatial configurations was

translated by five pixels at a time across VisNet’s ‘retina’ within a 3 by 2 grid. In other words, this time the sequences included horizontal as well as vertical shifts on the network’s ‘retina’.

After training was completed, the stimuli used during testing consisted of images showing the hand and a novel textured object in five different positions relative to the hand as shown in Figure 5.14.

5.11.2 Model parameters

In this experiment we continued using an up-scaled version of the model ‘retina’ (i.e. 256×256). The rest of the parameters remained unchanged.

5.11.3 Training

The training procedure for this experiment consisted of presenting VisNet with images of the hand embedded within 48 different natural scenes containing a variety of objects as shown in Figure 5.13.

As in previous simulations, the image sequences are meant to arise from a series of eye movements and the resulting shifts in the position of the hand and visual objects on the 256×256 ‘retina’. During each of the image sequences, the fixed spatial configuration of the visual scene is translated five pixels at a time. In previous simulations we have been shifting the images only horizontally, however in order to increase the ecological validity

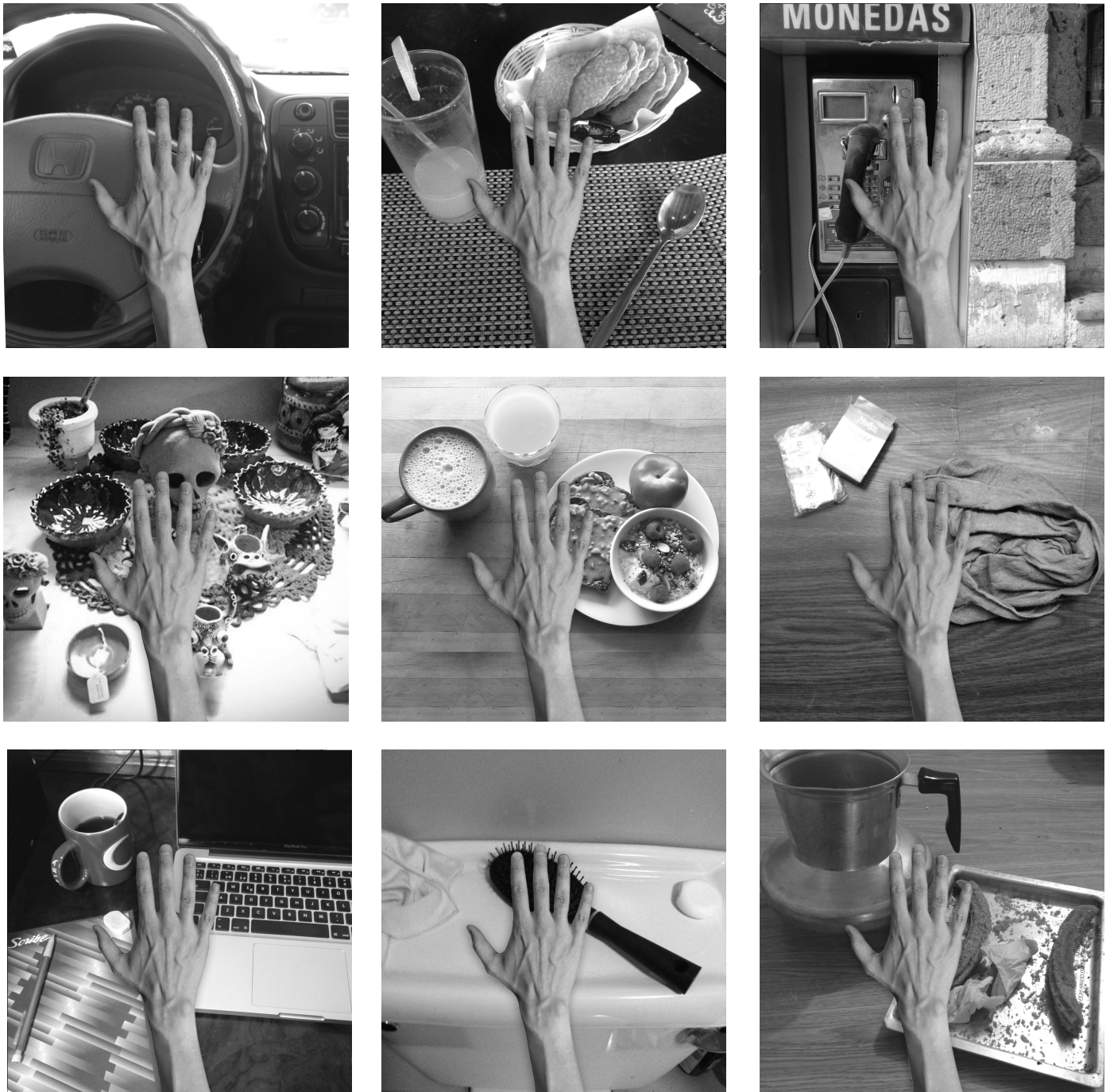


Figure 5.13: Realistic training images used for simulations with the hand presented against natural backgrounds during training. The figure shows various examples of the hand presented against different natural backgrounds during training. The position of the hand within each of the backgrounds is unchanged during the eye movements.

of this simulation, we also include a vertical shift of five pixels.

During the presentation of every frame of the image sequence we calculate the activation of individual neurons and their firing rates. The synaptic weights between layers are updated according to the trace rule (see Equation 4.2). This process is then repeated for every layer of the network. The learning rates used were 0.1, 0.1, 0.1 and 0.1 in each layer.

5.11.4 Testing

As in previous simulations, the synaptic weights were not changed during the testing phase. In order to test whether VisNet has been able to develop translation invariant neurons with a single, localised, hand-centred receptive field, image sequences of the hand and only a single textured object are presented in five different hand-centred locations. The outputs of the network are computed during the presentation of each hand-object configuration in each retinal location. Figure 5.14 shows an example of the five testing images presented to the network. Each of these images were tested in six different retinal locations presented on a 3 by 2 grid.

5.12 Results

Figures 5.15 and 5.16 show the response profiles of five neurons in the output layer of VisNet before training and after training, respectively. Following the same conventions of previous chapters, each of the five columns of plots contains the firing responses of a

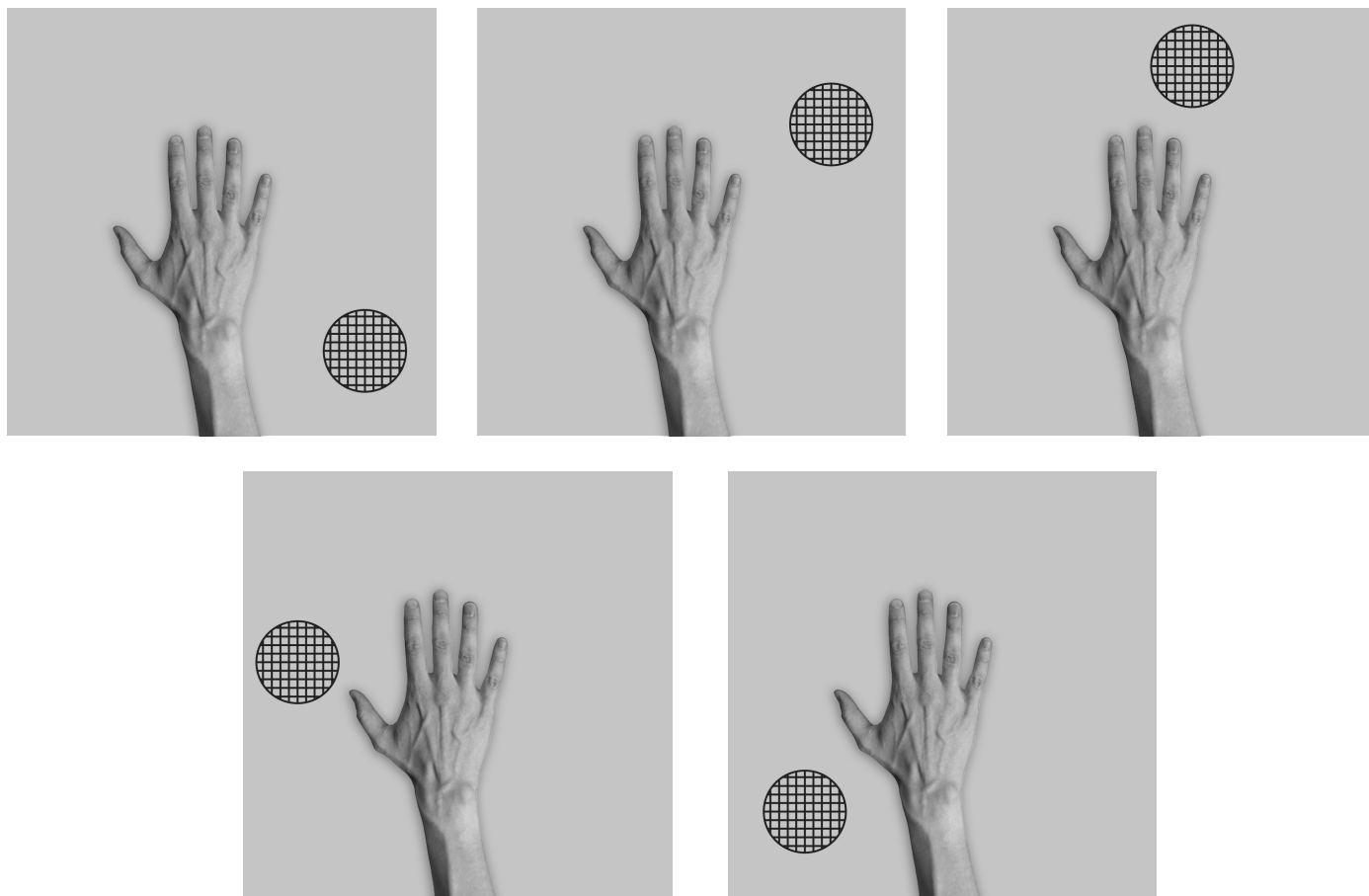


Figure 5.14: Testing with a plain background and one textured stimulus. These are the five test images used after the network had been trained with the hand presented against natural background scenes. That is, the network was tested with a single textured stimulus presented in five different hand-centred locations. Each hand-object configuration was tested in the six different retinal locations that the hand was trained in.

particular output cell, these are labelled at the top of the column. The five rows show the responses of the cells in response to each of the five hand-object configurations presented during testing. Each plot shows the responses of the given cell to the particular hand-object configuration over six different retinal locations.

Before training (Fig. 5.15) none of the cells respond exclusively to any of the hand-object configurations; in fact they respond rarely. However, after training, in Figure 5.16 we can see that each of the five cells learned to respond exclusively to one specific hand-object configuration, and that these responses were invariant to different retinal locations.

As in previous studies, an information analysis was carried out to investigate how these hand-object configurations are represented across the whole population of output cells. Figure 5.17 shows the single and multiple cell information measures for the output (fourth) layer neurons before and after training the network on images of the hand presented against natural backgrounds. The information analysis was performed by testing the network on the five hand-object configurations shown in Figure 5.14, where each such configuration was presented in six retinal locations.

Figure 5.17 (top) shows the single cell information measures for the output layer of neurons. We can see here that, before training none of the cells reached the maximum information. However, after training, 49 neurons reached the maximal single cell information of 2.32 bits. This means that these 49 output cells responded selectively to a single localised position of the test object with respect to the hand, and that this response was

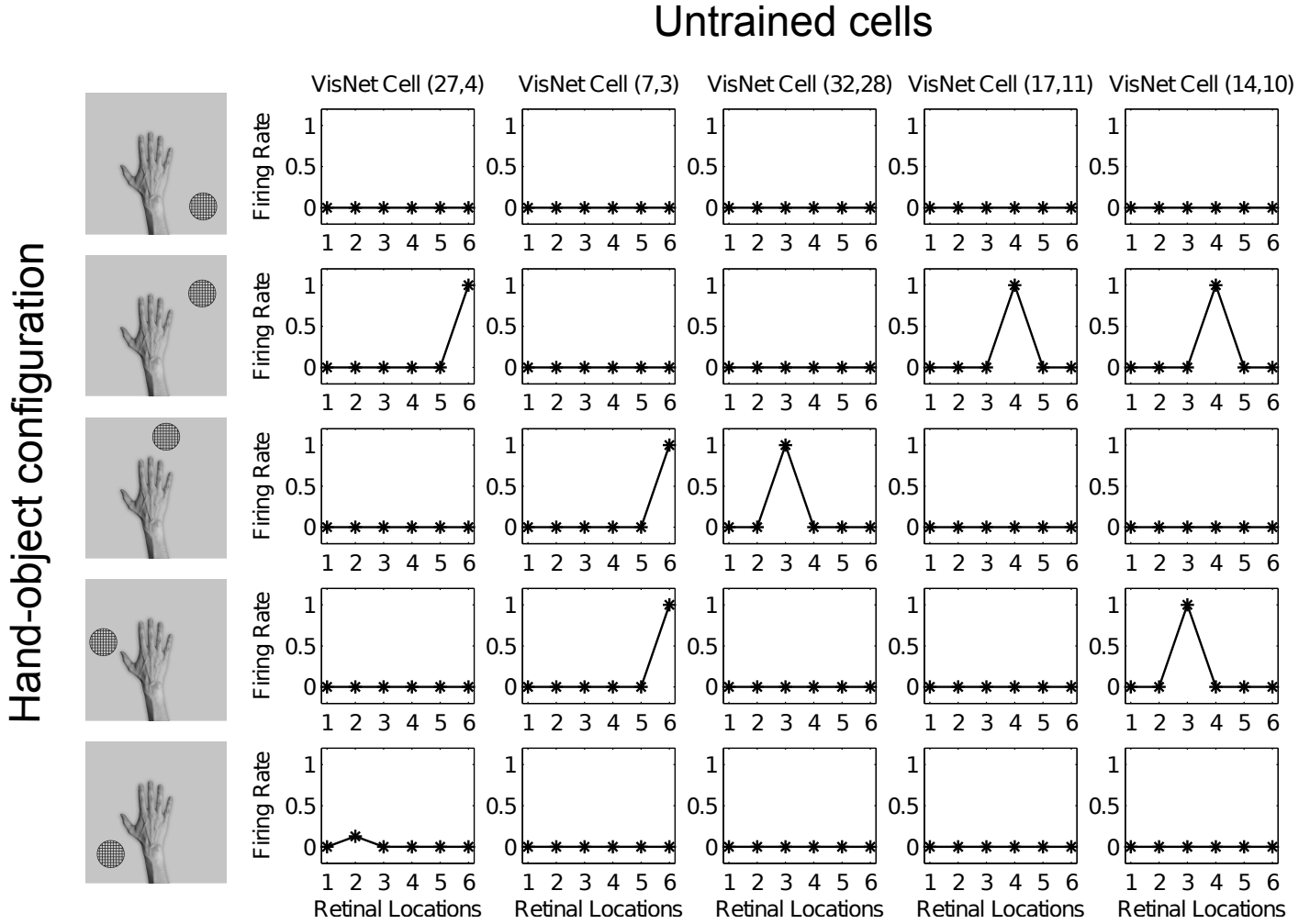


Figure 5.15: Response profiles of five neurons in the top (fourth) layer of VisNet before training. Each of the five columns shows the firing responses of a particular cell. Each row shows the responses of the cells to one of the five hand-object configurations over all six different retinal locations shown along the abscissae. It can be seen that each of the five cells initially responds randomly to each of the hand-object configurations over the different retinal locations.

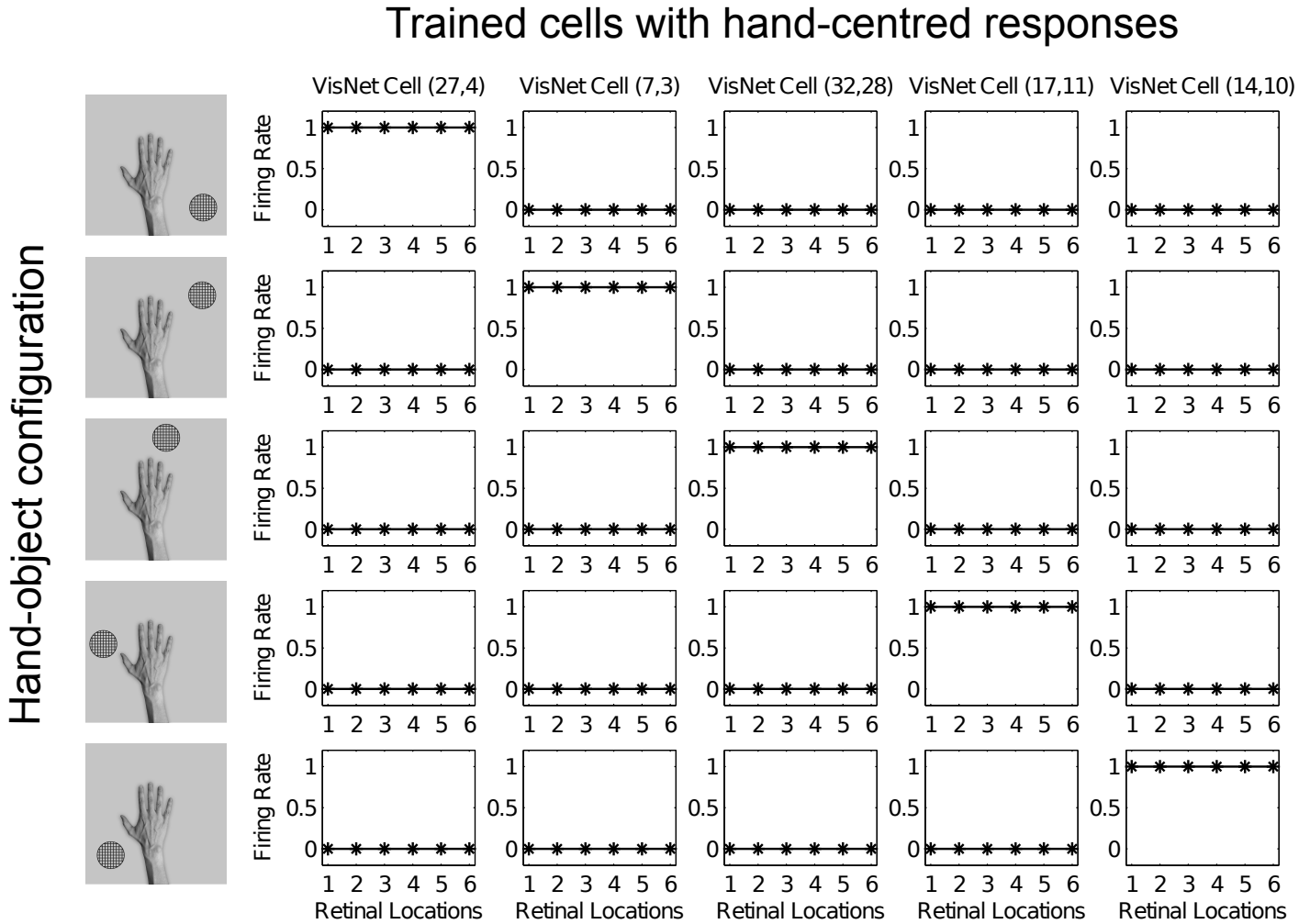


Figure 5.16: Response profiles of same five neurons in the top (fourth) layer of VisNet after training. Response profiles of the same five neurons from Fig. 5.15 after training on the images shown in Fig. 5.14. Conventions as in Fig. 5.15. It can be seen that each of the five cells has learned to respond selectively to just one of the hand-object configurations, and responds to that configuration over all six retinal positions shown along the abscissae. Moreover, each of the five hand-object configurations is represented by one of the cells.

invariant to retinal location.

In Figure 5.17 (bottom) it is evident that, before training, the multiple cell information does not reach the maximal value of 2.32 bits. However, after training we can see that the multiple cell information asymptotes to the maximal value, which means that all of the possible hand-object configurations are successfully represented by separate cells in the output layer.

In fact, the five cell response profiles after training shown in Figure 5.16 already confirmed that the network was able to represent each of the five hand-object configurations. The multiple cell analysis simply reaffirms that all five hand-object configurations are represented invariantly across all retinal locations by separate output neurons.

For this simulation we again traced the strengthened connections from each one of the output cells through successive layers to the input Gabor filters driving that cell. In Figure 5.18 we can see the Gabor input filters with strengthened connections to a trained output neuron that had learned to respond to one of the hand-centred locations. On the left side of Figure 5.18 we can see the Gabor filters that are most strongly driving the responses of the particular output cell. This cell is representing a subset of Gabor filtered inputs corresponding to the hand, as well as a subset of inputs representing a localised region near the hand. The right side of Figure 5.18 shows the image of the hand with the hand-centred receptive field of the neuron shown in blue.

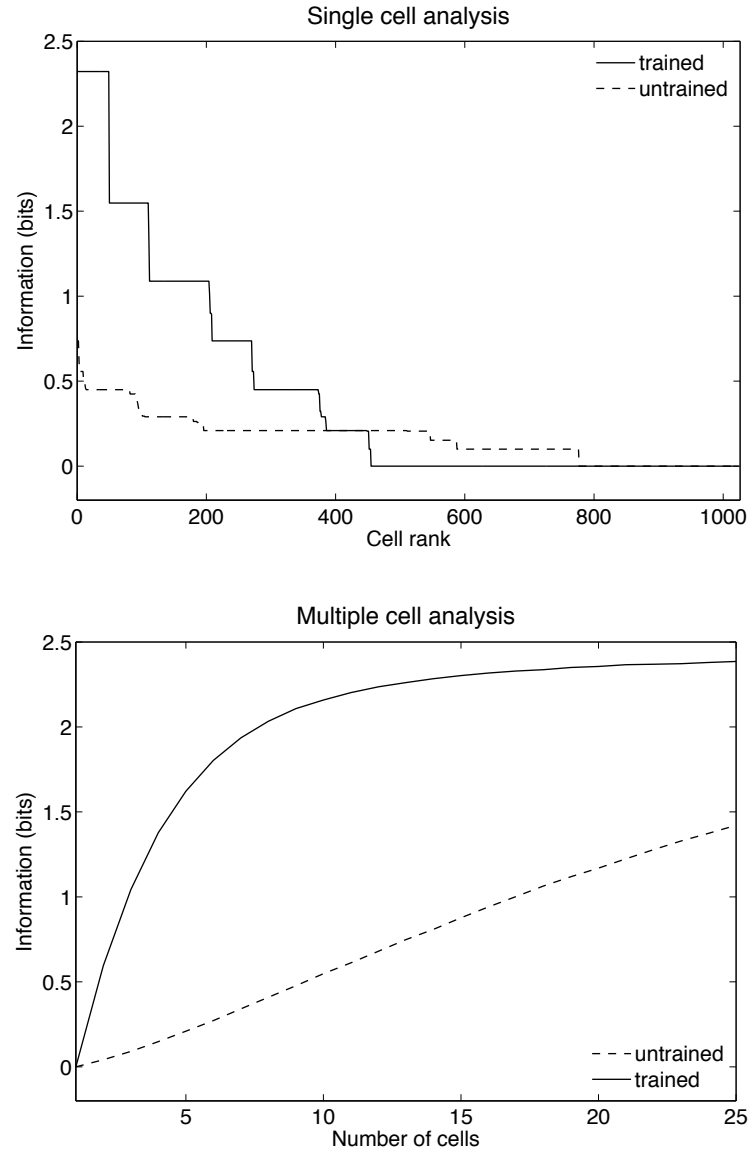


Figure 5.17: Information analysis carried out before and after the network has been trained on images of the hand presented against natural backgrounds. The information analysis was carried out by testing the network on the five hand-object configurations shown in Figure 5.14, where each such configuration was presented in six retinal locations. The upper plot shows the amount of single cell information carried by individual output cells in rank order. After training, it was found that 49 cells reached the maximum amount of single cell information of 2.32 bits. These cells responded perfectly to just one of the five tested hand-object configurations, and responded to that configuration across all six different retinal locations. In the untrained condition no cells reached maximal information. The lower plot shows the multiple cell information measures calculated across 25 cells with maximal single cell information. It can be seen that, after training, the multiple cell information asymptotes to the maximal value of 2.32 bits. This confirms that all five tested hand-object configurations are represented by the output cells.

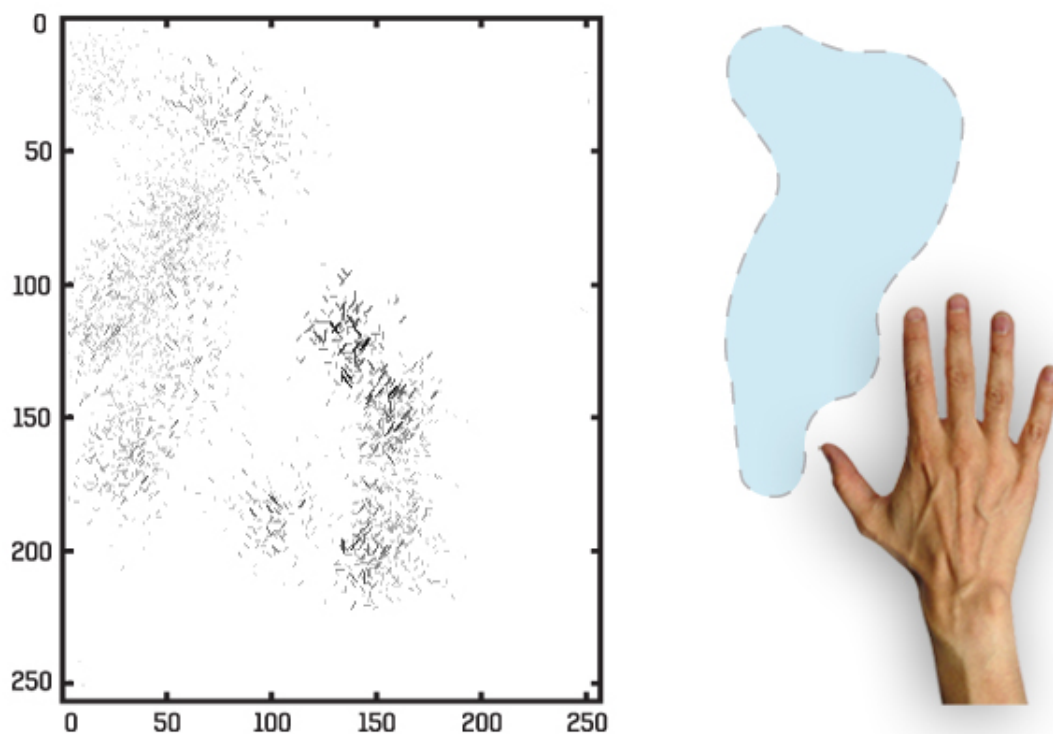


Figure 5.18: Tracing back the synaptic connections from a trained output cell to the input Gabor filters after the network was trained on images of the hand presented against natural scenes. The left side shows the input Gabor filters that an output cell has learned to respond to after training. In this image the Gabor filters with the strongest connectivity through the layers to the output cell are plotted, where each Gabor filter is weighted by the strengths of the feed-forward connections from that filter through the successive layers to the output neuron. The right side shows the image of the hand with the hand-centred receptive field of the neuron shown in blue.

5.13 Discussion

In the simulations presented in this chapter we have investigated whether VisNet could still self-organize and develop neurons with single, localized hand-centred receptive fields, as the network is trained in more realistic training conditions. Much of VisNet's research in the past has been carried out with relatively controlled training regimes and stimuli. However, in each subsection of this chapter, we have systematically improved the realism of the visual training stimuli in order to test the robustness of the proposed learning mechanism that relies on a combination of statistical decoupling between object locations and trace learning in order to drive the development of hand-centred visual representations.

The self-organisation of localised, hand-centred responses in the network is actually driven by a complex combination of factors including the competitive nature of the layers in the network, the learning rule used to modify the synapses, and the visual training stimuli. For example, although the trace learning rule binds input patterns together using temporal continuity, the trace rule incorporates a purely Hebbian associative component. Because of this, the trace rule may also bind input patterns together on the basis of spatial similarity using continuous transformation learning. Both of these learning mechanisms may operate concurrently. In fact, there is an interesting interplay between all of these mechanisms which allows the network to form localised hand-centred responses instead of forming object-specific or background-specific representations on which the network has been trained. The interplay is instrumental in allowing the network to extract relevant

features and generalize them to different yet sufficiently similar hand-object configurations.

We have shown how some neurons learn to respond to particular spatial configurations of the hand and an object location. Such neurons represent the location of a visual object in the reference frame of the hand. This learning process exploits the statistical decoupling that will exist between different hand-centred object locations across many different images. Furthermore, these neuronal responses can become invariant across different retinal locations by trace learning. This learning rule binds together input patterns which tend to occur close together in time. If the eyes typically saccade around a visual scene faster than the hand moves, then trace learning will bind together the same hand-object configuration across different retinal locations.

In Section 5.9 we investigated how the shape selectivity of neurons was affected by the number of objects that the network was trained on at a particular hand-centred location. Whenever a new object is shown at a particular hand-centred location, then there will likely be some overlap of features with previous objects presented at that location. In this case, it is likely that some of the same output cells will fire again to the presence of the new object. These cells would get their synaptic weights from the features of the new object strengthened. As the network is trained on more and more objects at the same hand-centred location, this subset of cells gradually learn to respond to most object features at that location and hence lose their shape selectivity. Our simulations suggest the possibility that hand-centred neurons in area 5 and other parts of the posterior parietal cortex may

in fact display a range of different degrees of object shape selectivity. The responses of some neurons may be still somewhat selective to shape, while other neurons respond to almost all objects placed within their hand-centred receptive field. Such a heterogeneous population of neurons was in fact observed in our simulations.

In Sections 5.3 and 5.7 we began to address how the network might develop neurons with single, localised, hand-centred receptive fields if it is trained on more realistic images containing multiple objects presented simultaneously with the hand. Specifically, we showed that by presenting the objects in many different pairs of hand-centred locations during training, this could facilitate the statistical decoupling between different object locations which would delineate the receptive fields to represent a localised region. This allowed us to train the network with more than one object presented at the time with the hand.

In Section 5.10 we further increased the realism of the simulations by training VisNet on images of the hand presented against natural visual scenes. Unlike the previous simulations where the hand-centred object locations were carefully controlled, this time the objects could appear in any location around the hand. One of the particularly interesting differences in the results was that instead of developing clearly delineated visual receptive fields (depending on the stimuli shown), the localised receptive fields near the hand had idiosyncratic shapes as shown in Figure 5.18.

Altogether, the results from this chapter demonstrate how individual output cells could

develop single, localised, hand-centred visual receptive fields which are invariant to retinal location, even when the network is trained on more realistic visual scenes with multiple objects presented simultaneously with the hand or even with the hand presented against complex natural backgrounds. However, despite the notable improvements in the realism of the images on which VisNet was successfully trained, the dynamics of the eye movements were still controlled artificially by pre-specified shifts. In the last experiment we incorporated both vertical and horizontal shifts. Nevertheless, the richness and complexity of the dynamics of natural eye movements from human subjects has not yet been explicitly incorporated to guide the retinal shifts in VisNet during training. The next chapter will present a new time-accurate differential model of VisNet, in which for the first time, the retinal shifts of the network will be guided according to data recorded from the natural eye-movements of human subjects collected with an eye-tracker during a hand-object manipulation task.

”Cognitive psychologists must make a greater effort to understand cognition as it occurs in the ordinary environment and in the context of natural purposeful activity. This would not mean an end to laboratory experiments, but a commitment to the study of variables that are ecologically important.”

— Neisser, U., *Principles and Implications of Cognitive Psychology* (1976), p.7.

6

The visual development of hand-centred receptive fields in a time-accurate model of VisNet trained with experimentally recorded human gaze changes

In this chapter we present experiments in which, for the first time, the retinal shifts in VisNet are driven by gaze changes recorded from participants. The purpose of this chapter is to confirm that the statistics of natural eye and head movements are capable of driving the development of hand-centered neuronal response characteristics. The participants undertook an experimental task in which they had to complete a jigsaw. During this, the participants were free to move their eyes and head. A camera mounted on the head captured images of the hand and jigsaw, while eye movements were recorded by an eye-tracking device. This combination of data allowed us to reconstruct the retinal images as humans

undertook the jigsaw task. In order to feed the sequences of retinal images to the network while faithfully representing the temporal dynamics of gaze changes recorded from human subjects, we implemented a new time-accurate version of VisNet which allowed us to feed the input data continuously. In this chapter we show how this differential model could develop hand-centred visual representations using trace learning. In the data recorded from the participants, we found that the subject's gaze often shifted through a sequence of locations around a fixed spatial configuration of the hand and jigsaw pieces. In this case, trace learning will bind these retinal images together onto the same subset of output neurons. In this way, trace learning encourages cells to respond to particular hand-object configurations across different retinal locations. The simulation results showed that before training the output cells responded in a random and unstructured way, while after training some output cells had learned to respond to specific hand-centered locations irrespective of retinal location. This led to an increase in the amount of single cell and multiple cell information that the output cells carried about the presence of visual objects in different hand-centered locations. Individual output cells fired selectively to the hand and a puzzle piece in a fixed spatial configuration across a large number of retinal views. These simulated firing properties successfully replicate the response characteristics of hand-centered neurons previously found in the primate brain during single unit recording studies. The results from this study have been submitted for publication (Galeazzi et al., *submitted*).

6.1 Introduction

In Chapter 4 I proposed a computational hypothesis of how *trace learning* may allow neurons to develop selective responses to the location of visual targets with respect to the hand that are invariant to shifts in retinal position (Galeazzi et al., 2013). I proposed that for a portion of the time humans shift their eyes around static visual scenes that contain their hand with other nearby objects in a fixed spatial configuration. In this case, trace learning will bind together these retinal images onto the same subset of higher layer neurons, which will then respond to particular hand-object configurations regardless of retinal position. Such cells effectively encode the hand-centered locations of visual targets, as reported in neurophysiology studies. This hypothesis was tested in our unsupervised, self-organising neural network model, VisNet, of the primate visual system. Simulations confirmed the plausibility of this hypothesis, and showed how different output cells learned to respond selectively to different object positions with respect to the hand (Galeazzi et al., 2013). In Chapter 5, I have demonstrated the ability of our model to develop hand-centered visual representations even when it is trained using highly realistic images, in which the hand is seen against natural scenes with multiple objects present at the same time (Galeazzi et al., 2015).

In the current chapter, I show how hand-centered visual receptive fields can develop during visually guided learning in VisNet when the model is driven by gaze changes recorded from human participants. The purpose of this study is to show how the statistics of natural

eye and head movements are capable of driving the development of such hand-centered neuronal response characteristics. The participants undertook an experimental task in which they had to complete a jigsaw. During this, the participants were free to move their eyes and head. A camera mounted on the head captured images of the hand and jigsaw, while eye movements were recorded by a head-mounted eye-tracking device. This combination of data allowed us to reconstruct the retinal images as humans undertook the jigsaw task. These sequences of retinal images were then fed into the simulations of the neural network model during self-organisation of its synaptic connectivity using trace learning.

The patterns of eye and head movements to change our gaze are closely related to the nature of the task (Yarbus et al., 1967). In a hand-object manipulation task, sequences of gaze changes are triggered for multiple purposes. Generally, we have to first explore the visual scene in order to gather information about the properties and location of the objects we want to manipulate, as well as using visual information to guide and control the movement.

In natural tasks of visually-guided movements there is a certain amount of variability in how we coordinate the movements of our head and eyes together with the movements of our hands. Eye, head and hand coordination seem to be dependent not only on the nature of the task, but also on the different strategies used by the participants. However, a number of studies have reported latencies between the initiation of the eye movements and the movement of the hand, where in most cases the eye movements seem to initiate

and lead the movements of the hand (Angel et al., 1970; Mather and Fisk, 1985; Prablanc et al., 1986). For example, eye movements in a natural task involving making a cup of tea were observed to initiate 560ms before the movement of the hand (Land et al., 1999). Interestingly, the eye-movements in this study were found to heavily monitor every single aspect of the activity despite the fact of being a relatively automated task. Differences in latency can also be affected depending on the attentional demands of the task (Pelz et al., 2001), as well as if it is automated and repetitive. There are tasks where these latencies can be significantly reduced (Abrams et al., 1990).

We hypothesized that while a subject is solving a jigsaw puzzle, there would be periods of time in which the hand would remain stationary while the participants explored the visual scene composed of the hand and an object near the hand (i.e. a jigsaw puzzle piece). It has been previously shown that trace learning is robust enough that it can allow a competitive network to develop non-retinal representations even if the training regime is composed of a wide variety of stimulus dynamics, as long as the stimulus dynamics required for self-organization occur for a small portion of the time (Mender and Stringer, 2014). Therefore, trace learning can still function even if the majority of the time the dynamics of gaze changes and hand movements are entirely different to what is required in our self-organizing model.

In the data recorded from the participants, we used the periods of time in which the subject's gaze often shifted through a sequence of locations around a fixed spatial configu-

ration of the hand and one of the jigsaw pieces to train our network. Trace learning in our model simulations was then able to bind these retinal images together onto the same subset of output neurons. In this way, trace learning encouraged cells to respond to particular hand-object configurations across different retinal locations. Thus, after training, some output cells responded to specific hand-centered locations irrespective of retinal position.

Previous research with VisNet has represented time in discrete processing steps, in which each time step corresponds to an unspecified interval of time. Furthermore, the different views or retinal shifts during training have always been controlled artificially by pre-specified shifts. However, in order to feed video images to the network that faithfully represent the temporal dynamics of gaze changes recorded from participants, we implemented a new time accurate differential formulation of the VisNet model. This is important for two reasons. Firstly, reality unfolds continuously and a discrete scheme does not allow the model to faithfully and realistically represent the continuous temporal flow of visual inputs as they occur in natural circumstances. Secondly, given that we are assuming that a temporal trace in the neuronal dynamics is what enables the binding of temporally concurrent views of the same hand-centred configuration, it is important to explicitly define the dynamical quantities and parameters that would govern the network. For example, in a differential version of trace learning, the trace value would be exponentially decaying over time while the neuron is not active. Therefore, these simulations would help us to establish whether trace learning could cope effectively and allow the network to form representations

of hand-centred configurations which are invariant across retinal shifts, based on how we explore a visual scene in a hand-object manipulation task.

The model simulations reported in this chapter thus confirm that retinal image sequences arising from natural movements of the eyes and head recorded from participants are able to drive the development of the kind of hand-centered visual representations that have been found experimentally in the primate brain during single unit recording studies (Bremner and Andersen, 2012; Buneo and Andersen, 2006).

6.2 Methods

6.2.1 Experimental Task

Participants

Six adults (2 males, 4 female, mean age 22.3 years, SD 2.6 years) from the University of Leicester participated in this study. All of them were naïve and with normal or corrected vision. One of the female participants was excluded from the experiment during the calibration phase due to the lack of a reliable signal from the pupil to the eye tracking equipment. This study will only show the results from the five remaining participants.

Apparatus

Eye movements were recorded with the mobile eye tracker ASL MobileEye (Applied Science Laboratories; Bedford, MA, USA) at a sampling rate of 30 Hz. This video-based eye tracker is mounted on a pair of lightweight safety glasses containing two cameras. One of the cameras is directed to the right eye, which is tracked using the pupil center and corneal reflection after calibration. The second camera records the scene. We aligned the scene camera with the line-of-sight of the participant and captured a video image of their approximate visual field. The visual range of the camera is approximately 50 degrees horizontal and 40 degrees vertical. We calibrated the scene camera to the pupil recording in order to accurately track the gaze direction within the scene. The numerical data is provided by the software in a .csv file that contains the x and y coordinates of the pupil radius in eye image pixels as well as the direction of the gaze with respect to the scene image.

Materials

In this study, we used a jigsaw puzzle manufactured by the UK company Gibsons[®]; the puzzle was made of good-quality cardboard, with 100 pieces. The puzzle displayed a painting from Trevor Mitchell called ‘A Hop, Skip and a Jump’. A completed puzzle measured 49cm x 34cm (approx 19 x 13 inches). The participants did not have any prior experience with the puzzle.

Procedure

The participants sat on a chair with a table in front of them. The participants were informed that they would be completing a jigsaw puzzle that was on the table covered by a cloth. The participants were asked to adjust their posture and distance from the table according to their own preference. At that moment we adjusted the camera recording the scene and aligned it with the forward line-of-sight of the participant and captured a video image of their approximate visual field. The scene camera was calibrated with the pupil recording in order to be able to track the gaze direction within the scene frames. The scene camera calibration required participants to fixate on each of the 4 fixation crosses that were marked across the table. After calibration, subjects were asked to re-fixate in the previously instructed points to corroborate the calibration accuracy.

After the eye-tracking calibration was successful, the participants were informed about the details of the task. They were told that after the cue provided by the experimenters, the cloth covering the jigsaw puzzle would be removed. The participants were never shown the full picture that was depicted in the puzzle. The jigsaw was presented almost completed with only 5 missing pieces that had been previously removed by the experimenters. All of the pieces removed had exactly the same shape in order to make sure that subjects would not use the shape of the piece as a strategy to find the location of the missing piece.

The participants were instructed that one of the missing pieces would appear near their hands and that they needed to place it in its corresponding location once they were sure

they had found it. The puzzle piece could appear in one of three possible locations near the hand which was resting on the table in front them. These hand-centered locations were left of the hand, right of the hand, and in front of the hand. After the participants placed the piece in the corresponding puzzle location the trial was terminated. For the next trial the experimenters would remove five different pieces from the jigsaw puzzle and repeat the procedure. The presentation of the trials was counterbalanced across subjects, as well as the order in which the missing pieces were placed around the hand (i.e. left, right, in front).

Typically the participants would explore the visual scene and saccade for a few seconds between the hand and jigsaw piece and the possible locations in the puzzle. The average length of the trials varied between subjects, having an average of 10.09 seconds with a standard deviation of 6.7 seconds. In Figure 6.1 we can see an example of the gaze changes in one of the trials, with the jigsaw piece placed to the right of the hand. It can be seen that the human participant makes a series of saccades between the hand, jigsaw piece and the puzzle. We used software to identify and record the locations of visual fixations within the scene, which are shown as dots in Figure 6.1.

6.3 Differential model of VisNet

Previous research with VisNet has represented time in discrete processing steps, in which each time step corresponds to an unspecified interval of time. However, in order to feed video images to the network that faithfully represent the temporal dynamics of gaze changes

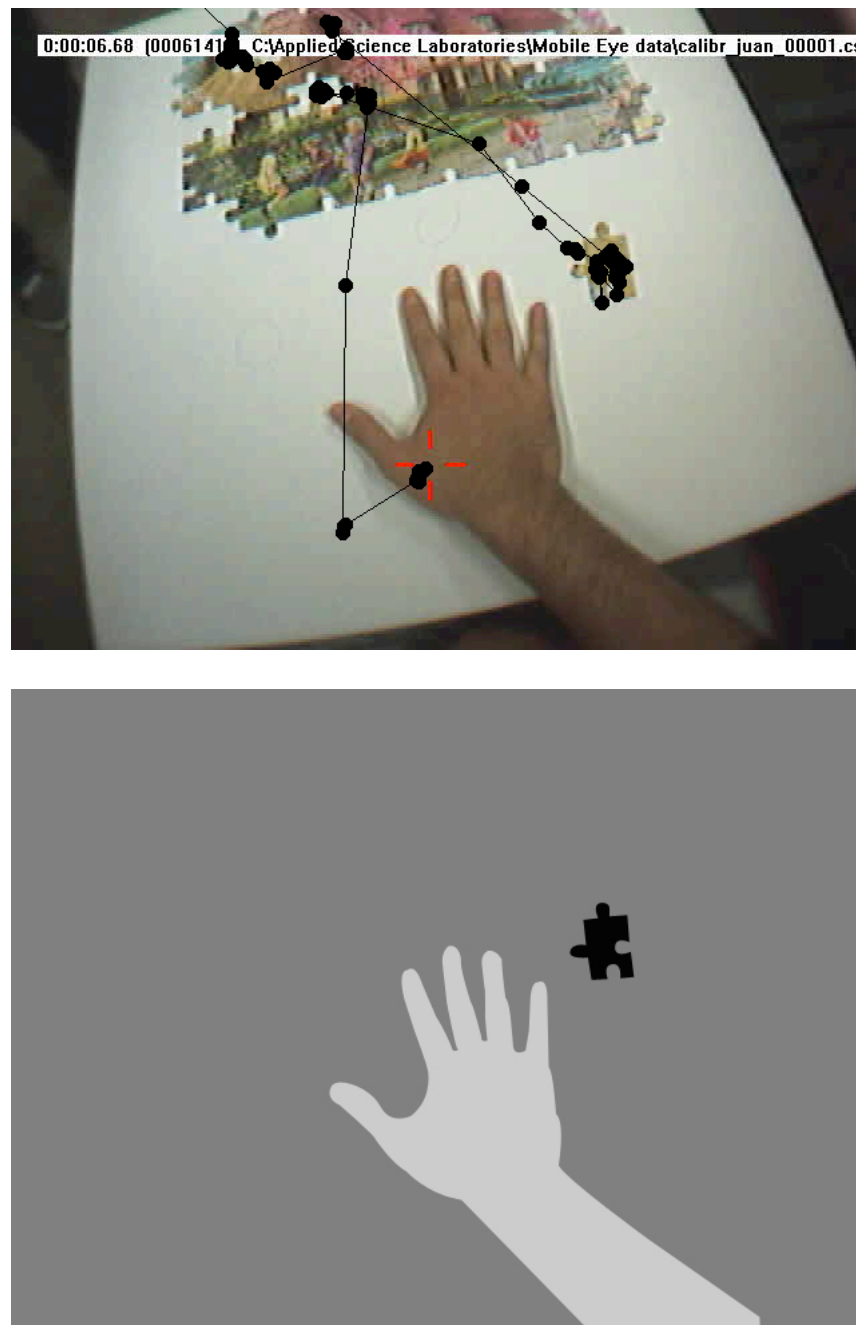


Figure 6.1: Example of the gaze changes during one of the trials and a sample of preprocessed frame. The upper image shows the dots that correspond to the locations of visual fixations extracted from numerical data provided by the software. In this trial the head remained stationary. Across participants the eye movements typically involved saccades back and forth between the puzzle piece and the potential puzzle locations in which the piece could fit. The lower image shows a sample of a preprocessed frame before it is presented to the network

recorded from human subjects, we required a new time accurate differential formulation of the VisNet model. This section provides a description of the architecture and equations that govern the time accurate version of VisNet used for these simulations.

6.3.1 Model architecture

Most of the architectural features of the model are similar to the previous discrete version of VisNet, which is described in (Rolls, 2008) and in Chapter 2 of this thesis. The time accurate version of VisNet also consists of four competitive layers of neurons with topologically organized feedforward synaptic connections between successive layers.

In this study we used a scaled up ‘retina’ in the input layer that has the dimensions 256×256 , as we did with the simulations presented in Chapter 5. Increasing the size of the ‘retina’ significantly improves the resolution and therefore the performance of the model. In order to maintain the resolution of the visual information throughout the layers, we also increased the dimensions of the intermediate and output layers to 256×256 . Scaling up the dimensions of all the layers of VisNet has been done in previous experiments (Rolls and Webb, 2014). As in previous work with VisNet, the input images are first pre-processed by an initial layer that contains a bank of Gabor filter outputs (see Chapter 2).

6.3.2 Activation

The neuronal activations and firing rates within each of the four layers is modelled as follows. Let $h_i(t)$ and $y_i(t)$ denote the activation and firing rate respectively of neuron i in layer ℓ at time t , and let $w_{ij}(t)$ be the weight of the synapse from neuron j in layer $\ell - 1$ to neuron i in layer ℓ . The activation of each cell i is calculated according to

$$\tau_h \frac{dh_i}{dt} = -h_i(t) + \sum_j w_{ij}(t) y_j(t) \quad (6.1)$$

where τ_h is an activation time constant which is constant for all cells. This equation is solved numerically using a Forward Euler finite difference method.

Within each layer competition is graded rather than winner-take-all, and is implemented in two stages at each numerical timestep. First, to implement lateral inhibition the activation of neurons within a layer are convolved with a spatial filter, I , where δ controls the contrast and σ controls the width, and a and b index the distance away from the centre of the filter

$$I_{a,b} = \begin{cases} -\delta e^{-\frac{a^2+b^2}{\sigma^2}} & \text{if } a \neq 0 \text{ or } b \neq 0, \\ 1 - \sum_{\substack{a \neq 0 \\ b \neq 0}} I_{a,b} & \text{if } a = 0 \text{ and } b = 0. \end{cases} \quad (6.2)$$

Next, contrast enhancement is applied by means of a sigmoid activation function

$$y = f^{sigmoid}(r) = \frac{1}{1 + e^{-2\beta(r-\alpha)}} \quad (6.3)$$

where r is the activation after lateral inhibition, y is the firing rate after contrast enhancement, and α and β are the sigmoid threshold and slope respectively. The parameters α and β are constant within each layer, although α is adjusted to control the sparseness of the firing rates. For the simplified case of neurons with binarised firing rates = 0/1, the sparseness is the proportion $\in [0, 1]$ of neurons that are active. For example, to set the sparseness to, say, 5%, the threshold is set to the value of the 95th percentile point of the activations within the layer. The parameters used for these simulations are shown in Table 6.1.

6.3.3 Trace Learning

Different views of the hand and a jigsaw piece in a fixed spatial configuration will tend to occur close together in time as the eyes quickly saccade around the scene. In order to bind together these temporally concurrent input patterns during learning we use a trace learning rule that incorporates a memory trace of recent neuronal activity. The trace value $\bar{y}_i$ for neuron i is governed by

$$\tau_{\bar{y}} \frac{d\bar{y}_i}{dt} = -\bar{y}_i + y_i \quad (6.4)$$

where $\tau_{\bar{y}}$ is a trace time constant that is common for all the cells.

We then update the strength of the synaptic weight from presynaptic neuron j to postsynaptic neuron i using a differential form of trace learning rule

$$\frac{dw_{ij}}{dt} = \alpha \bar{y}_i y_j \quad (6.5)$$

where α is the learning rate, y_j is the instantaneous firing rate of presynaptic neuron j , and $\bar{y}_i$ is the trace value of postsynaptic neuron i .

Finally, in order to prevent the unlimited growth of synaptic weights, we renormalize the synaptic weight vector of each postsynaptic neuron i on every numerical timestep as follows

$$\sum_j (w_{ij})^2 = 1.$$

Table 6.1 shows the model parameters used in the computer simulations.

Parameter	Layer 1	Layer 2	Layer 3	Layer 4
Dimensions	256x256	256x256	256x256	256x256
#Synapses	200	200	200	200
Learning rate (α)	0.1	0.1	0.1	0.1
Activation time constant (τ_h)	20ms	20ms	10ms	10ms
Trace time constant ($\tau_{\bar{y}}$)	500ms	500ms	500ms	500ms
Sigmoid threshold percentile	99.2%	98%	88%	95%
Sigmoid slope (β)	190	40	75	26
Filter radius (σ)	7	11	17	25
Filter contrast (δ)	1.5	1.5	1.6	1.4

Table 6.1: Parameters used in the computer simulations.

6.3.4 Numerical Simulation

The coupled ordinary differential equations 6.1, 6.4 and 6.5 cannot be solved analytically. Hence, we used a numerical integration technique known as the Forward-Euler method in order to find discrete approximations to the solution of these equations.

For example, using a Forward Euler scheme the activation would be computed by

$$h_i(t + dt) = h_i(t) - \frac{dt}{\tau} h_i(t) + \frac{dt}{\tau} \sum_j w_{ij}(t) y_j(t) \quad (6.6)$$

Where a new value of the variable $h_i(t + dt)$ is computed at time $t + dt$, where dt is the numerical timestep used in the Forward Euler method. In these simulations the numerical time step was set to a tenth of the neuronal time constant τ_h .

6.3.5 Preprocessing of stimuli for VisNet

Video frames from the scene camera were extracted using VLC media player. These frames were used as the templates for generating the input images presented to VisNet. The frames were converted to monochrome using the MATLAB function `rgb2gray` and resized to a 256x256 matrix to make it fit to VisNet's 'retina'. We have demonstrated in the previous chapter how VisNet can be trained using highly realistic images, in which the hand is seen against natural scenes with multiple objects present at the same time (Galeazzi et al., 2015). However, this requires a training regime in which the network is exposed to a large number of different backgrounds. Therefore, since the focus of the current study is

on testing the performance of the model incorporating natural movements from human subjects, with the exception of the hand and jigsaw piece, the backgrounds of the frames were filled with a 128 grayscale value as per standard VisNet preprocessing procedures. This procedure also filled missing image regions on VisNet's 'retina' arising as the camera images were shifted in line with eye movements recorded from the human subject. How the hypothesised learning mechanism may operate with the hand presented against complete natural background scenes is demonstrated in the previous chapter.

The .csv file generated by the ASL Mobile eye-tracker contained the x and y coordinates of the position of the gaze within the scene frame at a sampling rate of 30Hz. Missing data values were estimated using linear interpolation. Each of the image frames recorded by the scene camera was shifted across VisNet's retina using these coordinates. In this way, the retinal shifts in VisNet effectively matched the corresponding location of the participant's gaze in the visual scene.

As mentioned earlier, in order to prevent VisNet differentiating between different hand-centered locations by exploiting differences in the shapes of the jigsaw pieces, all experimental trials used pieces with the same shape. However, despite this, the illustrations on each jigsaw piece were different. So to prevent VisNet differentiating the positions of pieces based on the illustrations, the internal regions of all the pieces were filled with the same grayscale values. A sample of a preprocessed frame is shown in the lower image of Figure 6.1.

6.3.6 Training and Testing

As previously explained in Section 6.2.1, on each trial the participants would encounter a jigsaw piece near their hand, and they then had to decide where this piece fitted into the puzzle in front of them. There were three possible positions relative to the hand in which the jigsaw piece would appear (i.e. left, right, in front). The participants would typically perform a series of saccades to explore the visual scene while they attempted to solve the puzzle. The images recorded by the scene camera and eye movements recorded by the eye-tracking camera allowed us to reconstruct the retinal images actually seen by the participant. These images were processed as described above and then used to train our neural network model, VisNet.

The training procedure for VisNet consisted of presenting the model with successive processed frames extracted from the scene camera. The images would be centered on VisNet's 'retina' according to the location in the scene at which the participants were fixating. Given the time resolution of the eye-tracker (30Hz), we updated the retinal position of the image every 30 ms. In order to counterbalance the different training trials within each human subject, we adjusted the sequences to the same length. The activations and firing rates of neurons through successive layers of the network are calculated as described above in Section 6.3. The synaptic weights between layers are then updated according to the differential trace rule, see Equation 6.5.

Five independent VisNet simulations were run, where each simulation used data col-

lected from a different one of the five human participants. In each simulation, the processed images from the scene camera were presented shifting on VisNet’s retina, where the retinal shifts were guided by the eye movement data collected from the eye-tracking camera. One training epoch for the network consisted of presenting the retinal images constructed from all trials of the human participant. In each of the five simulations the network was trained over fifty epochs. Table 6.1 contains the parameters used for these simulations.

After the network is fully trained on all the trials of the corresponding human participant, the network is tested to determine whether it has develop localized hand-centered receptive fields. Image sequences of the different spatial configurations of the hand and the puzzle piece are presented over all the retinal locations where the spatial configuration was previously seen. In each simulation, the network was tested with the three spatial configurations of the hand and jigsaw piece corresponding to what the participants saw as they tried to solve the puzzle (i.e. with the jigsaw piece either on the left, or the right, or in front of the hand). Throughout the testing phase, the synaptic weights are unchanged. The neuronal outputs are computed and recorded during the presentation of each spatial configuration in each of the retinal locations.

6.4 Results

We ran five separate simulations using data recorded from each of the five different participants. In each simulation, we analysed the responses of the output (fourth) layer cells in

VisNet before and after the network was trained on the image sequences generated from the scene camera during the jigsaw puzzle task with the human subject. The network was tested before and after training to determine whether cells in the output layer learned to respond selectively to localized hand-centered receptive fields and responded invariantly as these hand-object configurations shifted across the different retinal locations.

Figure 6.2 shows the response profiles of three neurons in the output layer of VisNet before training in one of the simulations. Each of the three columns shows the firing responses of a particular output cell, which is labelled at the top of the column. The three rows of plots show the responses of the cells to each of the three hand-object configurations presented during testing. The top row shows the cell responses when a jigsaw piece is on the right side of the hand, the middle row shows the cell responses when the jigsaw piece is on the left side of the hand, and the bottom row corresponds to the jigsaw piece presented in front of the hand. Each individual subplot shows the responses of the given cell as the particular hand-object configuration is shifted over 100 retinal locations. We can see that before training the three cells respond in an unstructured manner to the different hand-object configurations.

Figure 6.3 shows the response profiles of the same three neurons in the output layer of VisNet after training. We can see that after training, each of the three cells has learned to respond much more selectively to one of the hand-object configurations. Cell 194, 129 (first column) is perfectly selective in that it responds to the same hand-object configuration

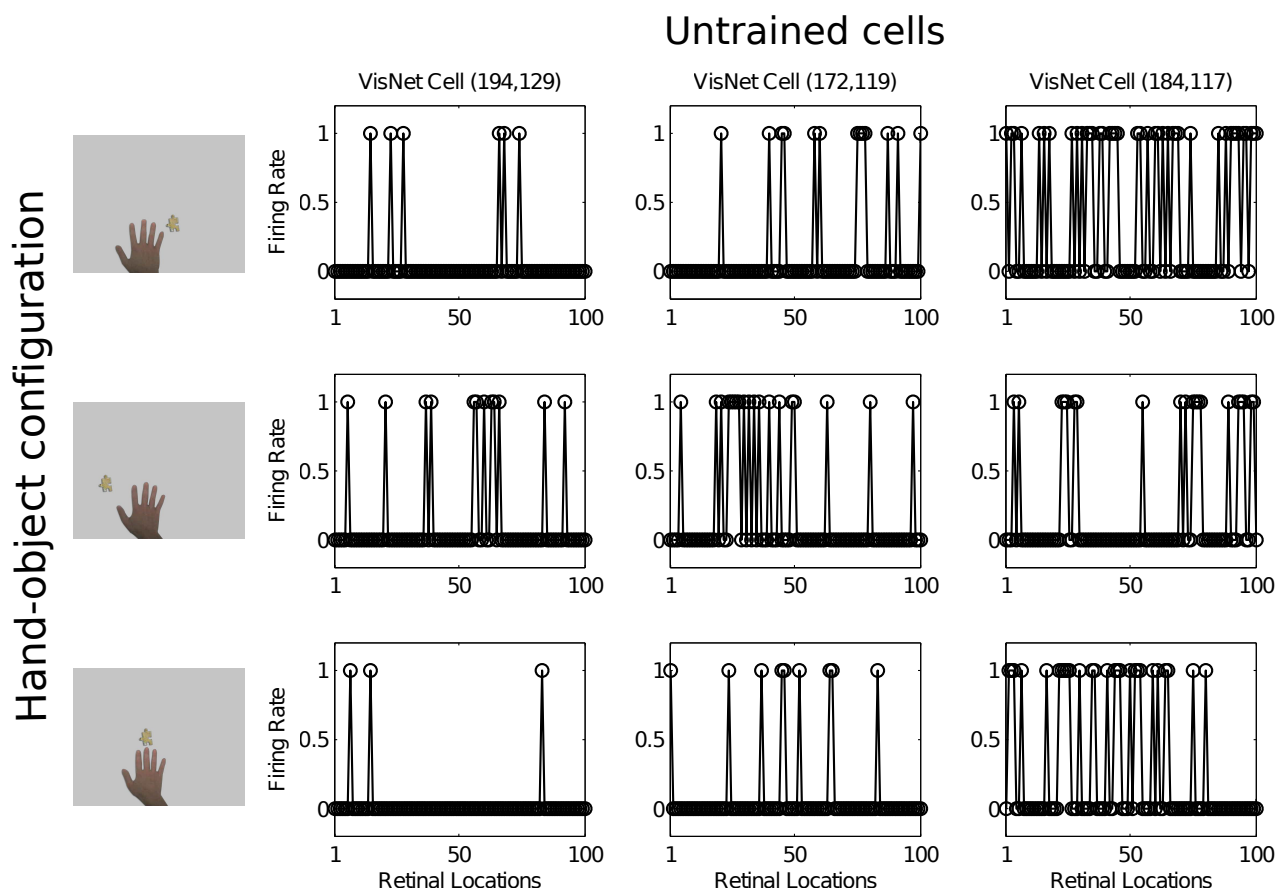


Figure 6.2: Response profiles of three neurons in the top (fourth) layer of VisNet before training. Each of the three columns shows the firing responses of a particular cell. Each row shows the responses of the cells to one of the three hand-object configurations over 100 different retinal locations shown along the abscissae. It can be seen that each of the three cells initially responds randomly to each of the hand-object configurations over the different retinal locations.

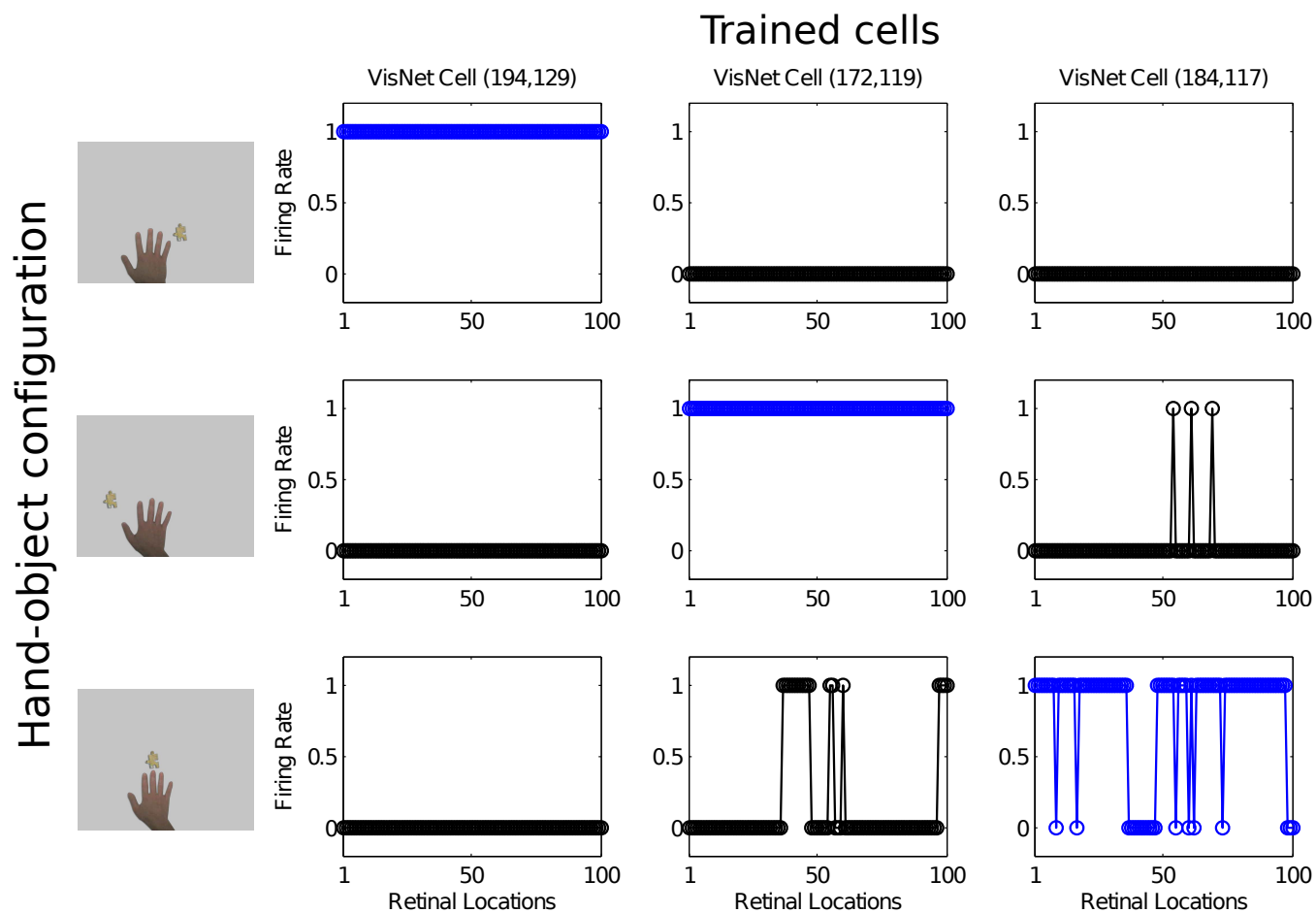


Figure 6.3: Response profiles of same three neurons in the top (fourth) layer of VisNet after training. Response profiles of the same three neurons shown in Figure 6.2 after training. Conventions as in Figure 6.2. It can be seen that each of the three cells responds predominantly to one of the hand-object configurations, and responds to that configuration over most of the 100 tested retinal positions shown along the abscissae.

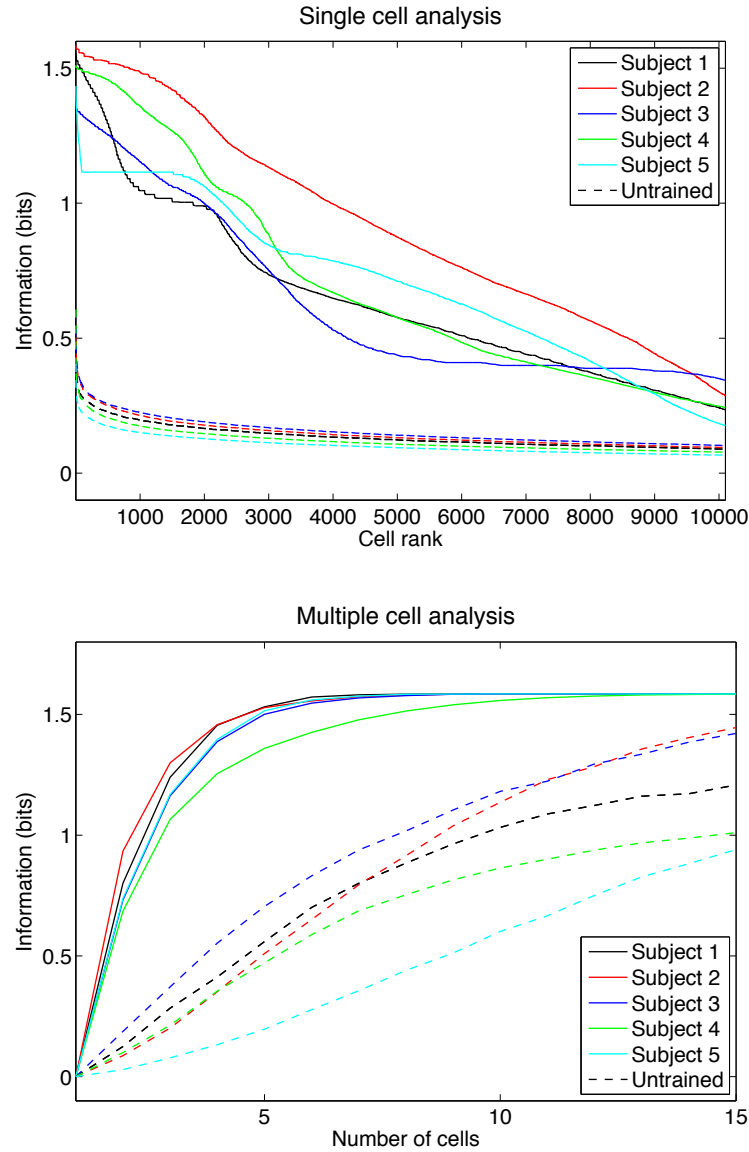


Figure 6.4: Information analysis. Information analysis for each of five VisNet simulations that were performed using data recorded from each of the five different participants. The five different simulations are represented by different coloured plots, where the dashed line shows performance before training and the unbroken line is after training. The upper plot shows the amount of single cell information carried by individual output cells in rank order. It can be seen that after training, the single cell information conveyed by individual neurons increased substantially in all test cases. In the untrained condition no cells reached near maximal information, whereas in the trained condition a portion of cells in each simulation reached high or even maximal information. These trained cells were highly selective to one of the three tested hand-object configurations, and responded to that configuration in most or all of the tested retinal locations. The lower plot shows the multiple cell information measures calculated across 15 cells with maximal single cell information. It can be seen that, after training, the multiple cell information asymptotes to the maximal value of 1.58 bits in all the simulations. This confirms that all three tested hand-object configurations are represented by the output cells.

across all of the tested retinal locations, and never responds to either of the other two hand-object configurations. The next two columns also show highly selective cells, each of which responds predominantly to one particular hand-object configuration across most or all retinal locations, and responds only rarely to any of the other configurations. The results from the other four simulations were similar.

Additionally, as a global measure of performance, we conducted information analysis on the responses of the output cells to all of the test image sequences. Information analysis was conducted separately for each of five VisNet simulations that were performed using data recorded from the five different participants. For each simulation, we studied the responses of the output (fourth) layer cells in VisNet before and after the network was trained on the images of hand-objects configurations generated from the human trials. Figure 6.4 shows the single and multiple information measures for each of the simulations.

The single cell information analysis, (Figure 6.4 top) shows that before training, the output cells conveyed very little information about the three different hand-object configurations (i.e. with the jigsaw piece either on the left, or the right, or in front of the hand). This means that, as was shown in the response profiles of Figure 6.2, the cells respond randomly or in a unstructured way to the different hand-object configurations. After training, we found that in all simulations there is a substantial increase in the amount of single cell information. As can be seen in Figure 6.4, a subset of cells in each simulation reached relatively high levels of information. This means that the trained cells developed

a high selectivity to one of the three hand-object configurations, and responded to that configuration in most or all of the tested retinal locations.

The multiple cell information analysis (Figure 6.4 bottom) shows that, before training, the multiple cell information does not reach the maximal value of 1.58 bits. However, after training we can see that multiple cell information asymptotes to the maximal value in all simulations. This means that all three of the hand-object configurations are successfully represented by separate cells in the output layer. In particular, even when individual cells do not encode the maximal amount of single cell information, the multiple cell analysis confirms that all three hand-object configurations can still be differentiated using just 15 cells that individually carry high levels of single cell information about one of the three configurations.

6.5 Discussion

The simulations presented in this chapter further investigated issues related to the ecological plausibility and viability of our trace learning hypothesis. The trace learning mechanism is able to bind together retinal input images that tend to occur close together in time. This is achieved by incorporating a memory trace of recent neuronal activity into the synaptic learning rule. Our modelling studies in previous chapters assumed that for much of the time humans are continually changing their gaze direction as they explore visual scenes containing fixed spatial configurations of the hand with other objects present. In this case,

trace learning may encourage individual cells to learn to respond to particular hand-object configurations across retinal shifts due to the rapidly changing gaze directions. However, in our earlier modelling studies we simply imposed artificially simulated gaze changes.

Therefore, the goal of the current chapter was to explore the validity of our assumption about the movements of the eyes, head and hands as humans explore visual scenes, and whether trace learning could exploit these natural movements to produce hand-centered visual representations. To do this, we collected relevant data from human volunteers as they solved a jigsaw puzzle. We recorded images of the hand and nearby jigsaw piece from a scene camera and eye movements from an eye-tracking camera, and then combined this data to reconstruct the retinal images seen by human participants as they tried to solve the jigsaw. These retinal images were then used to train our neural network model, VisNet, of the primate visual system.

It was found that these reconstructed retinal images, which arose from natural movements of the eyes and head as humans undertook the jigsaw task, were able to drive the development of neurons with hand-centered receptive fields in the output layer of the network model. In our simulations it was found that after training the network, there was an increase in the amount of single cell and multiple cell information that the output cells carried about the different hand-centered locations. The output cells learned to fire selectively to the hand and a puzzle piece in a fixed hand-object configuration over a large number of retinal views. In fact, the multiple cell analysis showed that in all simulations

a sample of 15 cells was sufficient to differentiate between the three different hand-object spatial configurations. Thus, this work confirms that our neural network model still develops hand-centered visual representations when trained on the retinal images of participants under ecological test conditions.

This is the first time that the flow of visual training images presented to VisNet is guided by actual gaze changes recorded from human participants, instead of imposing idealised image sequences based on theoretical assumptions about the statistics of eye and head movements. These findings are critical to show that the network is able to self-organise hand-centered visual representations successfully with retinal images arising from realistic movements of the eyes, head and hand. Our results suggest that despite the variability of the visual search strategies and gaze changes of individual participants, trace learning is robust enough to allow the network to form neurons with hand-centered receptive fields that are invariant to shifts in retinal location, as observed in neurophysiology studies (Bremner and Andersen, 2012; Buneo and Andersen, 2006).

Despite the significant improvements in terms of the ecological plausibility of the trace learning mechanism for the development of hand-centred representations, we have not yet addressed the fact that when visual localization of the hand is absent, the proprioceptive and motor signals provide a hand location signal that is computed in a body-centred reference frame, which would then require that the visual targets are also mapped into body-centred coordinates in order to derive the reach vector in the same coordinate frame

(Buneo and Andersen, 2012; Buneo et al., 2009). In the following chapter we will explore how visual targets can be mapped into body-centred coordinates using signals representing the position of the eyes and head.

Our brain is an organ of action that is directed toward practical tasks; it does not appear to have been built for discovering the ultimate causes of things, but rather for determining their immediate causes and invariant relationships.

—Santiago Ramón y Cajal, *Advice for a Young Investigator*

7

The potential role of proprioception in the development of hand-centred representations

The previous chapters explored the learning mechanisms that underlie the development of visual hand-centred receptive fields using a self-organising neural network model of the primate visual system. However, in these simulations the proprioceptive signals representing the position of the hand with respect to the trunk were not explicitly modelled. In this chapter we begin by presenting a computational framework that explores the potential role of proprioception in the development of hand-centred representations. We hypothesize that hand-centred visual representations may develop by combining proprioceptive signals representing the position of the hand with respect to the body with visual representations of the position of a target using a common frame of reference. In Section 7.2 we present a self-organising model showing how trace learning can drive the development of visual rep-

representations of the position of a visual target with respect to the body (trunk). In Section 7.3, we integrate these body-centred representations of the visual targets into a complete model architecture that develops hand-centred visual representations by combining them with proprioceptive signals representing the position of the hand with respect to the body.

7.1 Introduction: Overarching network architecture of hand-centred representations

As previously discussed in this thesis, the main differences between the ventral and dorsal visual pathways can be better explained in terms of the purpose of their outputs rather than differences in the processing of the visual inputs (Goodale and Milner, 1992). A detailed representation of objects for perceptual or identification purposes (i.e. 'vision for perception') does not require that the representation of the location of the perceived object is in a particular egocentric frame of reference, but rather the development of invariant representations that encode the relevant features of the objects and their spatial relation, which seems to be best represented using an allocentric or object-centred framework (Milner and Goodale, 2008). However, unlike the ventral visual pathway, the dorsal stream seems to represent the visual objects using a variety of body-centred and body-part centred frames of reference. It has been hypothesized that the main reason for transforming the initial retinocentric representation into a coordinate systems centred on different parts of the body, is to provide a representation that is more suitable to the effector of the planned

movement (i.e. 'vision for action'). Evidence from spatial neglect patients seems to be consistent with the notion of predominantly object-centred or allocentric frame of reference encoding in the ventral stream (Driver and Halligan, 1991; Humphreys et al., 2013), while egocentric deficits seem to be more correlated with injuries in the dorsal pathway (Karnath and Rorden, 2012).

The different reference frames in which visual information is encoded in each of the pathways seem to be related to the purpose that these visual representations serve. However, despite these differences, both visual pathways seem to be performing analogous computations. The ventral visual pathway has a hierarchical organization in which invariant representations of the visual objects are formed in the latter stages of processing. Similarly, the dorsal pathway organizes its information in body-centred and body-part centred coordinate systems that seem to be invariant to changes in posture (Spratling, 2009). If this is the case, the following question arises, why are these body-part centered representations formed predominantly within the dorsal pathway while object-centred and allocentric representations seem to appear more predominantly in the ventral pathway? I hypothesize that the nature of the postural proprioceptive signals entering the dorsal pathway may have a key role in differentiating the frame of reference in which visual information is represented along this pathway. These incoming proprioceptive signals representing the positions of different parts of the body will have an effect on the underlying computations and will drive the development of the visual representations of objects into

a variety of predominantly body-centred and body-part centred coordinate systems. In the previous chapters I have shown simulations in which the location of the hand was represented visually. Despite the fact that the proprioceptive location of the hand was not explicitly modelled, we hypothesized that in the brain, the incoming proprioceptive and motor efferent signals providing information about the location of the hand in space would be playing a key role in guiding the development of hand-centred visual representations along the visuomotor pathway. Consistent with this notion, there is evidence suggesting that the sight of objects near the hand biases visual processing towards the dorsal stream (Gozli et al., 2012). In Section 5.9, we showed an example of how incorporating a constant representation of the hand within a variety of visual scenes can affect the selectivity of the neuronal responses and increase the number of neurons representing a localised region within a hand-centred receptive field.

It is well known that the posterior parietal cortex (PPC) and premotor areas not only receive bottom-up visual signals (Caminiti et al., 1996; Galletti et al., 1996), but they also receive afferent signals from the somatosensory cortex and top-down motor signals (Prevosto et al., 2011; Sakata et al., 1973). Experimental studies have shown that hand-centred cells seem to maintain their firing properties when guided only by motor signals or proprioceptive signals in the absence of a visual cue of the hand (Buneo and Andersen, 2012; Graziano, 1999; Graziano et al., 1994b). This suggests that the hand-centred representations receive additional proprioceptive input signals specifying the position of the hand.

These internal representations of the hand position derived from proprioceptors could be used to update hand-centred representations in the PPC in the absence of a visual input from the hand. In this chapter we will present an overarching architecture in which we explore the potential role of a proprioceptive representation of the position of the hand in the development of hand-centred representations of the visual target.

7.1.1 Overarching network architecture of hand-centred representations

The overarching architecture is composed of five populations of neurons. Figure 7.1 shows the different components and their connectivity. One of the input layers provides a proprioceptive signal representing the position of the hand with respect to the trunk. The second input layer provides a visual signal representing the location of the visual targets with respect to the trunk. These two input layers have feed-forward synaptic connections projecting to a population of combination neurons in the combination layer. We propose that the inputs from the proprioceptive signals representing the position of the hand and the inputs from the visual representations of the position of a target are combined in this layer using a common body-centred frame of reference. This combination layer of neurons was modelled as a self-organising competitive layer. The competitive interaction between these neurons would encourage individual cells to learn to fire exclusively to particular combinations of hand position and target position in a body-centred reference frame.

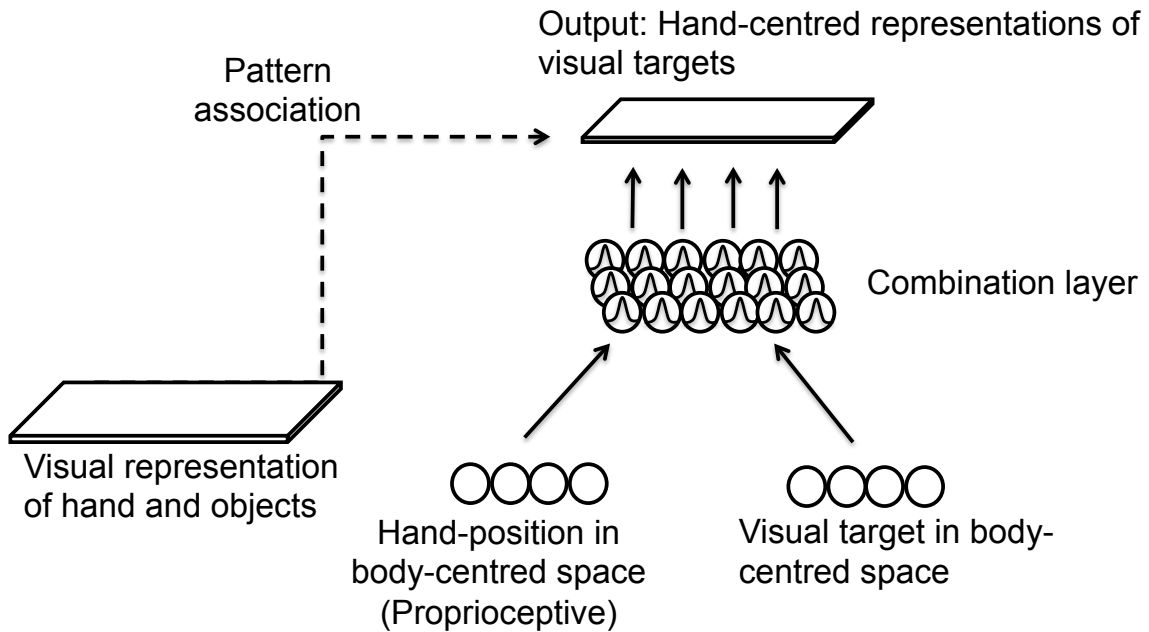


Figure 7.1: Overarching architecture of the neural network model incorporating proprioceptive signals. The overarching network presented in this chapter provides a theoretical account of how proprioceptive input signals representing the position of the hand with respect to the body can guide the development of hand-centred visual output neurons. This network has an input layer that receives a proprioceptive representation of the position of the hand in body-centred coordinates. Another input layer represents the location of the visual target in body-centred coordinates. Both of these layers have feed-forward synaptic connections to a combination layer. This combination population learns the different combinations of hand and visual target positions, and projects its outputs to a final output layer that learns to develop hand-centred representations of target location. Moreover, the final output layer can also receive parallel signals from visual representations of the hand and background target objects as used in the previous chapters. The visual representation of a hand and target objects can be associated with hand-centred output representations derived from proprioceptive inputs by pattern association learning due to their frequent and congruent co-activation. After learning, this allows the network to respond to a hand-centred configuration even when a visual cue of the hand is not available.

The combination layer has feed-forward synaptic connections projecting to the final output layer. Particular combinations of hand-position and target position corresponding to a specific hand-centred configuration represented by combination neurons could be then bound together across different body-centred locations by the output neurons using an invariance learning mechanism such as trace learning. This would allow the output layer to develop selective hand-centred responses across different locations of the hand with respect to the body, as long as the target and the hand are shown in the same hand-centred configuration.

It has been reported that some of the hand-centred cells can respond to a visual stimulus near the hand, even if a visual cue from the hand is absent (Graziano, 1999; Graziano et al., 1994b). In this case the four components shown on the right side of the network in Figure 7.1 will be able to represent a hand-centred configuration driven by the proprioceptive representation of the location of the hand and the visual representation of the target in body-centred space. However, in much of our sensory experience we will also have available a visual representation of the hand and surrounding objects. Furthermore, the visual and proprioceptive representations of the location of the hand will be spatially congruent. Thus, the fifth component of the network incorporates the visual representation of the hand and background objects as used in the previous chapters. This pathway will feed directly into the output layer forming a typical pattern association network. In this case, the sight of a hand-centred configuration in the left hand pathway of the network will be associated with

the activation of a hand-centred representation in the output layer, which is itself derived from a combination of proprioceptive signals and body-centred visual target locations.

In Section 7.2, I will begin by presenting a self-organising hypothesis showing how trace learning could drive the development of visual representations of the position of a visual target with respect to the body (trunk) based on the statistics of natural eye and head movements. In Section 7.3 we will show how these body-centred visual representations could be incorporated into the overarching model architecture described in Figure 7.1. We will demonstrate an example of how combining proprioceptive signals representing the position of the hand with respect to the body with visual representations of the position of a target also in a body-centred frame of reference could drive the development of hand-centred responses in the output layer of the model. A more detailed description of the operation of the network, as well as the parameters and training regimes used for these simulations will be provided in each subsection.

7.2 The development of visual neurons representing the locations of visual objects in a body (trunk) centred frame of reference

In order to represent the position of the target with respect to the hand in our overarching model, the visual representation of the targets and the localization of the hand need to be

represented within a common frame of reference. In this subsection of the chapter we will present a self-organising hypothesis on how body-centred representations of the position of a visual target could be developed using a trace learning mechanism. These body-centred responses form an important part of our overarching architecture (See Figure 7.2), allowing the combination of hand-positions and target positions to be formed.

7.2.1 Body-centred representations of visual targets in the brain

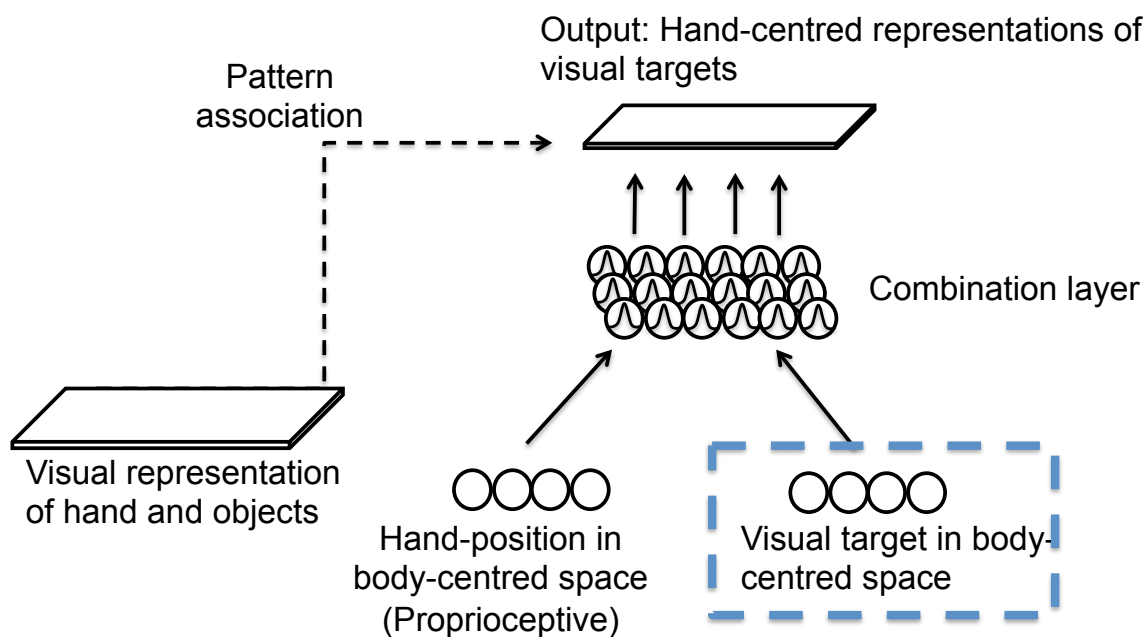


Figure 7.2: Representing the location of the visual target in body-centred coordinates. The simulations presented in this section provide a theoretical account for the development of cells that learn to represent the location of a visual target in a body-centred frame of reference.

As we have discussed earlier in this thesis, the location of visual targets seems to be gradually transformed from eye-centred to a variety of body-part centred representations

(Flanders et al., 1992; Soechting and Flanders, 1992). Neurons in the posterior parietal cortex and adjacent areas seem to combine visual information with postural signals representing the location of the eyes, head and limbs (Snyder et al., 1998) to develop a variety of non-retinal egocentric representations. In single cell recordings, we generally find a gradient of responses from predominantly eye-centred representations in early visuomotor areas to different non-retinal egocentric representations in more parietofrontal areas (e.g. Head-centred (Duhamel et al., 1997; Galletti et al., 1993), body-centred (Brotschie et al., 1995; Lacquaniti et al., 1995) and hand-centred representations (Bremner and Andersen, 2012)).

In this chapter a self-organising hypothesis of how the brain could be forming representations of visual targets in a body-centred or trunk centred frame of reference is presented. Moreover, the locations of the limbs are also thought to be mostly represented in a shoulder-centred or body-centred encoding scheme (Lacquaniti et al., 1995). In our overarching model the body-centred representations of the visual target will be later combined with a body-centred representation of the location of the hand in order to develop hand-centred responses. Furthermore, a common body-centred representation of the location of the visual target and the location of the hand can play an important role in motor planning, by allowing a direct computation of a reach vector in a common body-centred frame of reference (Buneo and Andersen, 2006; Buneo et al., 2009).

Different theoretical efforts to understand the process of reference frame transformations

(Pouget and Sejnowski, 1997; Zipser and Andersen, 1988), have been discussed in this thesis. However, most of these models make use of a back-propagation algorithm (Werbos, 1974) in which the error correction 'delta-rule' (Widrow et al., 1960) is implemented. These approaches have been regarded as biologically implausible (Rolls, 2008) and therefore cannot provide an account of how these properties may develop in the cortex. More recently, a few self-organising approaches to the development of head-centred and body-centred representations have been put forward. For example, Spratling (2009) explored a temporal binding hypothesis for developing head-centred and body-centred cells. In this model a hierarchy of two separate conjunctive and disjunctive layers are used to develop head-centred and body-centred responses. In the first stage of processing, head-centred output cells are developed from retinocentric representations and gain modulated signals representing the position of the eyes. The model then uses this head-centred output to learn combinations of different head-centred locations to develop output cells in a body-centred reference frame. More recently, colleagues from our research centre developed a self-organising learning hypothesis for the development of head-centered visual representations (Mender and Stringer, 2013) incorporating eye position gain fields in the input layer and using a trace learning mechanism. In this chapter we extend this work to present a new model in which body-centred representations can be directly developed from an input layer that incorporates an additional gain field representing the position of the head with respect to the trunk. Thus, in this model it is shown that development of explicit head-centred

responses is not necessarily a prerequisite to develop body-centred representations. As long as the input layer contains receptive fields that multiplicatively combine visual, eye position and head position signals, we can directly develop cells in the output layer that learn to encode the position of visual targets in a body-centred frame of reference using the statistics of eye and head movements during training.

7.2.2 Hypothesis

It was hypothesised that the following four core model components would permit neurons to learn to respond to visual objects in a body (trunk) centred reference frame through a biologically plausible process of visually guided learning:

- There is a population of input neurons that encode a combination of the retinotopic location of visual targets, eye position and head position at the single neuron level through receptive fields that multiplicatively combine visual, eye position and head position signals.
- There is a population of output neurons that compete with each other through mutual inhibitory interactions mediated by inhibitory interneurons.
- The feedforward synaptic connections between the input and output neurons are modified by a local synaptic trace learning rule that encourages individual output neurons to learn to respond to subsets of input patterns that tend to occur close together in time.

- During natural self-motion, there are periods of time when either the eyes are moving in the head or the head is moving on the body (trunk) or both the eyes and head are moving together, while the body remains stationary with respect to the visual environment and visual objects also remain stationary within the environment.

The core self-organisational principle is as follows. It is assumed that there will be periods of time when either the eyes are moving in the head or the head is moving on the body (trunk) or both the eyes and head are moving together, while the body remains stationary with respect to a static environment. Typical movement sequences seen in humans involve the head remaining stationary while the eyes move rapidly around a visual scene, and then the head is readjusted to a different position in which the eyes then perform further saccades. During these sequences of eye and head movements the body may remain stationary. With a few variations contingent on the tasks being executed, it is estimated that on average humans make more than three saccades per second (O'Regan, 1992). Thus, as we move our eyes and head to explore a visual scene, a stationary object will appear in different positions across the retina while remaining in the same position with respect to the body (i.e. trunk). As we have discussed previously, trace learning encourages neurons to learn to respond to inputs that tend to occur close together in time by incorporating a memory trace of the recent neuronal activity in the learning rule (Földiák, 1991). In this case, the trace learning rule will encourage individual output neurons to bind together sequences of inputs representing combinations of retinotopic visual target

location, eye position and head position corresponding to a particular body-centred target location within the environment. This will produce output neurons that have learned to respond to the presence of a visual target in a specific location with respect to the body. When an object appears in a new position with respect to the trunk or the trunk is itself moved to a different location within the environment, further combinations of eye and head movements will continue and the new body-centred target position will be represented by a different subset of output neurons. Hence, this learning process should allow the network to develop output cells that learn to respond to the position of objects in body-centred coordinates, irrespective of the position of the eyes and the head.

7.2.3 Model Architecture

The architecture of the neural network model consists of two layers of artificial neurons. The input layer of the network was composed of neurons that represented a multiplicative combination of the location of the visual target in retinal coordinates, as well as the position of the eyes with respect to the head and the position of the head with respect to the body (i.e. gaze configuration). In other words these input neurons were modelled as retinotopic cells with peaked gain fields representing preferred combinations of eye and head positions.

The retinal spaces represented by the input layer were within the range of $[-100^\circ, 100^\circ]$, the eye position was represented within the range of $[-20^\circ, 20^\circ]$ and the head position space was also represented within the range of $[-20^\circ, 20^\circ]$.

For this simulation we modelled the response functions of the input neurons using peaked gain fields of the position of the eyes and the position of the head. The following equation describes the firing rate function of the input layer neurons.

$$y_i^I = \exp\left(-\frac{(h - \lambda_i)^2}{2\rho^2}\right) \times \exp\left(-\frac{(e - \mu_i)^2}{2\varrho^2}\right) \times \exp\left(-\frac{(r - \nu_i)^2}{2\sigma^2}\right) \quad (7.1)$$

The response function presented in Equation 7.1 computes the product of three components. The first component represents the positional signal of the head, the second component represents the positional signal of the eyes, and the third component represents the position of the visual target in retinocentric coordinates.

This function represents the retinotopic position of a visual target that is gain modulated by a peaked Gaussian function of eye position and a peaked Gaussian function of head position. λ_i represents the head position at which the input neuron i responds maximally, where the standard deviation ρ determines the width of the Gaussian head position tuning curve. Similarly, μ_i represents the eye position at which the input neuron i responds maximally, while the standard deviation ϱ determines the width of the Gaussian eye position tuning curve. Lastly, ν_i represents the position on the retina at which the input neuron i responds maximally when a stimulus is presented there, while the width of the Gaussian retinal preference tuning curve is determined by the standard deviation σ . This response function allows the neurons to respond maximally to particular combinations of head position, eye position, and retinal position of a visual target.

The neurons in the input layer project feedforward synaptic connections to the output layer. The output layer was modelled as a competitive network which incorporates lateral competition between neurons (Rolls and Treves, 1998). The feedforward synaptic connectivity in the network was diluted, although each of the output neurons received the same number of connections from the input layer. The parameter ϕ determined the percentage of neurons in the input layer from which each output layer neuron received connections. The input and output layers, as well as the connections between them, did not have a topographical organization.

As in the previous chapters using a discrete representation of time, the activation h_i of each output neuron i was computed by

$$h_i = \sum_j w_{ij} y_j \quad (7.2)$$

where y_j is the firing rate of input neuron j , and w_{ij} is the strength of the synapse from neuron j to neuron i .

Lateral competition between output neurons was effected by means of an adjustable sigmoid activation function

$$y = f^{sigmoid}(h) = \frac{1}{1 + e^{-2\beta(h-\alpha)}} \quad (7.3)$$

where h is the cell activation, y is the firing rate after lateral competition, and α and

β are the sigmoid threshold and slope respectively. The parameters α and β are constant across all neurons within the output layer, although α is adjusted at each timestep to control the sparseness of the firing rates. For the simplified case of neurons with binarised firing rates = 0/1, the sparseness is the proportion $\in [0, 1]$ of neurons that are active. For example, to set the sparseness to 0.2, the threshold is set to the value of the 80th percentile point of the activations within the layer. The parameters used for these simulations are shown in Table 7.1.

In this study the model also implements *trace learning* in its synaptic connections, in which a temporal trace of the previous activity of the neuron is incorporated in the learning rule. The trace learning rule has been described in previous chapters (see Equation 4.2). It has been previously shown how trace learning may be used to bind subsets of input patterns that occur close together in time. In the simulations described below, there are sequences of eye and head movements during which the body remains stationary. In this case, trace learning should encourage output neurons to respond to combinations of retinotopic visual target location, eye position and head position corresponding to a particular body-centred target locations. Although, it should be noted that trace learning is robust enough to allow a competitive network to develop non-retinal egocentric representations even if the training regime is composed of a wide variety of stimulus dynamics, as long as the stimulus dynamics required for self-organization occur for a small portion of the time (Mender and Stringer, 2014).

Parameter	Symbol	Value
Number of target locations	-	6
Number of training epochs	-	50
Width of head position tuning curve	ρ	6°
Width of eye position tuning curve	ϱ	6°
Width of retinal tuning curve	σ	6°
Output neuron population size	-	100
Input neuron population size	-	44541
Trace value	η	0.8
Activation function slope	β	6.5
Activation function threshold	α	0.5
Sparseness percentile	-	80%
Synaptic connectivity	ϕ	5%

Table 7.1: Parameters of self-organising model of body-centred visual representations.

The synaptic weights were initialised randomly within the interval of $[0, 1]$. To restrict and limit the growth of each neuron’s synaptic weight vector, we also normalise its length at the end of each timestep during training as is usual in competitive learning (Rolls, 2008; Rolls and Treves, 1998).

7.2.4 Analysis

In order to examine the performance of the network we adapted some of the metrics proposed by Mender and Stringer (2013, 2014) for analysing the different frames of reference represented by the output neurons before and after training. For this study we included an assessment of the responses of the output neurons in body-centred coordinates.

In this case $\mathbf{R}$ will denote the matrix that contains the responses of an output neuron

during the testing phase. Specifically, $\mathbf{R}[i, j]$ will refer to the firing rate when the model is in the i^{th} gaze configuration g_i (i.e. eye and head configuration) and the visual target is present in the j^{th} body-centered location t_j . G refers to the number of gaze configurations during testing and T refers to the number of body-centered locations in which the visual targets appear during testing. Therefore the response vector when we test for body-centredness ($\mathbf{R}[i, 1], \dots, \mathbf{R}[i, T]$) refers to the responses of the network at the i^{th} gaze configuration. The gaze configurations and body-centered target locations are indexed ordered from left (negative) to right (positive).

Reference Frames

In this study we tested the responses of the output neurons in three possible frames of reference (i.e. Eye-centred, Head-centred and Body-centred). In order to quantify the reference frame in which the cell response can be better described we applied three different metrics, one for each of the three possible reference frames.

For example, an output neuron is considered to encode the position of a target in body-centred coordinates if it fires to the presence of a visual target in a particular position with respect to the trunk, irrespective of where the eye and head are positioned. In this case the body-centredness metric will be measuring the average correlation of the response vectors of an output neuron in all the different gaze configurations in which it has been tested.

The *body-centredness* metric was computed using the following equation

$$\Pi = \frac{1}{\binom{G}{2}} \sum_{1 \leq i_1 < i_2 \leq G} \frac{\sum_{j=1}^T (\mathbf{R}[i_1, j] - \bar{\mathbf{R}}[i_1]) (\mathbf{R}[i_2, j] - \bar{\mathbf{R}}[i_2])}{\sqrt{\sum_{j=1}^T (\mathbf{R}[i_1, j] - \bar{\mathbf{R}}[i_1])^2 \sum_{j=1}^T (\mathbf{R}[i_2, j] - \bar{\mathbf{R}}[i_2])^2}} \quad (7.4)$$

where

$$\bar{\mathbf{R}}[i] = \frac{1}{T} \sum_{j=1}^T \mathbf{R}[i, j] \quad (7.5)$$

As per standard correlation measures, the value could be between -1 and 1 . The maximal positive value would reflect a perfect correlation of the response vectors in body-centred coordinates.

Similar metrics were computed to quantify the degree of correlations of the responses of individual output neurons in the other frames of reference. For example during testing, the network was also presented with visual targets that followed the different positions of the head. The response vectors of this testing regime were used to compute a measure of head-centredness. Lastly, the model was tested presenting the visual targets following the movement of the eyes. This would mean that the response vector corresponded to the same eye-centred coordinates. The same correlation analysis was performed to compute the eye-centredness metric of individual neurons. Only responsive cells were included in this analysis.

Receptive Field Location

In order to determine the location of the body-centred receptive field of a particular output neuron, we computed the centre of mass of the body-centred response vector at each of the gaze configurations $i = 1, \dots, G$. We then calculated the average of the body-centred locations across all the gaze configurations using the following equation

$$\frac{1}{G} \sum_{i=1}^G \frac{\sum_{j=1}^T t_j \mathbf{R}[i, j]}{\sum_{j=1}^T \mathbf{R}[i, j]} \quad (7.6)$$

This averaging procedure has been previously shown to be optimal (Mender and Stringer, 2013, 2014).

Receptive Field Size

In order to determine the size of the body-centred receptive field of a particular output neuron we first had to isolate the responsive regions of that cell. In the following analysis, a neuron was considered responsive when it was firing above a threshold rate C , which was set to be half the maximal firing rate of the neuron across all gaze configurations. For each gaze configuration, this threshold was used to isolate the responsive body-centered regions of the given neuron. The total size of all such regions was added up, estimating the total receptive field size at the given gaze configuration. The final receptive field size of the given neuron was the average of all these estimates from the different gaze configurations.

In order to isolate the responsive regions of a neuron for the i th gaze configuration, we computed a body-centred piecewise linear response function $R_i : [-60^\circ, 60^\circ] \rightarrow [0, 1]$ from the response vector using linear interpolation. By finding the solutions x^* to the equation $C = R_i(x^*)$, we could then verify in which regions of the body-centred space, the neuronal responses that were interpolated passed the threshold C . In our simulations there was generally only a single, clearly identifiable responsive region. The sizes of these responsive regions were then averaged across the different gaze configurations i . On a few occasions, a body centered neuron did not actually respond in one of the gaze configurations.

7.2.5 Training the Network

The training procedure for these experiments consisted of presenting the model with input patterns that represented a combination of the position of the eyes with respect to the head, the position of the head with respect to the body (trunk), and the retinotopic location of a visual object. In particular, our trace learning hypothesis regarding the development of body-centred neurons requires that the network is presented with a series of training sequences in which the eyes and head move to effect changes in gaze direction, while the visual object remains static with respect to the body. Furthermore, experimental studies have previously shown that the eyes move more frequently than the head (Einhäuser et al., 2007; Freedman and Sparks, 1997). Therefore, in our model simulations, the eyes were simulated performing sequences of several rapid saccades, where each saccade sequence

was followed by a single postural adjustment of head position. After a series of such eye and head movements, the visual object would then be moved to a new body-centred location, and the whole training process was then repeated. This procedure ensured that the network was trained with sequences of eye and head movements during which the visual object remains static with respect to the body, as required by our hypothesis.

The visual objects were presented at six different body-centred locations during training $[-56^\circ, -34^\circ, -26^\circ, 12^\circ, 42^\circ$ and $60^\circ]$; these objects were presented to the network one at a time. The position of the eyes was moved randomly anywhere within the range $[-20^\circ, 20^\circ]$. While the position of the head with respect to the trunk was moved between four pre-specified positions $[-12^\circ, -4^\circ, 4^\circ, 12^\circ]$. The respective ranges over which the eyes, head and visual object moved guaranteed that the object always stayed in sight.

The firing rates of individual neurons within the input layer were calculated for every new combination of eye and head position (i.e. gaze change) and visual target location with respect to the body. Next the activations and firing rates of the output neurons are computed. Subsequently, the synaptic weights between the input and output layers are updated according to the trace rule as in previous chapters. Trace learning in this case helps the network to bind together different combinations of head position and eye position (i.e. gaze configuration), as well as retinotopic location of the visual target, that occur close together in time. This encourages output neurons to learn to represent to the location of a visual target irrespective of where the head and the eyes are positioned, effectively forming

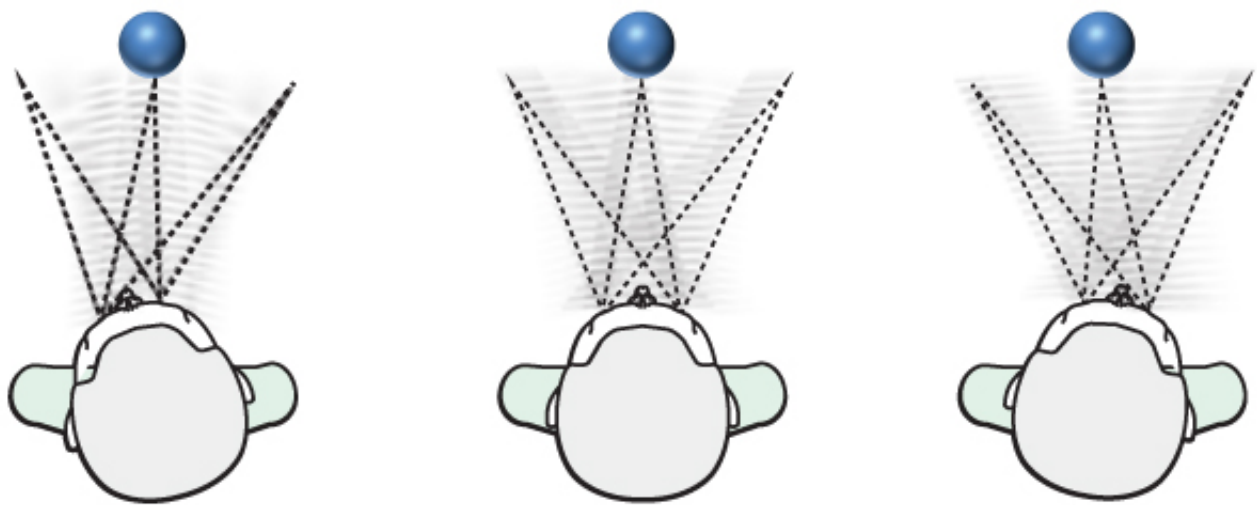


Figure 7.3: Schematic examples of eye and head movements used during training in simulations of the development of body-centred visual neurons. During training we presented a static visual target in the same body-centred location while the network explored the visual scene using different combinations of head and eye movements. Experimental studies have shown that primates move their eyes more frequently than their head. Therefore, in the model simulations carried out here, the eyes performed sequences of rapid saccades followed by occasional head-movements. The figure shows three different head positions with respect to the trunk. In each head position a series of eye movements are performed. During these eye and head movements, the object remains stationary with respect to the body. The details of the training regime are described in section 7.2.5

a body-centred representation of the visual target. The output layer is implemented as a competitive network; the competitive interaction between output neurons allows the network to represent the different body-centred target locations in distinct sub-populations of neurons within the output layer.

For this experiment, one training epoch consisted of the presentation of a visual object in six body-centred locations, where each of these object locations was seen in a total of 84 different gaze configurations (i.e. 4 head positions and 21 eye positions per head position). Figure 7.2.5, provides an illustration of how different combinations of eye and head positions can occur while an object remains in the same body-centred position. Our network was trained over fifty epochs in total. The learning rate used was 0.1. The rest of the training parameters can be found in Table 7.1.

7.2.6 Testing the Network

Throughout the testing phase, the synaptic weights were not changed.

In order to test for body-centred representations, the network was tested by presenting a visual target across 121 different body-centred locations covering the range $[-60^\circ, 60^\circ]$ in one degree increments, while the eyes and head remain fixed in a particular gaze configuration. The firing rates of the output neurons were computed and recorded during the presentation of the visual target in each of the tested body-centred locations. This was repeated for four different gaze configurations (Head -12° , Eye 10°), (Head 4° , Eye -12°),

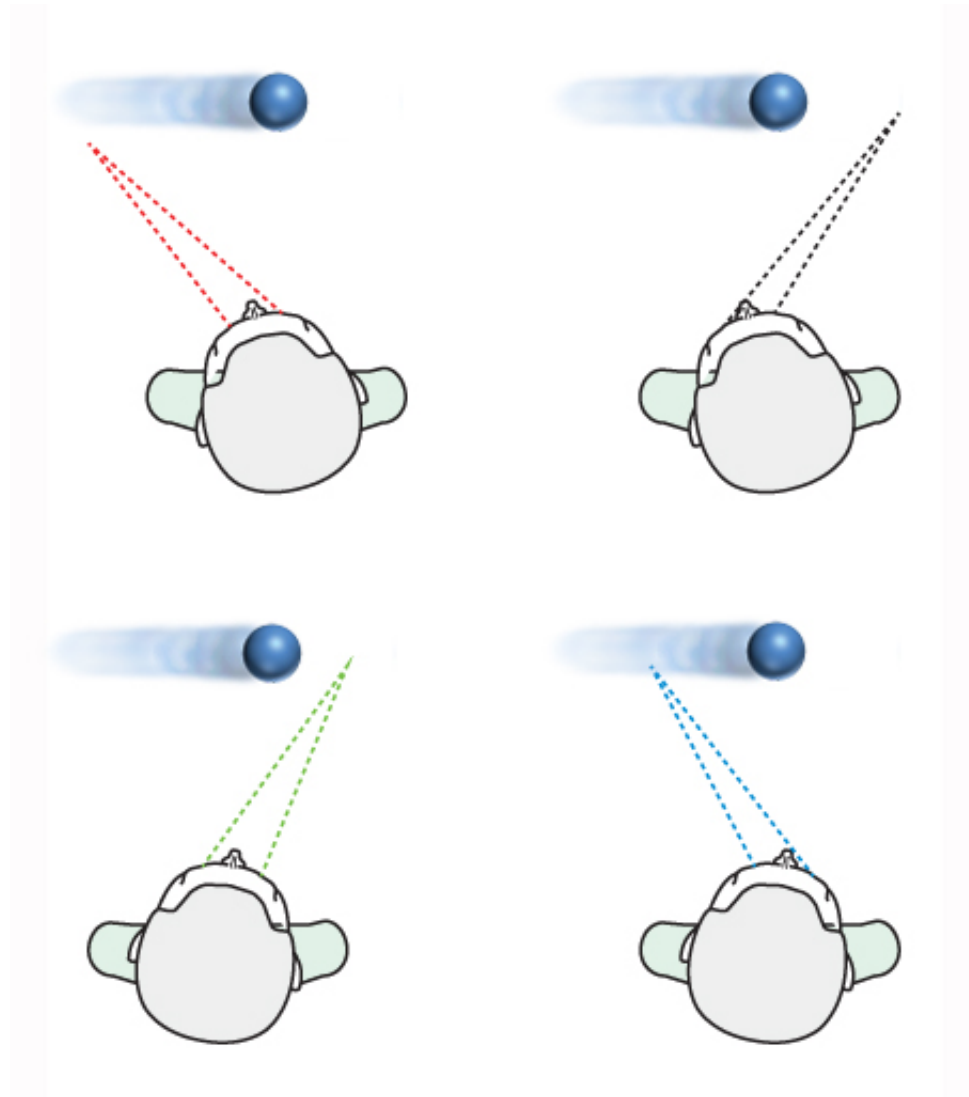


Figure 7.4: Schematic example of four gaze configurations used during testing in simulations of the development of body-centred visual neurons. The figure shows four different gaze configurations. Each gaze configuration involves a different combination of position of the eyes with respect to the head and position of the head with respect to the body (trunk). During testing, a static visual target is presented in 121 different body-centred locations covering the range $[-60^\circ, 60^\circ]$ in one degree increments, while the head and eyes remained fixed in a particular gaze configuration. This was repeated for four different gaze configurations (Head -12° , Eye 10°), (Head 4° , Eye -12°), (Head -4° , Eye -8°) and (Head 12° , Eye 4°). An output cell with a body-centred receptive field will fire maximally to the same body-centred location of the visual target across the four different gaze configurations. The details of the testing regime are described in section 7.2.6

(Head -4° , Eye -8°) and (Head 12° , Eye 4°). The response vectors collected from this testing regime were then used to calculate the corresponding reference frame metrics as well as to analyse the properties of the body-centred receptive fields that had developed. Figure 7.4 provides a schematic example of how the testing procedure is performed.

In order to test whether the output neurons had in fact developed response properties that were more head-centred, the network was also presented with visual targets in different fixed positions with respect to the head. The response vectors of this testing regime were used to compute a measure of head-centredness. Additionally, the model was tested by presenting the visual targets following the movement of the eyes. In this response vector, the locations of the visual targets throughout the different gaze configurations were exactly the same in eye-centred coordinates. This response vector was then used to compute the measure of eye-centredness of individual neurons.

7.2.7 Results

This experiment investigated how the self-organising competitive network model described in Section 7.2.3 may develop body-centred representations of the locations of visual targets through the hypothesised trace learning mechanism described in Section 7.2.2.

To test the network, a single visual target was presented in different positions with respect to the body under different gaze conditions corresponding to different combinations of eye position and head position. Subsequently, the firing responses of the output cells

before and after training were analysed. Figure 7.5 shows the response profiles of the same output cell before and after training. Each of the curves within each plot corresponds to four different gaze configurations, while a visual target is presented in each of the tested body-centred locations within a range of $[-60^\circ, 60^\circ]$. Figure 7.5 top, shows the responses of one the output neurons before training. We can see that this cell seems to be almost unresponsive to the objects presented in different positions across the body-centred space. However after training the network for 50 epochs it can be seen (Figure 7.5 bottom) that the same output neuron responds maximally in approximately the same body-centred position across all of the tested gaze configurations.

This output cell had a body-centredness correlation coefficient of 0.07 before training and 0.86 after training. These results confirmed that training led to the output neuron developing highly body-centred responses. The body-centred receptive field of this output cell after training is located at 10.23° with a size of 5.48° . The responses of the trained output neuron were not so highly correlated with the other two egocentric reference frames. Before training the cell had an eye-centredness correlation coefficient of 0.08 and a head-centredness coefficient of 0.31. After training, the eye-centredness correlation coefficient was 0.29 and the head-centredness coefficient 0.32. Thus, the body-centredness measure of the trained neuron was much greater than either its eye-centredness or head-centredness. Overall these results show that after training, the output cell started to represent the location of the visual target in a body (trunk) centred frame of reference.

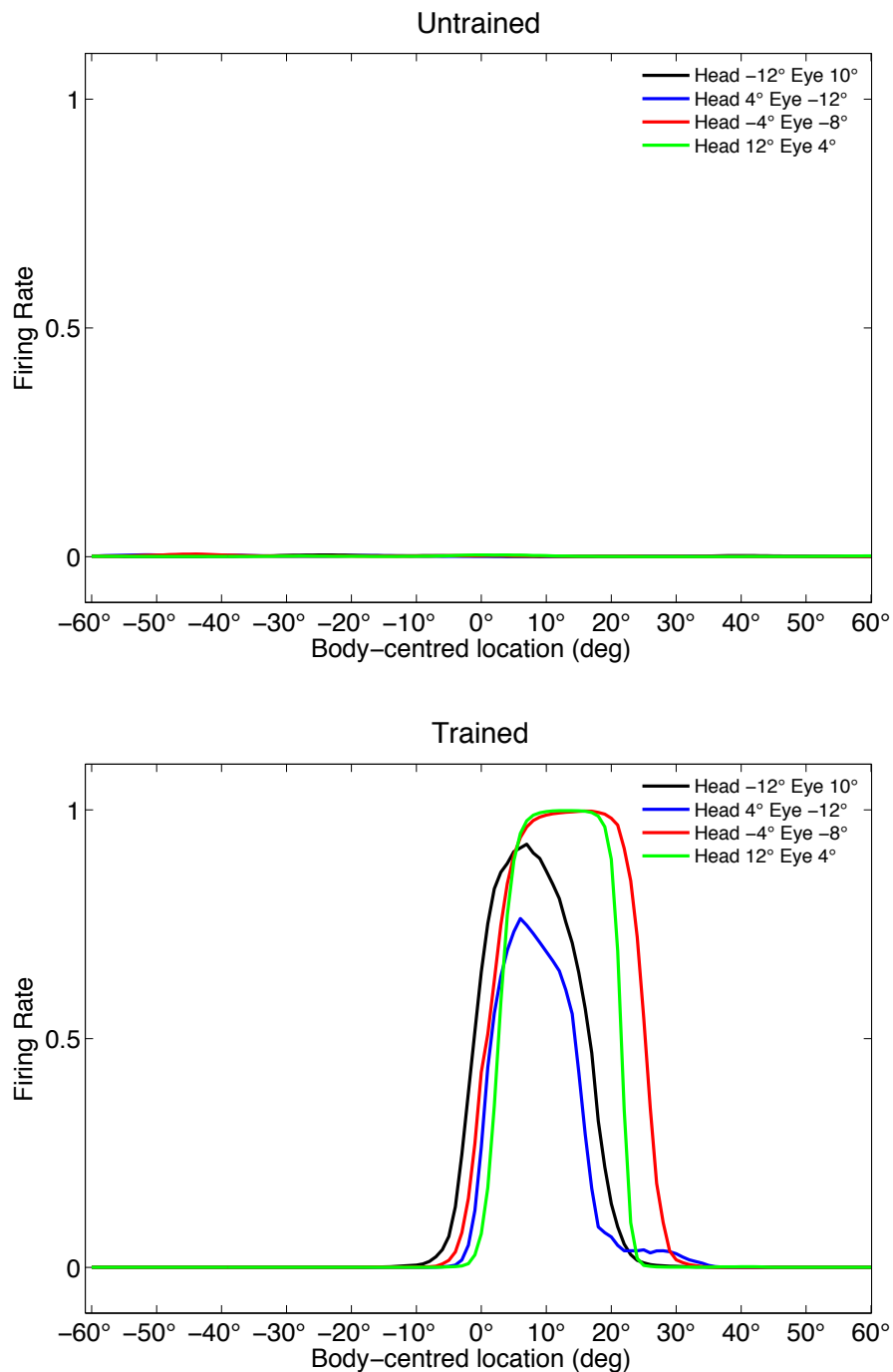


Figure 7.5: Firing rate responses during testing in simulations of the development of body-centred visual neurons. The upper plot shows the firing rate responses of an untrained output cell, while the lower plot shows the firing rate response of the same neuron after training. Each of the curves within the two plots corresponds to testing the network in one of the four different gaze configurations (specific combinations of head and eye position). Each curve shows the firing rate responses of the neuron as the visual target is presented in 121 body-centred test locations covering the range $[-60^\circ, 60^\circ]$. This cell shows the development of a body-centred receptive field after training.

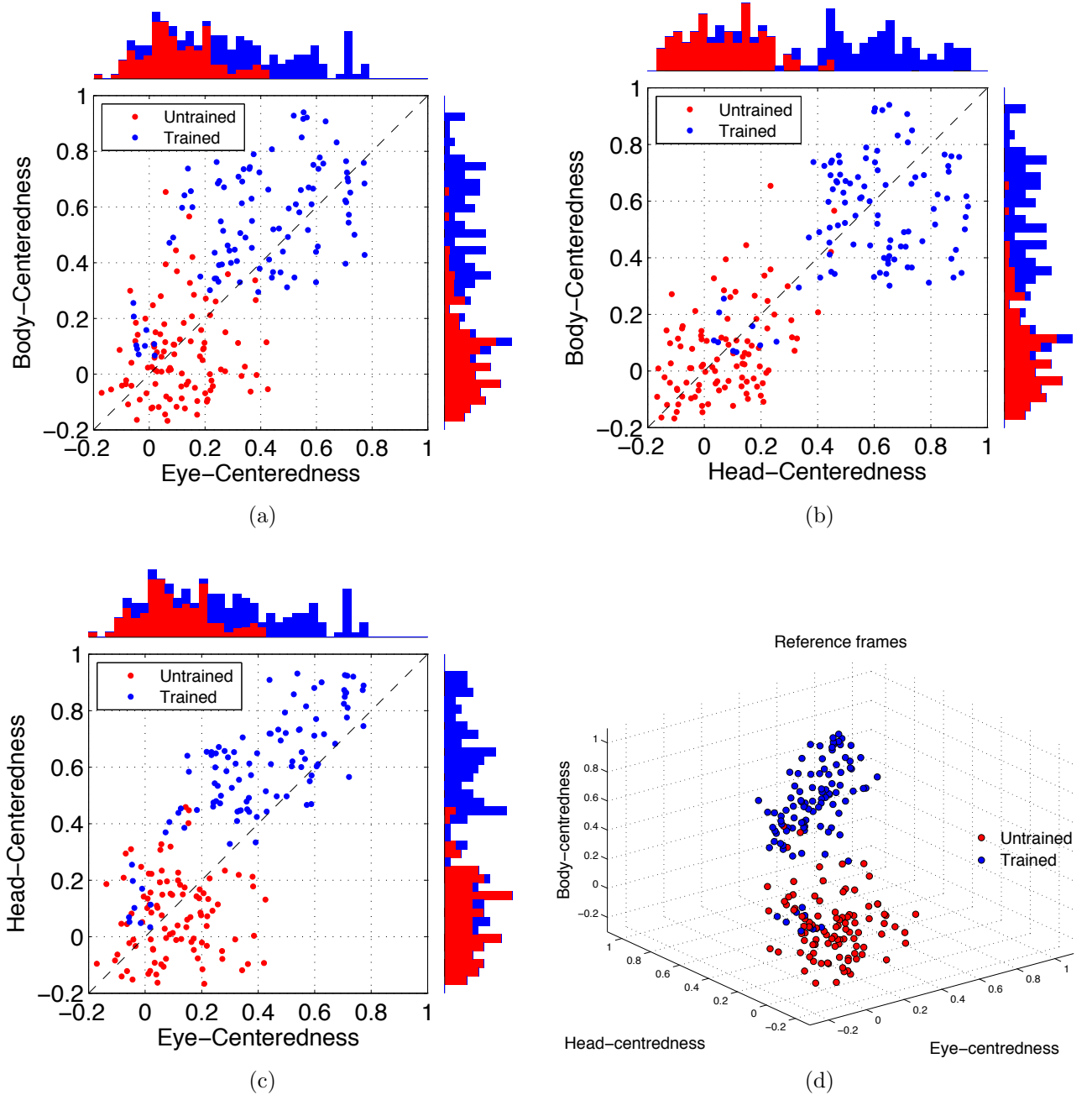


Figure 7.6: Population analysis of the frame of reference of the output neurons in simulations of the development of body-centred visual neurons. Plots (a), (b), (c) and (d) are scatterplots showing the reference frame response characteristics of all neurons in the output layer, where each neuron is plotted as a point corresponding to that neurons particular combination of body-centeredness, head-centeredness and eye-centeredness. Each of the plots (a), (b) and (c) shows a direct comparison between two different reference frame metrics. Figure (d) displays these comparisons in a three dimensional plot. The results are presented before training and after training. Overall it can be seen that before training most of the cells were not responding in a clearly identifiable frame of reference. However, after training most of the cells started to respond in a non-retinocentric frame of reference. In particular, many trained cells have predominantly body-centred responses.

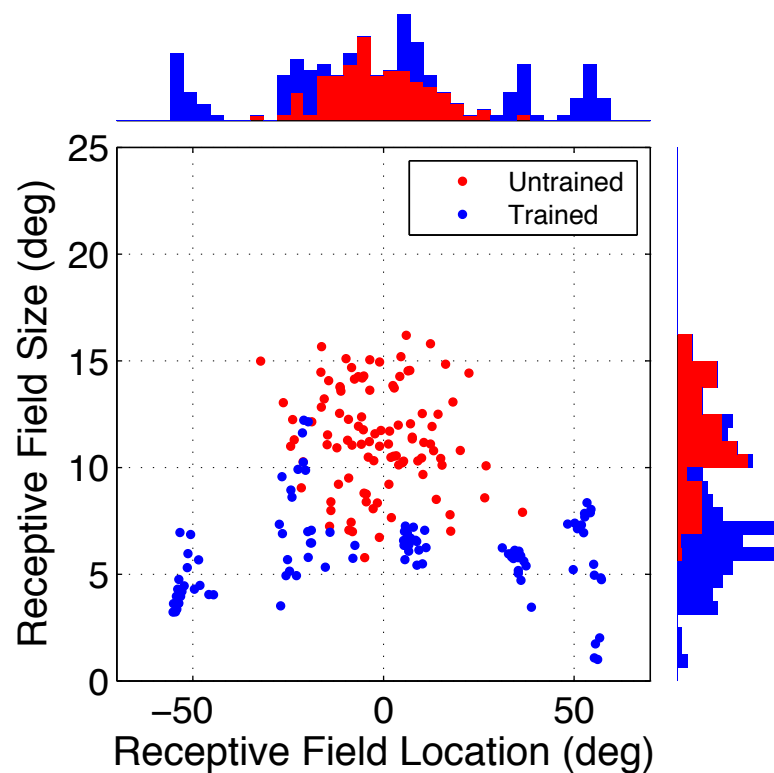


Figure 7.7: Population analysis of the sizes and locations of the receptive fields of output neurons in simulations of the development of body-centred visual neurons. This scatter plot shows the size and location of the body-centred receptive fields of output neurons before and after training. Each point in the scatterplot represents a different output neuron. It can be seen that, after training, the locations of the body-centred receptive fields are grouped according to the six body-centred locations in which the network was trained.

In Figure 7.6 I present the reference frame analyses of the entire population of neurons in the output layer. As can be seen in Figure 7.6, before training the output cells had body-centredness values that were close to zero. However, after training, the output cells increased their body-centred correlation coefficients with values much closer to 1. The body-centredness correlation coefficient increased from an average of 0.06 before training to 0.54 after training.

The population analysis also shows an increase in both the eye and the head-centredness values after training, though more notably in the case of the head-centredness measures. Some neurons develop strongly head-centred responses because, during training, there are periods when the eyes are moving rapidly while the head remains stationary. This kind of movement of the eyes and head has previously been shown to lead by a similar trace learning mechanism to the development of output neurons with head-centred responses (Mender and Stringer, 2013, 2014). Specifically, it is possible that during the periods in which the eyes move while the head remains fixed with respect to the trunk, trace learning will encourages some of the output cells to bind together the different retinal views of the object for the same head position. In this situation, trace learning is working in a somewhat similar fashion to that in which body-centred responses develop. These factors account for the presence of cells with high values for either the body-centred or head-centred reference frame metrics. Thus, the kind of training used in the current simulations, in which the eyes saccade around a visual scene while the head and trunk are stationary, followed by

occasional head movements while the trunk is stationary, leads to a mixture of output neurons with responses in both head-centred and body-centred reference frames.

Figure 7.7 shows the size and location of the body-centred receptive fields of output neurons before and after training. It can be seen that, before training, the neuronal responses have a wide range of receptive field sizes and the receptive field locations appear scattered loosely around the center of the plot in body-centred coordinates. However, after training, the neurons developed receptive fields that were clustered more tightly around the different body-centred locations in which the network was trained. It can be seen that the average size of the receptive fields was decreased from 11.39° before training to 6.01° width after training, thus demonstrating that the neurons learned to respond more selectively to particular positions in body-centred space after training.

These results confirm that the trace learning mechanism hypothesised in Section 7.2.2 enables the network to develop neurons in the output layer with localised, body-centered visual receptive fields. These cells encode the location of visual targets in the same body-centred position irrespective of the gaze configuration (i.e. eye position and head position). Body-centred cells will fire maximally as long as the visual target is present in the same position with respect to the trunk. We hypothesise that such body-centred visual neurons may play a key role in how the dorsal visual system could develop hand-centred visual representations using proprioceptive input signals encoding the position of the hand with respect to the body. This proposed role for body-centred visual neurons is demonstrated

through computer simulation in the next section.

7.2.8 Summary: Body-centred representations

In this chapter I presented an overarching architecture in which the potential role of a proprioceptive representation of the position of the hand in the development of hand-centred representations is explored. This overarching network (see Figure 7.1) contains a population of neurons that will be representing the location of the visual targets with respect to the body (Figure 7.2).

It has been hypothesized that the brain is able to represent the location of visual targets using a body-centred frame of reference (Brotchie et al., 1995; Lacquaniti et al., 1995). I have presented here a new self-organising hypothesis for the development of body-centred visual representations using the trace learning mechanism. I implemented a new unsupervised, self-organising network composed of two layers. The input layer was composed of neurons that encoded the position of the eyes and the position of the head as well as the location of the visual target in the retina, this representation was coupled multiplicatively forming mixed gain fields response profiles in which the neurons would respond maximally to particular combinations of head position, eye position, and retinal position of a visual target. The output layer was implemented as a competitive network and the synaptic weight were updated using a trace learning rule.

In my learning hypothesis (See Section 7.2.2) I use combinations of eye and head move-

ments while a static object is presented to the network. The network exploits the temporal co-occurrence of different input patterns and a form of trace learning would allow the network to associate different views of the object as well as integrating different combinations of eye and head positions in which this object has been seen. This allows the network to develop output cells that learn to respond to the position of objects in body-centred coordinates, irrespective of the position of the eyes and the head.

In this architecture I will use these body-centred representations of the location of the targets and combine them with a proprioceptive representation of the location of the hand to develop hand-centred representations. The representation of the location of visual targets in body-centred coordinates have been hypothesized to serve an important role in motor planning. In order to reach for an object in space, the motor system is thought to calculate a reach vector as the difference between the initial position of the effector and the position of the target or the endpoint of the movement (Buneo and Andersen, 2006). To compute this reach vector efficiently, it has been suggested that the representation of the location of the hand and the location of the target can be subtracted using a common frame of reference (Andersen, 1995). There are many ways and reference frames in which the brain can compute the reach vector. Recoding the location of visual targets from eye-centred coordinates to a body-centred frame of reference, allows a direct computation of the reach vector from a body-centred representation of the location of the arm (Buneo et al., 2009) which would be a more natural frame of reference for the proprioceptive representation of

the limbs. We suggest that this type of transformation is more consistent with the increase of non-retinal receptive fields anchored to parts of the body as the recording sites are moved closer to areas within the later stages of the visuomotor pathway in the primate brain. Moreover, body-centred and body-part centred frames of reference seem to provide more suitable information for converting the motor plan into the necessary activations of muscles in the execution of movements.

7.3 The development of visual neurons representing the hand-centred locations of objects through com- bining inputs from body-centred visual neurons and proprioceptive signals representing the body- centred position of the hand

Introduction

In the first chapters of this thesis, I explored different visually guided learning mechanisms that could allow the network to develop hand-centred visual representations. In this final section I am exploring the potential role of proprioception in the development of hand-centred representations. The term *proprioception* was coined by Sherrington (1907) to

refer to the receptors that provided information about the position and movement of the different parts of the body. The internal representation of the location and movement of the limbs is derived primarily from the activity of muscle spindles conveying information about the length of muscles which is encoded according to the joint angles (Gandevia et al., 1992; Proske and Gandevia, 2012). The internal estimation of position of the limbs seems to be computed from changes in the discharges of muscle spindle populations coming from a variety of sets of muscles (Ribot-Ciscar et al., 2003). As I have mentioned in section 7.1.1, the overarching network (see Figure 7.1) contains a population of input neurons that will provide proprioceptive signals representing the location of the hand with respect to the body. We will show how hand-centred output representations could be developed using this model architecture by combining a visual representation of the position of the targets with a proprioceptive representation of the location of the hand using a common body-centred frame of reference. The co-activation of a particular combination of hand-position and visual target location representing a specific hand-centred configuration, will be learned by a subset of combination neurons. All possible spatial configurations will be learned by different subsets of combination neurons. Then, a trace learning rule operating in the feedforward connections from the combination layer to the output layer will enable individual output neurons to learn to respond to specific spatial configurations invariantly across different body-centred positions.

7.3.1 Stimuli and Methods

7.3.2 Model architecture and parameters

Figure 7.1 shows the general architecture of our overarching network. The architecture uses two input layers that project their feed-forward synaptic connections into a single layer of combination cells. One of the input layers is a proprioceptive signal representing the location of the hand with respect to the body (in this case a trunk-centred and a shoulder-centred encoding will be equivalent). The firing rate responses of the individual neurons in this layer were modelled with a Gaussian activity profile where the peak corresponded to the preferred body-centred location of the hand for that cell. The second input layer represents a visual signal encoding the location of the visual target also in body-centred coordinates. Similarly, we modelled the body-centred preferences of these visual input neurons by imposing a Gaussian activity profile where the peak corresponded to the preferred visual target location in body centred coordinates. We have already shown in Section 7.2 how these preferences to body-centred locations of a visual target could self-organize with visually guided training.

The neurons from the two input layers project feedforward synaptic connections to an intermediate combination layer, which was modelled as a competitive network incorporating lateral competition between neurons (Rolls and Treves, 1998). The input layers had full connectivity with the combination layer. As in the previous chapters using a discrete representation of time, the activation h_i of each combination neuron i was computed by

Equation 7.2 and lateral competition was effected by means of a sigmoid activation function with an adjustable threshold α as given by Equation 7.3. The competitive interaction between the combination neurons would encourage individual cells to fire exclusively to particular combinations of body-centred hand position and object position. At this stage of processing we implemented a Hebb rule in the synaptic connections between the input layers and the combination layer. The Hebb rule has been described previously in Chapter 3 (see Equation 3.1). In this case, the Hebbian learning mechanism will form output neurons that learn to respond selectively to specific combinations of body-centred hand position and target position.

The combination layer was fully connected to the output layer. The final output layer was also modelled as a competitive network. Particular combinations of hand position and target position corresponding to a specific hand-centred configuration could be then bound together across different body-centred locations by output neurons using an invariance learning mechanism. This would allow the output layer to develop hand-centred visual responses across different locations of the hand with respect to the body, as long as the target and the hand occur in the same hand-centred configuration. For this simulation study, in order to bind the different input patterns corresponding to a hand-centred configuration the synaptic weights between the combination layer and the output layer were updated according to the trace rule (see Equation 4.2), which binds together input patterns that tend to occur close together in time.

Parameter	Symbol	Value
<i>Visual Input Layer</i>		
Population size	N^V	4
<i>Proprioceptive input layer</i>		
Population size	N^P	4
<i>Combination Layer</i>		
Population size	N^C	24
Activation function slope	β^C	60
Visual input connectivity rate	ϕ^V	100%
Proprioceptive input connectivity rate	ϕ^P	100%
Sparseness percentile	-	90%
<i>Output Layer</i>		
Population size	N^O	4
Activation function slope	β^R	60
Combination layer connectivity	ϕ^C	100%
Sparseness percentile	-	50%
<i>Training parameters</i>		
Training epochs	-	100
Eta value	η	0.8
Learning rate	α	0.05

Table 7.2: Parameters of self-organising overarching network model of hand-centred visual responses shown in Figure 7.1

The values of the synaptic weights between all the different layers were always initialized randomly with values between $[0, 1]$, which was followed by immediate renormalisation of all the synaptic weight vectors. During training, particular body-centred configurations of the hand with an object are presented to the network. The firing rates of the visual input neurons and proprioceptive input neurons are then set according to their given Gaussian response functions of target position and hand position, respectively. Next, activity is propagated up through the combination layer and the output layer. After

all of the firing rate responses have been computed, learning can be implemented in the feedforward synaptic connections between layers. As mentioned earlier the synaptic weights between the input layers and the combination layer were updated according to a standard Hebb rule, while the synaptic weights between the combination layer and the output layer were updated according to the trace rule. I normalised the length of the synaptic weight vectors at the end of each timestep during training as is usual in competitive learning (Rolls, 2008; Rolls and Treves, 1998). Table 7.2 shows the parameters used in this study.

7.3.3 Training

The training procedure for these experiments consisted of presenting the model with an input that represented the hand position and an input representing the location of the visual target in body-centred coordinates. In the simplified model simulations performed below the input layers represent four different possible hand positions with respect to the body H_1 , H_2 , H_3 and H_4 and four possible positions of the visual target also with respect to the body T_1 , T_2 , T_3 and T_4 . The six input patterns generated for training corresponded to the presentation of a static visual target in two possible positions relative to the hand (i.e. left and right), each of which occurred in three separate positions with respect to the body. In Figure 7.8 we present a schematic representation of the 6 training patterns. The top row shows three configurations H_1T_1 , H_2T_2 and H_3T_3 with the target always on the right of the hand. The bottom row shows three configurations H_2T_1 , H_3T_2 and H_4T_3 with

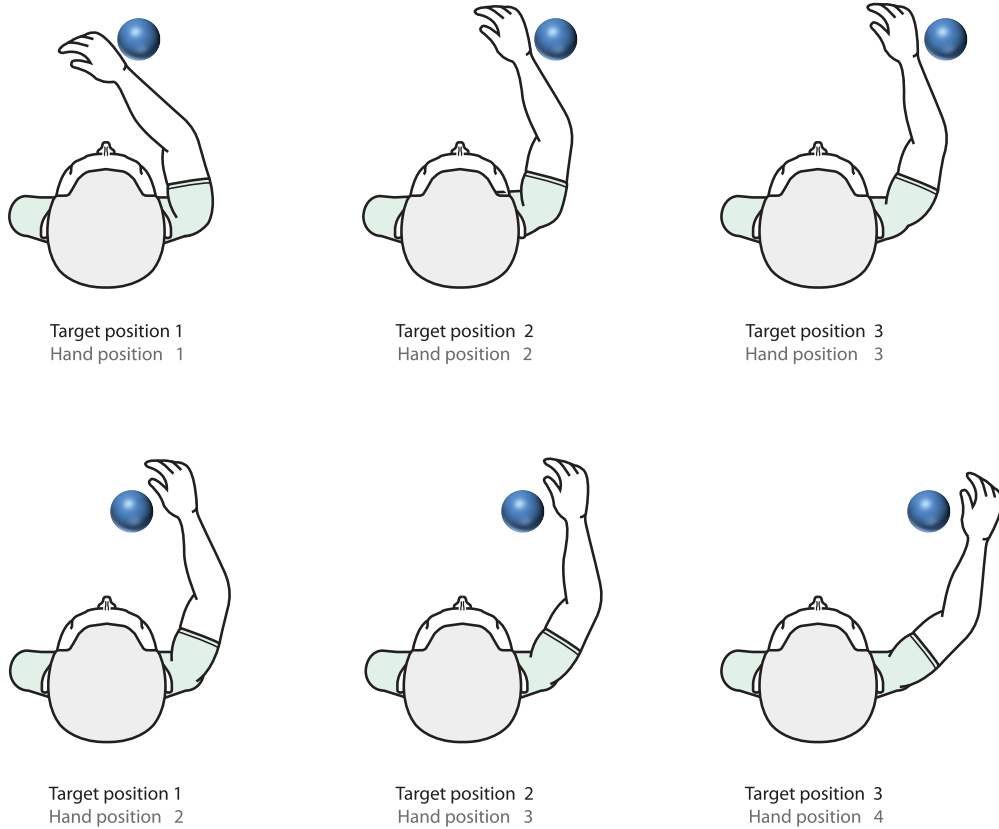


Figure 7.8: Schematic examples of hand-object configurations using for training and testing. The proprioceptive input layer is able to encode four possible hand positions H_1 , H_2 , H_3 and H_4 with respect to the body, while the visual input layer can encode four possible target positions T_1 , T_2 , T_3 and T_4 with respect to the body. During training, we presented the hand and target objects in the following six spatial configurations. The top row shows three configurations H_1T_1 , H_2T_2 and H_3T_3 with the target always on the right hand. The bottom row show three configurations H_2T_1 , H_3T_2 and H_4T_3 with the target on the left of the hand. Thus, during training we presented the visual target in two possible positions relative to the hand (i.e. left and right), with each such spatial configuration occurring in three body-centred positions. The details of the training regime are described in Section 7.3.3 and for the testing regime in 7.3.4

the target on the left of the hand. Therefore, a total of 6 input patterns were generated for the training regime.

The activation of individual neurons within the input layer and their firing rates were calculated every time one of the six possible hand-object configurations was presented. The synaptic weights between the input layer and the combination layer were updated according to a standard Hebb rule which was followed by renormalisation of the synaptic weight vectors. The synaptic weights between the combination layer and the output layer were updated according to the trace rule again followed by weight vector normalisation. In this case trace learning is used to enable the network to develop representations of hand-object configurations in the final output layer, that are position invariant with respect to the body. The combination layer and the final output layer are both implemented as competitive networks. In the combination layer, the competitive interaction between neurons allows the network to represent individual spatial configurations of the hand and object, while the competition within the final output layer combined with trace learning allows the neurons to respond to particular hand-object configurations across different positions of the hand with respect to the body.

In these simulations, one training epoch consisted of the presentation of the six input patterns shown in Figure 7.8. During each training epoch, we first presented the three spatial configurations H_1T_1 , H_2T_2 and H_3T_3 , with the target on the right of the hand. Then we next presented the three spatial configurations H_2T_1 , H_3T_2 and H_4T_3 with the

target on the left of the hand, clustered together in time. This was necessary to enable the trace learning rule to bind together the three configurations with the target on the right of the hand onto a subset of output cells, and bind the three other configurations with the target on the left of the hand onto a second separate subset of output cells. Our network was trained over 100 epochs in total and the training was performed one layer at a time, similar to the VisNet standard procedures. The learning rate used was 0.05. The rest of the training parameters can be found in Table 7.2.

7.3.4 Testing

The network was tested by presenting the same six hand-object configurations shown during training (see Figure 7.8). The firing rates of the output neurons are computed and recorded during the presentation of each of the tested configurations. Throughout the testing phase, the synaptic weights were not changed.

7.3.5 Results

The goal of this simulation was to develop cells in the final output layer that responded selectively to a specific hand-centred configuration (i.e. with the target either on the left or right of the hand) irrespective of where the hand is located with respect to the body. Unlike the simulations in previous chapters the position of the hand is now represented using proprioceptive signals. It was hypothesized that hand-centred representations may

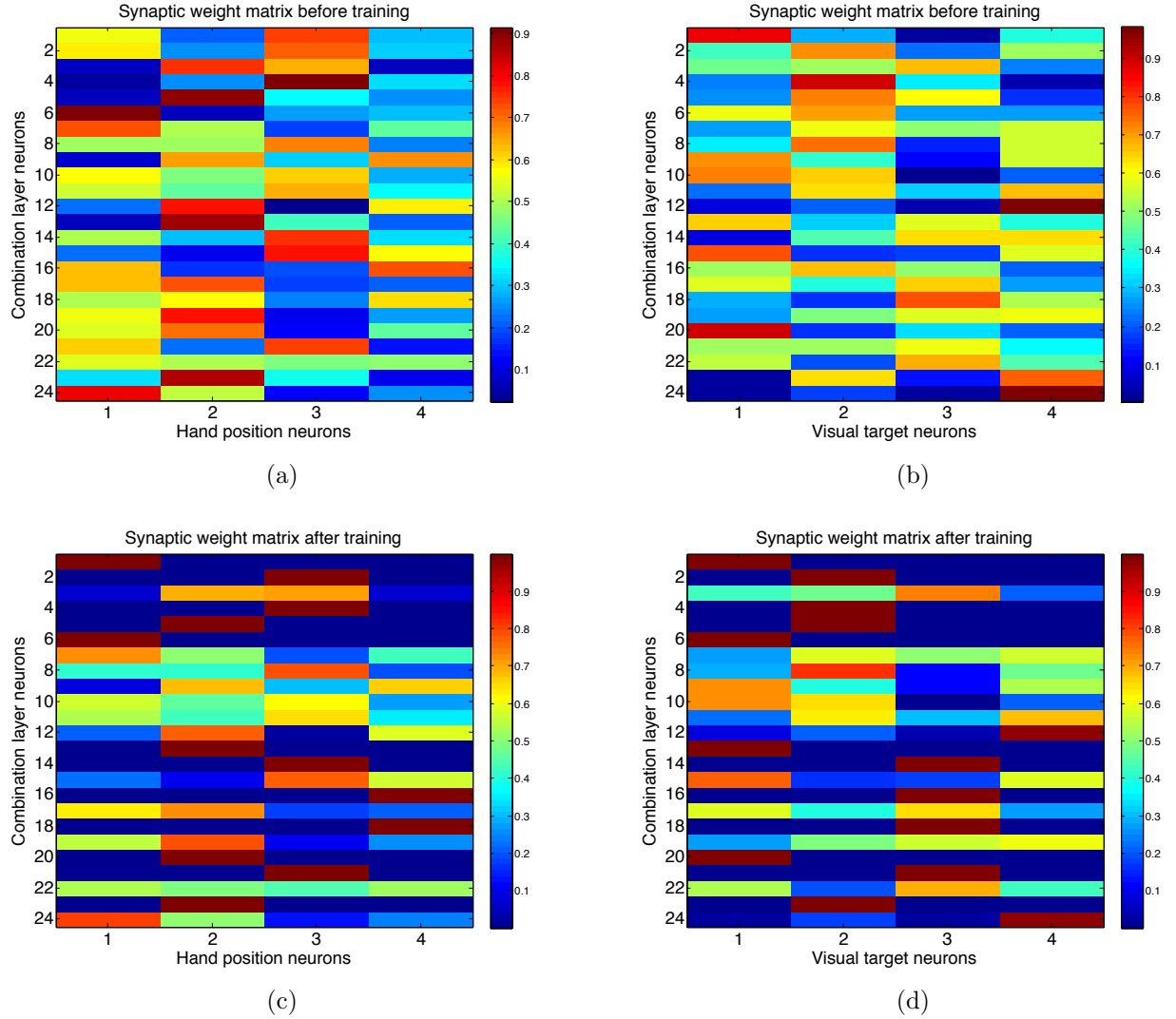


Figure 7.9: Feed-forward synaptic weight matrices from the proprioceptive hand position input layer and the visual target position input layer onto the intermediate combination layer before and after training. The plots show the afferent synaptic weights from the two input layers onto the population of combination neurons. Plots (a) and (c) correspond to the synaptic weight matrices from the hand position input layer onto the intermediate combination layer. Plots (b) and (d) correspond to the synaptic weight matrices from the input layer of cells representing the body-centred location of the visual target onto the intermediate combination layer. The two upper plots show the feed-forward synaptic weight matrices before training and the lower plots shows the same synaptic weight matrices after training. The strength of the synaptic weights are indicated by the colour, the colourbar shows the corresponding values.

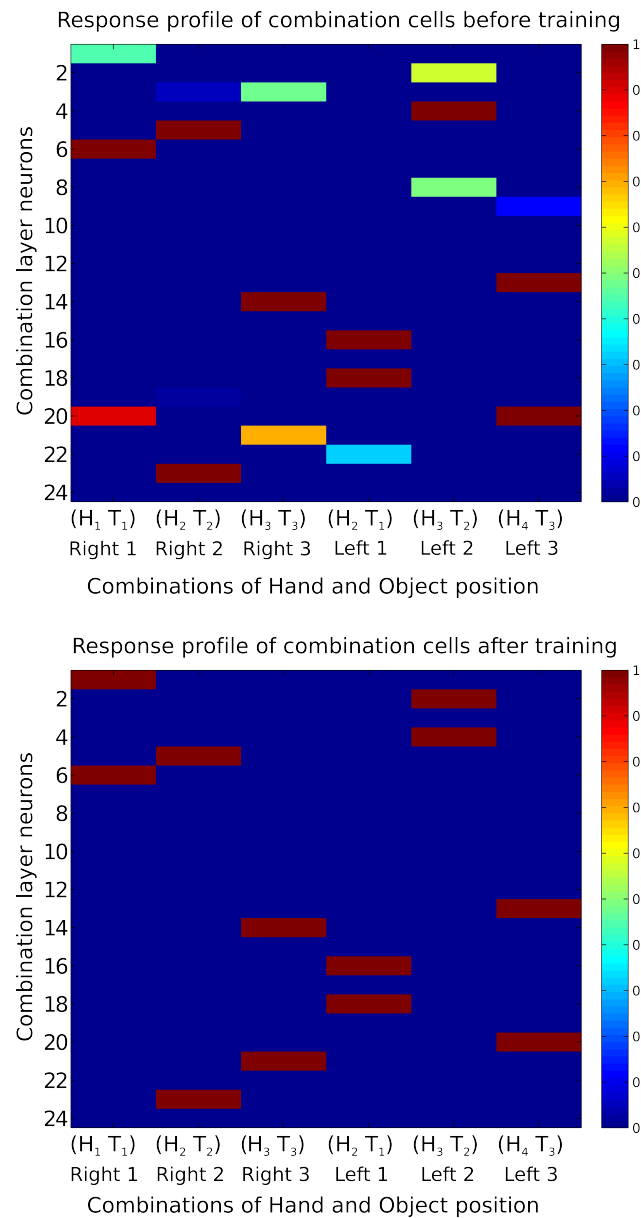


Figure 7.10: Firing rates of cell in the combination layer before and after training. The upper plot shows the firing rate responses of the untrained combination cells during testing. The lower plot shows the firing rate responses of the combination cells after training during testing. The values along the abscissa correspond to each of the 6 hand-object combinations presented during testing. The values along the ordinate correspond to each of the 24 cells in the combination layer. The strength of the responses are indicated by the color, the colorbar shows the corresponding values.

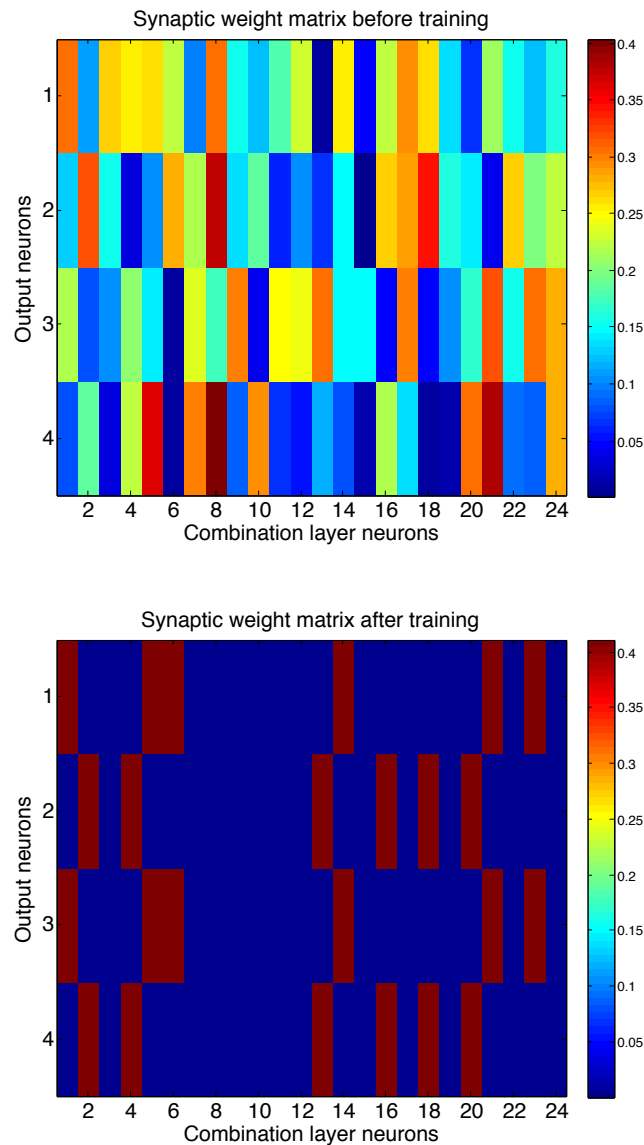


Figure 7.11: Feed-forward synaptic weight matrix from the combination layer to the final output layer before and after training. The plots show the afferent synaptic weights from the combination layer onto the population of output neurons. The upper plot shows the feed-forward synaptic weight matrix before training and the lower plot shows the same synaptic weight matrix after training. The values along the abscissa correspond to each of the cells from the combination layer. The values along the ordinate correspond to each of the four cells in the output layer. The strength of the synaptic weights are indicated by the colour, the colourbar shows the corresponding values.

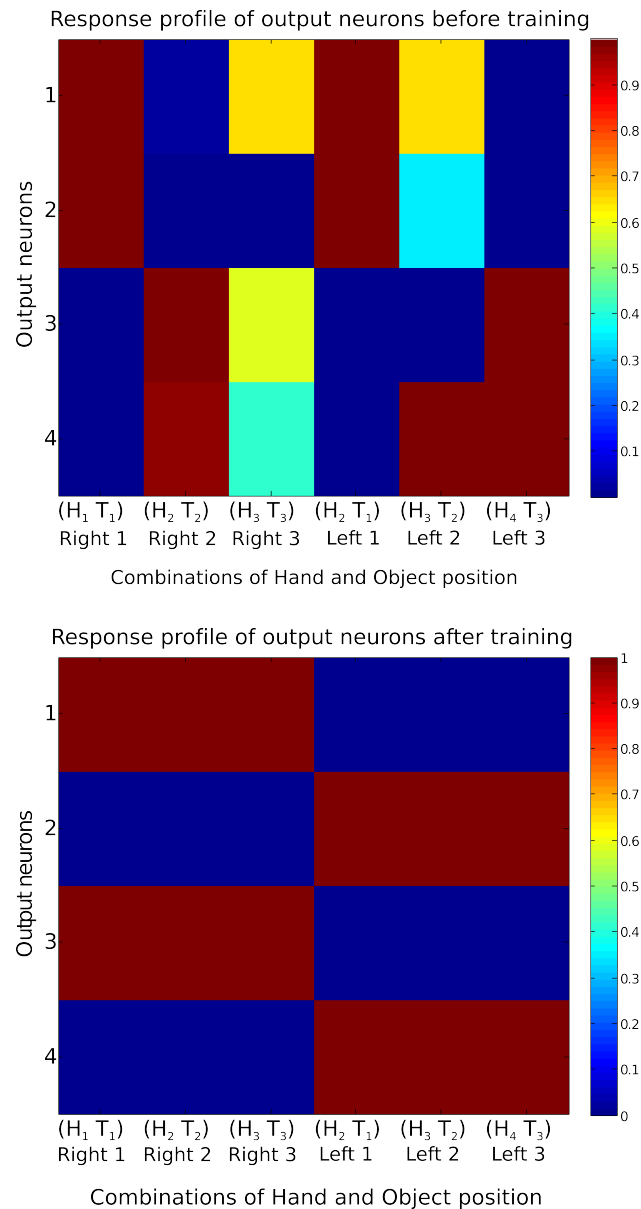


Figure 7.12: Firing rates of the final output layer before and after training. The upper plot shows the firing rate responses of the untrained output cells during testing. The lower plot shows the firing rate responses of the output cells after training during testing. The values along the abscissa correspond to each of the hand-object combinations presented during testing. The values along the ordinate correspond to each of the four cells in the output layer. The strengths of the responses are indicated by the colour, the colourbar shows the corresponding values. The lower plot confirms that the four output neurons each learned to respond to one of the two positions of the target with respect to the hand (i.e. either left or right), and responded invariantly across different positions of the hand with respect to the body.

develop by combining proprioceptive signals representing the position of the hand with respect to the body with a visual representations of the position of a target also in a body-centred frame of reference.

To test the network, we presented a total of six stimuli, three of the stimuli corresponded to the presentation of a static visual target to the right of the hand and the other three corresponded to the presentation of a static visual target to the left of the hand. The firing responses of the output cells before and after training were analysed.

Figure 7.9 shows the feed-forward synaptic weight matrices from the proprioceptive input population representing the hand position onto the combination layer, before training (a) and after training (c). In the same figure we can see the feed-forward synaptic weight matrices from the input population representing the body-centred locations of the visual target onto the combination layer, before training (b) and after training (d). In both cases, the synaptic weights were initialized randomly. However after training, we can see that in some cases the synaptic connections from a particular combination of hand position and target position onto a combination cell were more strengthened. This can be confirmed by the response profiles of neurons in the combination layer before and after training (see Figure 7.10). We can see how each of the cells from the combination layer has learned to respond maximally to a particular combination of a hand location with respect to the body and a target location also in body-centred coordinates.

In the next stage of the hierarchy, Figure 7.11 shows the effect of training on the feed-

forward synaptic weight matrix from the combination layer onto the population of output neurons. The upper plot shows the feed-forward synaptic weight matrix before training and the lower plot shows the same weight matrix after training. Here we can see how before training, the synaptic weights were initialized randomly. However after training, the synaptic weights from combination neurons that represent different combinations of hand position and target position corresponding to a visual target presented to the right relative to the hand (i.e. H_1T_1 , H_2T_2 and H_3T_3) are strengthened onto a particular subset of output cells (i.e. Cells 1 and 3). Similarly, the combination neurons representing a visual target presented to the left relative to the hand (i.e. H_2T_1 , H_3T_2 and H_4T_3) are strengthened onto a different pair of output cells (i.e. Cells 2 and 4).

Figure 7.12 shows the comparison in matrix form, between the firing rates of the final output layer before and after training. The values along the abscissa correspond to each of the hand-object combinations presented during testing. The values along the ordinate correspond to each of the four cells in the output layer. The strengths of the responses are indicated by the colour, the colourbar shows the corresponding values.

Figure 7.12 (top) shows the response profiles of the cells in the output layer before training. It can be seen that each of the four cells initially respond in an unstructured manner to the different hand-object configurations.

However, in the lower plot of Figure 7.12 we can see how after training the output cells respond selectively to the same hand-object configuration across all of the tested hand

locations with respect to the trunk. Half of the output cells responded to the object shown to the right relative to the hand and never responded when the object was presented to the left, while the other half of the output cells responded selectively to the object shown to the left relative to the hand and never responded when the object was presented to the right relative to the hand. This result confirms that the four output neurons each learned to respond to one of the two positions of the target with respect to the hand (i.e. either left or right), and responded invariantly across different positions of the hand with respect to the body.

Altogether, the firing rate responses of the individual cells in the output layer demonstrate that our network is able to develop neurons that respond to hand-centred configurations irrespective of where the hand is located with respect to the body using a proprioceptive representation. These cell representations are developed by combining proprioceptive signals representing the position of the hand with respect to the body with body-centred visual representation of the position of a target.

7.3.6 Discussion

Visual cells that learnt to develop a localized receptive field with respect to the hand were simulated in the previous chapters. Each of these cells would fire maximally irrespective of where on the retina the cell's preferred hand-object configuration appeared. Furthermore, these visual cells would also fire maximally irrespective of where the hand is located with

respect to the body, as long as the same hand-centred configuration appeared somewhere on the retina. However, in previous chapters we did not explicitly model a proprioceptive input representation of the hand. In this case an output cell would have to learn to fire maximally to a specific hand-object configuration irrespective of where the hand is located with respect to the body.

In this last subsection I have shown how representations of hand-object spatial configurations could develop by combining proprioceptive signals representing the position of the hand with respect to the body with visual representations of the position of a target also with respect to the body. I showed how these signals could be combined in the combination layer by using a common body-centred frame of reference. Particular subsets of combination neurons would represent combinations of hand and target positions with respect to the body corresponding to a particular hand-object configuration (e.g. target on the left or right of the hand). The subsets of combinations representing a particular hand-object configuration were bound together by output neurons using a trace learning mechanism. This allowed the network to develop selective hand-centred responses across different hand positions with respect to the body (i.e. invariant to the position of the hand with respect to the body). The results show that the output cells were able to discriminate and respond selectively to one of the two tested hand-object configurations (i.e. Left or Right) across different hand positions with respect to the body.

In the overarching architecture presented in this chapter and shown in Figure 7.1, I

presented an hypothesis of how hand-centred representations can be developed by combining vision with proprioception. The proprioceptive representations of the location of the hand allows our model to account for the experimental studies showing that hand-centred cells seem to maintain their firing properties in the absence of a visual cue of the hand (Graziano, 1999; Graziano et al., 1994b). However, the location of the hand is typically represented using inputs from a variety of available sensory modalities, including both vision and proprioception. We learn the combination of inputs from different sensory modalities that correspond to a particular position of the hand. The visual and proprioceptive representations of the location of the hand will be mostly congruent throughout our everyday experience. In our overarching architecture I suggest that the visual representation of a hand-centred configuration could be associated by hebbian pattern association learning with the hand-centred output representations derived by combination of proprioceptive hand position signals and visual target position signals in a body-centred reference frame (see Figure 7.1.1).

The complex interaction of visual and proprioceptive signals at different stages of visuomotor processing has been widely studied. However, it is still unclear how and when the visual and proprioceptive inputs converge and how these signals are subsequently computed (Graziano, 1999; Rossetti et al., 1995; Sober and Sabes, 2003, 2005). As we have discussed above in this thesis, a number of studies have manipulated systematically the availability of and congruency between the visual and the proprioceptive representation

of the hand position. There are studies suggesting that the visual representation of the position of the hand tends to bias a proprioceptive representation when the different signals are incongruent (Farnè et al., 2000; Holmes and Spence, 2005; Makin et al., 2007). However, other studies have suggested the dominance of proprioception (van Beers et al., 2002). To reconcile these findings, it has been suggested that the reliability and congruency of the sensory modalities in which the limbs and targets are perceived could play a role in determining the type of signal or representation that dominates (McGuire and Sabes, 2009, 2011; Rossetti et al., 1995; Sober and Sabes, 2003, 2005). Furthermore, the visual and proprioceptive representations of the location of the hand might have distinct roles at different stages of goal-directed movements (Sarlegna and Sainburg, 2009), as well being weighted differently in different regions of the brain. For example, the anterior IPS seems to be modulated more strongly by proprioceptive representations of the hand position, while the posterior IPS shows a stronger visual contribution (Makin et al., 2007).

In the next chapter we will present a more extended discussion of the potential role of different sensory and positional signals and their co-activation as an underlying mechanism in the development of body-part centred receptive fields, as well as presenting some suggestions for further investigations.

That is the beginning of knowledge—the discovery of something we do not understand.

—Frank Herbert, *God Emperor of Dune*.

8

Discussion

In this chapter I provide a summary of the main findings presented in this doctoral thesis and propose suggestions for future work.

8.1 Summary of main findings

The location of visual targets is initially encoded in a retinotopic frame of reference. However, this information is transformed in later stages of processing into different frames of reference that are more suitable to guide our behaviours and that provide perceptual stability (Flanders et al., 1992; Soechting and Flanders, 1992). A number of neurophysiological studies in the posterior parietal cortex and premotor areas of the primate brain have reported a continuum of complex non-retinal reference frames, including cells that

respond to the location of objects with respect to the hand (Bremner and Andersen, 2012; Graziano and Gross, 1998; Graziano et al., 1997, 1994b). These cells would respond to a visual object within the hand-centred receptive field, irrespective of where on the retina this spatial configuration is seen as well as where the hand located with respect to the body. Some of these representations are thought to be relevant for facilitating hand-object interactions (Holmes, 2013; Makin et al., 2008).

The main focus of this thesis was to explore how neurons in the primate brain could self-organize to develop such hand-centred representations. This investigation was carried out using self-organizing neural network models, in order to provide a theoretical framework of how these cells might develop during unsupervised visually guided learning.

In Chapter 3 I presented our first self-organising hypothesis of how VisNet may develop hand-centered visual representations in the output layer using *Continuous Transformation* (CT) learning (Stringer et al., 2006). This learning mechanism relies on the spatial similarity of the input patterns as a particular spatial configuration of hand and object shifts over the retina. In Chapter 3 I showed that CT learning seemed to be able to develop output neurons that respond to specific locations of a visual target object with respect to the hand when it is initially trained with a small number of hand-centred locations. However, when the number of hand-centred training locations was increased, CT learning decreased its spatial selectivity. It became clear that the highly associative aspect of this learning mechanism had unwanted deleterious effects. CT learning tended to produce output cells

that responded to large regions of hand-centred space. This was an indication that CT learning was too associative when the model trained on a continuum of target locations, which would not allow the network to develop localised representations. However, I argued that the results obtained with CT learning could potentially provide an account for the development of a continuous map of large surrounding areas of the visual space around the hand. Particularly, in the case of cells with large receptive fields surrounding large parts of the body. Nevertheless, in my view, it was clear that a CT learning mechanism would not be the best candidate for accounting for the development of more localised hand-centred representations.

In Chapter 4 I presented a second self-organizing hypothesis using a learning mechanism called *trace learning* (Földiák, 1991; Rolls, 1992). In this hypothesis the learning rule incorporated a memory trace of recent neuronal activity in order to encourage neurons to learn to respond to input patterns that tend to occur close together in time. I assumed initially that there would be periods of time in which the eyes would tend to shift about a visual scene rapidly while the hand would remain stationary near an object. It has been previously shown that trace learning is robust enough that it can allow a competitive network to develop supra-retinal representations even if the appropriate body movements only happen for a small portion of the time (Mender and Stringer, 2014). Therefore, trace learning could still produce hand-centred representations even if for the majority of the time the dynamics of gaze changes and hand movements are entirely different to what is

required in this self-organizing hypothesis.

I showed in Chapter 4 how trace learning bound together successive retinal images corresponding to the same hand-object configuration, thus forming localised hand-centred output representations. The results presented in Chapter 4 showed how the network was capable of generating output cells that are tuned to specific target locations in a hand-centred frame of reference. Furthermore, unlike the CT learning hypothesis, the network was now able to develop localised hand-centred receptive fields even as the number of hand-object configurations increased during training, almost approaching a continuum of target locations.

In Chapters 5 and 6 I further advanced the trace learning hypothesis by gradually increasing the realism of the visual training conditions. In Chapter 5 I trained the network using more natural visual scenes consisting of multiple targets presented at the same time around the hand instead of just an isolated object in a particular location with respect to the hand. I presented three sets of experiments, and I introduced the principles of the *statistical decoupling* that allows the output cells of the network to develop single localised receptive fields even when the network is trained with more than one object near the hand at a time. First, VisNet was trained on idealised images consisting of pairs of circular objects presented simultaneously near the hand. The results showed that VisNet was able to develop neurons in the output layer that respond exclusively to single, localised hand-centred locations even when the network was always presented with objects seen

together in multiple hand-centred locations during training. This showed that a statistical decoupling between different hand-centred locations is possible even when learning occurs in the presence of multiple objects around the hand.

In a second study within the same chapter (see Section 5.5), VisNet was trained with computer generated images consisting of a realistic image of a hand and a variety of natural objects, presented within different textured backgrounds during training. The network was then tested with just one textured object near the hand in order to verify if the output cells are capable of building hand-centered representations with a single localised receptive field. Similarly, these simulations also showed that VisNet was able to develop neurons in the output layer that respond exclusively to single, localised hand-centred locations, even when the network was trained with a variety of natural objects and the configurations were shown against more realistic textured backgrounds. In an additional simulation (see Section 5.9) I showed how some of the cells with hand-centered receptive fields decreased their shape selectivity and started responding to a larger proportion of objects within a localised region of hand-centered space as the number of objects presented in overlapping locations during training increases. This happened because the associative learning in the feedforward connections increased the tuning of an output cell to respond to features of an object presented in that hand-centred location. Then when a new object is presented in the same hand-centred location, the two objects may share some features in common. The activation of these common features encourages the same subset of output neurons

to respond again to the new object. Subsequently, the features of the new object become associated with the same subset of output cells. This process will gradually decrease the object selectivity of these output cells and this is consistent with the notion of forming localised hand-centred receptive fields instead of object-specific representations. These results could also account for the difference in the proportion of cells that are highly selective to shape between the ventral and dorsal streams as well as being consistent with the experimental evidence showing a lower spatial resolution and poor feature binding and object recognition in near hand space (Gozli et al., 2014, 2012).

In the third study presented in Chapter 5, I further explored these same learning principles using natural visual scenes collected by volunteers. Unlike the previous simulations where the hand-centred object locations were carefully controlled, this time the objects could appear in any location around the hand. One of the particularly interesting differences in the results was that instead of developing clearly delineated visual receptive fields, the localised receptive fields near the hand had idiosyncratic shapes. Taken together, the results in Chapter 5 provided an important step in showing how single, localised, hand-centered receptive fields could emerge under more ecologically realistic visual training conditions.

The ecological plausibility of the trace learning hypothesis was further explored in Chapter 6 using eye and head movements recorded from human participants as they completed a jigsaw. A head mounted camera captured images of the hand and jigsaw, while eye

movements were recorded by an eye-tracking device. This combination of data allowed us to reconstruct the retinal images seen as humans undertook the jigsaw task. These retinal images were then fed into a new time accurate differential formulation of the VisNet model in order to faithfully represent the temporal dynamics of the recorded gaze changes. These results provided an important step in demonstrating how localised, hand-centred receptive fields can emerge through visual experience that is shaped by the natural eye and head movements of human participants.

Lastly, in Chapter 7 I presented a complete model architecture that develops hand-centred visual representations by incorporating proprioceptive input signals representing the position of the hand with respect to the body.

I began by providing a self-organizing hypothesis of how trace learning can first help to develop a body-centred representation of the locations of visual targets. The model incorporated retinotopic visual input signals that were gain-modulated by the position of the eyes and the head. In my learning hypothesis I used combinations of eye and head movements while a static object was presented to the network. This allowed the network to develop output cells that learned to respond to the position of objects in body-centred coordinates, irrespective of the position of the eyes and the head.

Subsequently I showed how these body-centred representations of the location of visual targets could be used in combination with the proprioceptive representation of the location of the hand in order to develop hand-centred output representations. I presented a simple

demonstration of how this could occur in the overarching network architecture. This network had an input layer receiving a proprioceptive representation of the position of the hand in body-centred coordinates and another input layer represented the location of the visual target in body-centred coordinates. Both of these layers had feed-forward synaptic connections to a combination layer. The combination population learned the different combinations of hand and visual target positions, and projected its outputs to a final output layer that also used trace learning to develop hand-centred representations of target location. The results showed that the output cells would respond selectively to a hand-centred configuration across different hand positions with respect to the body (i.e. hand posture invariant).

Moreover, in this network architecture I hypothesized that the final output layer could also receive parallel signals from visual representations of the hand and background target objects as used in the previous chapters. The visual representation of a hand and target objects could then be associated with hand-centred output representations derived from proprioceptive inputs by pattern association learning due to their frequent and congruent co-activation. After learning, this would allow the network to respond to a hand-object configuration even when a visual cue of the hand is not available.

8.2 Future work

In this thesis I primarily focused on the development of hand-centred visual representations and I presented a number of self-organising hypotheses to account for the development of such responses. In the first experimental chapters, the location of the hand was represented visually. However, towards the end of the thesis I presented an hypothesis of how these hand-centred responses might develop using a proprioceptive signal representing the location of the hand.

There are a number of potential future directions for modelling the development of body-centred and body-part centred visual representations. For example, I acknowledge that some of the complexity in the connectivity and interactions between the different stages of visual processing was not explicitly represented in our model of the primate visual system. There are important back-projections and integration of multisensory signals at different stages of processing that may play an important role in the development of hand-centred representations. For example, the delayed-reach paradigm used in many of these studies suggests that these neurons may continue to maintain their activity even when the visual target disappeared from view, as long as they keep planning to move their arm to the intended target location. This suggests that the hand-centred representations may be receiving additional signals conveying a memory of the target location or alternatively the cells could be maintaining the representation of a selected motor plan. These additional features could be modelled by adding recurrent collateral synapses, that could keep a

persistent packet of activity even in the absence of any external input.

Furthermore, in our simulations of the visually-guided development of hand-centred neurons we have not yet attempted to identify specific layers of the models with particular brain areas in the primate visual system. The architecture and connectivity of the relevant parts of the visuomotor system is very rich, with hand-centred neurons occurring in multiple different brain areas. Future work should develop models with more complex architectures representing more closely the known neuroanatomy of the visual pathways as well as integrating the relevant multisensory information at different stages of processing. This would help us to elucidate the interaction between the different signals that give rise to some of the known properties of hand-centred representations.

It has been reported that there is a wide range of heterogeneous responses in different parts of the posterior parietal cortex. Some of the representations do not seem to have a clear frame of reference, as well as showing gain modulation by a variety of positional signals from different parts of the body. It is currently unclear how all of these different representations develop. In this case, we also suggest that introducing more complex architectures representing more closely the neuroanatomy of the regions involved. For example, including reciprocal connections between the different stages of processing as well as recurrent connection within individual areas, may help us to account for some of the heterogeneous and idiosyncratic response properties that have been reported.

In much the same spirit, future work should also aim at incorporating more biophysi-

cally accurate models of individual neurons (e.g. Hodgkin and Huxley and spiking neural networks). However, these models tend to be computationally very expensive for large-scale simulations using multilayered neural networks, specially for self-organising hypotheses requiring modifiable synapses.

In every theoretical endeavour there is inevitably a compromise on the level of abstraction required to produce a simple enough model that allows one to carefully control and systematically isolate the effect of specific variables, and yet at the same time being able to maintain a faithful and meaningful representation of the phenomena that we are trying to understand. This thesis was carried using an established model of the primate visual system, VisNet, as well as developing a novel overarching network architecture using competitive networks. With the available computational resources during this doctoral project, firing-rate based models offered feasible computational costs and consequent run times, while at the same time offering an acceptable level of abstraction to test biologically plausible, self-organising hypotheses of the development of hand-centred representations. There are valuable general principles of self-organization presented in this thesis that can provide an insight into how the statistics of our visual experience and consequently the incoming sensory signals can determine the kind of representations that have been proposed to exist in different regions of the primate brain.

The hand-centred representations in parieto-frontal circuits are thought to play a role in guiding movement towards target objects. These neurons, by providing a representation of

the location of the target with respect to the hand, are thought to play a role in encoding the vectors for planning a reach towards the visual target (Bremner and Andersen, 2012). The act of reaching, however, is a complex process that integrates a wide range of sensory and motor information to plan and control movement trajectories. It is important to emphasize that the self-organising hypotheses presented in this thesis are intended to provide an account of how cells may self-organize to develop hand-centred representations from eye-centred input, using local learning rules, and do not address directly the dynamics of reaching. Nevertheless, these hand-centred responses are thought to play an important role in movement planning and future work should attempt to integrate these representations into a full motor model.

Model evaluation: How the brain utilizes information from hand-centred cells

At present, much of the research done with VisNet has focused on providing self-organising mechanisms that allow neurons in the output layer to learn to selectively represent a stimulus (in this case hand-centred configurations), across different retinal locations. These neurons would convey maximal information and would be the most informative to reach a decision about which stimulus was presented. However, the way this information is propagated in subsequent stages of processing has not been explicitly explored. In the brain, there is clearly no homunculus selecting the cells according to the information they contain about a known stimulus. We suggest that a learning mechanism using information from error feedback signals could underlie this process, allowing subsequent stages of processing

to strengthen the synaptic connections from the most-informative cells.

Hand-centred representations are thought to play a role in guiding movement towards target objects. In the context of how we learn to accurately control our movements, the output signals from motor planning stages could be either strengthened or inhibited depending on the errors that they cause in goal directed behaviour. Efference copy signals from motor commands seem to predict the motor consequences of our actions as well as making predictions of the sensory consequences of such motor commands. Differences in estimates and observations produce error feedback and learning is thought to be driven by sensory prediction errors (Shadmehr et al., 2010). The most informative cells would provide the best estimates and this would produce a strengthening of such connections. Conversely, feedback signals would weaken the synaptic connections from uninformative cells.

As we mentioned earlier, motor dynamics involve complex processes that integrate a wide range of sensory and motor information. Therefore, it is beyond the scope of this thesis to attempt to provide an account on how a full sensorimotor system self-organizes. However, future work should attempt to show how this process takes place in a biologically plausible architecture using local learning rules. Particularly, if we attempt to provide a full account of sensorimotor organization.

Expanding the overarching network architecture: Integration and interaction of different signals

This thesis concluded with the suggestion of an overarching self-organizing network architecture in which hand-centred configurations could be developed using visual and proprioceptive signals. However, there are alternative architectures and network interactions in which these kinds of representations could be generated. For example, in my overarching network (see Figure 7.2) I included a constant signal carrying a visual representation of the hand and the target in addition to the proprioceptive representations. However, this signal could feed into the network at different stages of processing. Furthermore, the potential competition and interactions when these representations are incongruent or in conflict was beyond the scope of this thesis and hence not explored. I propose to expand this work in order to explore in more detail the complex interaction of visual and proprioceptive signals at different stages of motor planning.

Additionally, some of the hand-centred cells representing peri-hand space have been reported to be bimodal, with overlapping and spatially congruent tactile and visual receptive fields (Graziano, 1999; Graziano and Gross, 1998; Graziano et al., 1997, 1994b; Hyvärinen and Poranen, 1974; Rizzolatti et al., 1981). In this case, a constant tactile feedback could provide an additional sensory signal into the model. In future work I propose to explicitly model the effects of tactile stimulation in order to explore the role of touch in hand-centred representations. The tactile input will be generally activated whenever an object that is

in a particular position relative to the hand makes contact with the skin. This tactile activation will be co-active with particular combinations of postural and visual signals in the overarching network. The co-activation of tactile stimulation and the visual input from a hand-centred configuration has been shown to develop this spatial congruency between visual and tactile representations using artificial neural networks (Magosso et al., 2008, 2010).

The co-activation of different signals representing combinations of multisensory inputs and positional signals, could be driving the development of some of the more abstract, multi-modal and invariant properties of body-part centred representations that appear in later stages of visuomotor processing. We suggest that this underlying principle of co-activation can be operating to bind together the multiple sensorimotor signals that correspond to similar body-part centred locations and similar motor outcomes. This is consistent with the observed illusions of body ownership (e.g. rubber-hand illusion) in which the correlation in the activation of bottom-up sensory signals can be manipulated to induce changes in the representation of a body-part location (Tsakiris and Haggard, 2005).

In summary, I propose that an extension to my overarching network architecture should incorporate in more detail some of the known aspects of the neurophysiology of the sensorimotor system, including known reciprocal connections between the different brain areas and recurrent connections within individual areas. This would also require the integration of the relevant multisensory and motor signals at the corresponding stages of processing. I

suggest that more biophysically accurate models of neurons should be explored, as long as it does not require including an unnecessary level of detail that may make the model overly complicated and thus inadequate to advance our understanding of the self-organization of these neuronal response properties. Lastly, I would also suggest that the future modelling hypotheses should be tested and driven as much as possible with signals from natural eye, head and body movements recorded from primates and human participants at different stages of development, in order to provide a more ecologically plausible simulation of the natural development of sensorimotor integration.

The need for more neuropsychophysiological work

In terms of avenues for future neurophysiological work, it is still yet unclear how the hand-centred responses reported by Bremner and Andersen (2012) in area 5d relate to other hand-centred representations reported earlier in the context of peri-hand space (Graziano, 1999; Graziano and Gross, 1998; Graziano et al., 1997, 1994b; Hyvärinen and Poranen, 1974; Rizzolatti et al., 1981). Furthermore, there is still an open debate on the best way of identifying the frame of reference in which the individual cells seem to represent sensory information, or even if this functional distinction is necessary at all. The majority of the cell responses recorded in these areas are quite complex and heterogeneous, making it difficult to strictly categorize them in particular frames of reference. A reference frame is an actual abstraction or estimation based on how the cell responses fit a particular parametric model given a very contrived task.

Most of the tasks from which these responses are recorded involve quite a passive and limited set of experimental conditions, which makes it difficult to disentangle not only the frame of reference in which a particular cell encodes the location of a target, but also how visual, proprioceptive, tactile and motor signals are weighted and integrated during the task. For example, in some studies the head of the monkey is restrained, which makes it difficult to assess a difference between a head-centred, body-centred or shoulder-centred representation. On the other hand, PPC neurons may very well have interesting dynamical properties as recently explored (Bremner and Andersen, 2014), in which the frame of reference of a cell could be varying during different moments of a task. It has been argued that the responses of the cells could also be affected by the context of the tasks. The congruency between the different sensory modalities in which the limbs and targets are perceived and the reliability of these different sensory representations seem to also play an important role in the type of encoding scheme of the neuron.

However, there are conflicting results that make it difficult to determine the roles and dynamic interactions between the different representations. For example, on the one hand it was claimed that the presence or absence of visual cues of the initial hand position did not seem to have a strong effect on the response profiles of some cells in the PPC (Buneo and Andersen, 2012). It is important to note that in the experiment performed by Buneo and Andersen (2012), the visual and proprioceptive cues were always spatially congruent, which makes it hard to disentangle the weight of each representation. Other

studies have also suggested that the reference frame in which the spatial representation of the hand and target is computed is largely independent of the modality of the incoming sensory information (e.g. McGuire and Sabes, 2011). On the other hand, it has been recently shown that even when in some regions the predominance of either gaze-centred or body-centred encoding seems to be largely unaffected by differences in sensory modality of incoming signals, other areas like the human left anterior intraparietal sulcus (aIPS) seem to flexibly switch between gaze or body-centred encoding depending of the sensory modality of the incoming signals (Leoné et al., 2015). This would support the notion of an intermediate sensory-dependent transition between the predominantly eye-centred representations and predominantly body-centred representations.

Taken together, the abundance of conflicting results clearly suggest the need for a more rigorous and thorough examination of the dynamical properties of cell responses, controlling and systematically manipulating a wider range of variables.

References

- Abrams, R. A., Meyer, D. E., and Kornblum, S. (1990). Eye-hand coordination: oculomotor control in rapid aimed limb movements. *Journal of Experimental Psychology: Human Perception and Performance*, 16(2):248.
- Andersen, R. A. (1995). Encoding of intention and spatial location in the posterior parietal cortex. *Cerebral Cortex*, 5(5):457–469.
- Andersen, R. A., Bracewell, R. M., Barash, S., Gnadt, J. W., and Fogassi, L. (1990). Eye position effects on visual, memory, and saccade-related activity in areas lip and 7a of macaque. *The Journal of Neuroscience*, 10(4):1176–1196.
- Andersen, R. A., Essick, G. K., and Siegel, R. M. (1985). Encoding of spatial location by posterior parietal neurons. *Science*, 230(4724):456–458.
- Andersen, R. A., Essick, G. K., and Siegel, R. M. (1987). Neurons of area 7 activated by both visual stimuli and oculomotor behavior. *Experimental Brain Research*, 67(2):316–322.
- Angel, R., Alston, W., and Garland, H. (1970). Functional relations between the manual and oculomotor control systems. *Experimental neurology*, 27(2):248–257.
- Barash, S., Bracewell, R., Fogassi, L., Gnadt, J., and Andersen, R. (1991a). Saccade-related activity in the lateral intraparietal area. i. temporal properties; comparison with area 7a. *Journal of Neurophysiology*, 66(3):1095–1108.

- Barash, S., Bracewell, R., Fogassi, L., Gnadt, J., and Andersen, R. (1991b). Saccade-related activity in the lateral intraparietal area. ii. spatial properties. *Journal of Neurophysiology*, 66(3):1109–1124.
- Batista, A. and Andersen, R. (2001). The parietal reach region codes the next planned movement in a sequential reach task. *Journal of Neurophysiology*, 85(2):539–544.
- Batista, A., Buneo, C., Snyder, L., and Andersen, R. (1999). Reach plans in eye-centered coordinates. *Science*, 285(5425):257–260.
- Batista, A. P., Santhanam, G., Byron, M. Y., Ryu, S. I., Afshar, A., and Shenoy, K. V. (2007). Reference frames for reach planning in macaque dorsal premotor cortex. *Journal of neurophysiology*, 98(2):966–983.
- Bays, P. M. and Wolpert, D. M. (2007). Computational principles of sensorimotor control that minimize uncertainty and variability. *The Journal of physiology*, 578(2):387–396.
- Bellan, V., Gilpin, H. R., Stanton, T. R., Newport, R., Gallace, A., and Moseley, G. L. (2015). Untangling visual and proprioceptive contributions to hand localisation over time. *Experimental brain research*, 233(6):1689–1701.
- Bickerton, W. L., Samson, D., Williamson, J., and Humphreys, G. W. (2011). Separating forms of neglect using the apples test: validation and functional prediction in chronic and acute stroke. *Neuropsychology*, 25(5):567.
- Block, A. (1890). Expériences sur les sensations musculaires. *Revue Scientifique*, 45:294–301.
- Blohm, G. and Crawford, J. (2009). Fields of gain in the brain. *Neuron*, 64(5):598–600.
- Blohm, G., Keith, G., and Crawford, J. (2009). Decoding the cortical transformations for visually guided reaching in 3d space. *Cerebral Cortex*, 19(6):1372–1393.

- Bosco, A., Breveglieri, R., Chinellato, E., Galletti, C., and Fattori, P. (2010). Reaching activity in the medial posterior parietal cortex of monkeys is modulated by visual feedback. *The Journal of Neuroscience*, 30(44):14773–14785.
- Botvinick, M., Cohen, J., et al. (1998). Rubber hands’ feel’touch that eyes see. *Nature*, 391(6669):756–756.
- Bowen, A., McKenna, K., and Tallis, R. C. (1999). Reasons for variability in the reported rate of occurrence of unilateral spatial neglect after stroke. *Stroke*, 30(6):1196–1202.
- Bremmer, F., Duhamel, J.-R., Ben Hamed, S., and Graf, W. (2002). Heading encoding in the macaque ventral intraparietal area (vip). *European Journal of Neuroscience*, 16(8):1554–1568.
- Bremner, L. R. and Andersen, R. A. (2012). Coding of the reach vector in parietal area 5d. *Neuron*, 75(2):342–351.
- Bremner, L. R. and Andersen, R. A. (2014). Temporal analysis of reference frames in parietal cortex area 5d during reach planning. *The Journal of Neuroscience*, 34(15):5273–5284.
- Brockmole, J. R., Davoli, C. C., Abrams, R. A., and Witt, J. K. (2013). The world within reach effects of hand posture and tool use on visual cognition. *Current Directions in Psychological Science*, 22(1):38–44.
- Brotchie, P., Andersen, R., Snyder, L., and Goodman, S. (1995). Head position signals used by parietal neurons to encode locations of visual stimuli. *Nature*, -(–):232–235.
- Brozzoli, C., Dumatte, M. L., Pavani, F., Frassinetti, F., and Farne, A. (2006). Neglect and extinction: within and between sensory modalities. *Restorative neurology and neuroscience*, 24(4):217–232.

- Brozzoli, C., Gentile, G., and Ehrsson, H. H. (2012). That's near my hand! parietal and premotor coding of hand-centered space contributes to localization and self-attribution of the hand. *The Journal of Neuroscience*, 32(42):14573–14582.
- Brozzoli, C., Gentile, G., Petkova, V. I., and Ehrsson, H. H. (2011). Fmri adaptation reveals a cortical mechanism for the coding of space near the hand. *The Journal of Neuroscience*, 31(24):9023–9031.
- Brozzoli, C., Pavani, F., Urquizar, C., Cardinali, L., and Farne, A. (2009). Grasping actions remap peripersonal space. *Neuroreport*, 20(10):913–917.
- Buneo, C. and Andersen, R. (2006). The posterior parietal cortex: sensorimotor interface for the planning and online control of visually guided movements. *Neuropsychologia*, 44(13):2594–2606.
- Buneo, C. and Andersen, R. (2012). Integration of target and hand position signals in the posterior parietal cortex: effects of workspace and hand vision. *Journal of neurophysiology*, 108(1):187–199.
- Buneo, C., Jarvis, M., Batista, A., and Andersen, R. (2002). Direct visuomotor transformations for reaching. *Nature*, 416(6881):632–636.
- Buneo, C., Soechting, J., and Squire, L. (2009). Motor psychophysics. *Encyclopedia of neuroscience*, pages 1041–1045.
- Buxbaum, L., Ferraro, M., Veramonti, T., Farne, A., Whyte, J., Ladavas, E., Frassinetti, F., and Coslett, H. (2004). Hemispatial neglect subtypes, neuroanatomy, and disability. *Neurology*, 62(5):749–756.
- Caminiti, R., Ferraina, S., and Johnson, P. (1996). The sources of visual information to the primate frontal lobe: a novel role for the superior parietal lobule. *Cerebral Cortex*, 6(3):319–328.

- Caminiti, R., Innocenti, G. M., and Battaglia-Mayer, A. (2015). Organization and evolution of parieto-frontal processing streams in macaque monkeys and humans. *Neuroscience & Biobehavioral Reviews*, 56:73–96.
- Chang, S., Papadimitriou, C., and Snyder, L. (2009). Using a compound gain field to compute a reach plan. *Neuron*, 64(5):744–755.
- Chang, S. and Snyder, L. (2010). Idiosyncratic and systematic aspects of spatial representations in the macaque parietal cortex. *Proceedings of the National Academy of Sciences*, 107(17):7951–7956.
- Chieffi, S., Allport, D., and Woodin, M. (1999). Hand-centred coding of target location in visuo-spatial working memory. *Neuropsychologia*, 37(4):495–502.
- Cohen, Y. E. and Andersen, R. A. (2002). A common reference frame for movement plans in the posterior parietal cortex. *Nature Reviews Neuroscience*, 3(7):553–562.
- Colby, C. and Goldberg, M. (1999). Space and attention in parietal cortex. *Annual review of neuroscience*, 22(1):319–349.
- Cox, D. D., Meier, P., Oertelt, N., and DiCarlo, J. J. (2005). 'breaking' position-invariant object recognition. *Nature neuroscience*, 8(9):1145–1147.
- Craske, B. and Crawshaw, M. (1975). Shifts in kinesthesia through time and after active and passive movement. *Perceptual and motor skills*, 40(3):755–761.
- Culham, J. and Kanwisher, N. (2001). Neuroimaging of cognitive functions in human parietal cortex. *Current opinion in neurobiology*, 11(2):157–163.
- Derrington, A. and Lennie, P. (1984). Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. *The Journal of Physiology*, 357(1):219–240.

- Desmurget, M., Rossetti, Y., Jordan, M., Meckler, C., and Prablanc, C. (1997). Viewing the hand prior to movement improves accuracy of pointing performed toward the unseen contralateral hand. *Experimental Brain Research*, 115(1):180–186.
- di Pellegrino, G., Làdavas, E., and Farné, A. (1997). Seeing where your hands are. *Nature*, 388(6644):730–730.
- Driver, J. and Halligan, P. W. (1991). Can visual neglect operate in object-centred co-ordinates? an affirmative single-case study. *Cognitive Neuropsychology*, 8(6):475–496.
- Duffy, C. J. (1998). Mst neurons respond to optic flow and translational movement. *Journal of neurophysiology*, 80(4):1816–1827.
- Duhamel, J.-R., Bremmer, F., BenHamed, S., and Graf, W. (1997). Spatial invariance of visual receptive fields in parietal cortex neurons. *Nature*, 389(6653):845–848.
- Duhamel, J.-R., Colby, C. L., and Goldberg, M. E. (1998). Ventral intraparietal area of the macaque: congruent visual and somatic response properties. *Journal of neurophysiology*, 79(1):126–136.
- Eguchi, A., Mender, B. M., Evans, B. D., Humphreys, G. W., and Stringer, S. M. (2015). Computational modeling of the neural representation of object shape in the primate ventral visual system. *Frontiers in computational neuroscience*, 9.
- Einhäuser, W., Schumann, F., Bardins, S., Bartl, K., Böning, G., Schneider, E., and König, P. (2007). Human eye-head co-ordination in natural exploration. *Network: Computation in Neural Systems*, 18(3):267–297.
- Evans, B. D. and Stringer, S. M. (2012). Transformation-invariant visual representations in self-organizing spiking neural networks. *Frontiers in computational neuroscience*, 6.

- Fadiga, L., Caselli, L., Craighero, L., Gesierich, B., Oliynyk, A., Tia, B., and Viaro, R. (2013). Activity in ventral premotor cortex is modulated by vision of own hand in action. *PeerJ*, 1:e88.
- Farnè, A., Demattè, M. L., and Làdavas, E. (2005). Neuropsychological evidence of modular organization of the near peripersonal space. *Neurology*, 65(11):1754–1758.
- Farnè, A., Pavani, F., Meneghello, F., and Làdavas, E. (2000). Left tactile extinction following visual stimulation of a rubber hand. *Brain*, 123(11):2350–2360.
- Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1(1):1–47.
- Ferrera, V. P., Nealey, T. A., and Maunsell, J. (1994). Responses in macaque visual area v4 following inactivation of the parvocellular and magnocellular lgn pathways. *The Journal of neuroscience*, 14(4):2080–2088.
- Ffytche, D. and Zeki, S. (1996). Brain activity related to the perception of illusory contours. *Neuroimage*, 3(2):104–108.
- Filimon, F., Nelson, J., Huang, R., and Sereno, M. (2009). Multiple parietal reach regions in humans: cortical representations for visual and proprioceptive feedback during on-line reaching. *The Journal of Neuroscience*, 29(9):2961–2971.
- Fitzpatrick, D., Lund, J. S., and Blasdel, G. G. (1985). Intrinsic connections of macaque striate cortex: afferent and efferent connections of lamina 4c. *J Neurosci*, 5(12):3329–3349.
- Flanders, M., Tillery, S., and Soechting, J. (1992). Early stages in a sensorimotor transformation. *Behavioral and Brain Sciences*, 15(02):309–320.

- Fogassi, L., Gallese, V., Fadiga, L., Luppino, G., Matelli, M., and Rizzolatti, G. (1996). Coding of peripersonal space in inferior premotor cortex (area f4). *Journal of neurophysiology*, 76(1):141–157.
- Fogassi, L., Raos, V., Franchi, G., Gallese, V., Luppino, G., and Matelli, M. (1999). Visual responses in the dorsal premotor area f2 of the macaque monkey. *Experimental brain research*, 128(1-2):194–199.
- Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3:194–200.
- Freedman, E. G. and Sparks, D. L. (1997). Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys. *Journal of neurophysiology*, 77(5):2328–2348.
- Galeazzi, J. M., Mender, B. M., Paredes, M., Tromans, J. M., Evans, B. D., Minini, L., and Stringer, S. M. (2013). A self-organizing model of the visual development of hand-centred representations. *PloS one*, 8(6):e66272.
- Galeazzi, J. M., Minini, L., and Stringer, S. M. (2015). The development of hand-centered visual representations in the primate brain: A computer modeling study using natural visual scenes. *Frontiers in Computational Neuroscience*, 9:147.
- Galeazzi, J. M., Navajas, J., Mender, B. M. W., Quiroga, R. Q., Minini, L., and Stringer, S. M. The visual development of hand-centered receptive fields in a neural network model of the primate visual system trained with experimentally recorded human gaze changes. submitted.
- Galeazzi, J. M. and Stringer, S. M. (2013). Neural network modelling of hierarchical motor function in the brain. In *Computational and Robotic Models of the Hierarchical Organization of Behavior*, pages 293–317. Springer.

- Galletti, C., Battaglini, P., and Fattori, P. (1995). Eye position influence on the parieto-occipital area po (v6) of the macaque monkey. *European Journal of Neuroscience*, 7(12):2486–2501.
- Galletti, C., Battaglini, P. P., and Fattori, P. (1993). Parietal neurons encoding spatial locations in craniotopic coordinates. *Experimental Brain Research*, 96(2):221–229.
- Galletti, C., Fattori, P., Battaglini, P., Shipp, S., and Zeki, S. (1996). Functional demarcation of a border between areas v6 and v6a in the superior parietal gyrus of the macaque monkey. *European Journal of Neuroscience*, 8(1):30–52.
- Galletti, C., Gamberini, M., Kutz, D., Fattori, P., Luppino, G., and Matelli, M. (2001). The cortical connections of area v6: an occipito-parietal network processing visual information. *European Journal of Neuroscience*, 13(8):1572–1588.
- Gandevia, S., McCloskey, D., and Burke, D. (1992). Kinaesthetic signals and muscle contraction. *Trends in neurosciences*, 15(2):62–65.
- Gattass, R., Gross, C., and Sandell, J. (1981). Visual topography of v2 in the macaque. *Journal of Comparative Neurology*, 201(4):519–539.
- Gentile, G., Petkova, V. I., and Ehrsson, H. H. (2011). Integration of visual and tactile signals from the hand in the human brain: an fmri study. *Journal of Neurophysiology*, 105(2):910–922.
- Glover, S. (2004). Planning and control in action. *Behavioral and brain sciences*, 27(01):57–69.
- Goodale, M. A. and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25.
- Gordon, J., Ghilardi, M. F., Cooper, S. E., and Ghez, C. (1994). Accuracy of planar reaching movements. *Experimental Brain Research*, 99(1):112–130.

- Gozli, D. G., Ardron, J., and Pratt, J. (2014). Reduced visual feature binding in the near-hand space. *Attention, Perception, & Psychophysics*, 76(5):1308–1317.
- Gozli, D. G., West, G. L., and Pratt, J. (2012). Hand position alters vision by biasing processing through different visual pathways. *Cognition*, 124(2):244–250.
- Graziano, M. S. (1999). Where is my arm? the relative role of vision and proprioception in the neuronal representation of limb position. *Proceedings of the National Academy of Sciences*, 96(18):10418–10421.
- Graziano, M. S., Andersen, R. A., and Snowden, R. J. (1994a). Tuning of mst neurons to spiral motions. *The Journal of neuroscience*, 14(1):54–67.
- Graziano, M. S. and Botvinick, M. (2002). How the brain represents the body: insights from neurophysiology and psychology. *Common mechanisms in perception and action: Attention and performance XIX*, pages 136–157.
- Graziano, M. S. and Cooke, D. F. (2006). Parieto-frontal interactions, personal space, and defensive behavior. *Neuropsychologia*, 44(6):845–859.
- Graziano, M. S., Cooke, D. F., and Taylor, C. S. (2000). Coding the location of the arm by sight. *Science*, 290(5497):1782–1786.
- Graziano, M. S. and Gross, C. G. (1993). A bimodal map of space: somatosensory receptive fields in the macaque putamen with corresponding visual receptive fields. *Experimental Brain Research*, 97(1):96–109.
- Graziano, M. S. and Gross, C. G. (1998). Spatial maps for the control of movement. *Current opinion in neurobiology*, 8(2):195–201.
- Graziano, M. S., Gross, C. G., Taylor, C. S., and Moore, T. (2004). A system of multimodal areas in the primate brain. *Crossmodal space and crossmodal attention*, pages 51–67.

- Graziano, M. S., Hu, X., and Gross, C. (1997). Visuospatial properties of ventral premotor cortex. *Journal of Neurophysiology*, 77(5):2268–2292.
- Graziano, M. S., Yap, G. S., and Gross, C. G. (1994b). Coding of visual space by premotor neurons. *Science*, pages 1054–1054.
- Gross, C. G. (2008). Single neuron studies of inferior temporal cortex. *Neuropsychologia*, 46(3):841–852.
- Halligan, P. W. and Marshall, J. C. (1991). Left neglect for near but not far space in man. *Nature*, 350(6318):498–500.
- Hebb, D. (1949). The organization of behavior; a neuropsychological theory.
- Heilman, K. M. and Valenstein, E. (2000). Mechanisms underlying hemispatial neglect. *Brain and Behaviour: Critical Concepts in Psychology*, 4(2):1279.
- Hertz, J., Krogh, A., and Palmer, R. (1991). *Introduction to the theory of neural computation*, volume 1. Westview press.
- Himmelbach, M. and Karnath, H.-O. (2005). Dorsal and ventral stream interaction: contributions from optic ataxia. *Journal of Cognitive Neuroscience*, 17(4):632–640.
- Holmes, N. P. (2013). Hand-centered space, hand-centered attention, and the control of movement. In *The Hand, an Organ of the Mind: What the Manual Tells the Mental*, page 57. MIT Press.
- Holmes, N. P. and Spence, C. (2005). Visual bias of unseen hand position with a mirror: spatial and temporal factors. *Experimental Brain Research*, 166(3-4):489–497.
- Horner, J., Massey, E. W., Woodruff, W. W., Chase, K. N., and Dawson, D. V. (1989). Task-dependent neglect: computed tomography size and locus correlations. *Neurorehabilitation and Neural Repair*, 3(1):7–13.

- Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106.
- Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243.
- Humphreys, G. W., Gillebert, C. R., Chechlacz, M., and Riddoch, M. J. (2013). Reference frames in visual selection. *Annals of the New York Academy of Sciences*, 1296(1):75–87.
- Hyvärinen, J. and Poranen, A. (1974). Function of the parietal associative area 7 as revealed from cellular discharges in alert monkeys. *Brain*, 97(4):673–692.
- Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of neurophysiology*, 73(1):218–226.
- Jackson, H. (1875). London hospital.: Clinical observations on cases of disease of the nervous system. *The Lancet*, 105(2681):85.
- Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, 58(6):1233–1258.
- Jones, L. A. and Lederman, S. J. (2006). *Human hand function*. Oxford University Press.
- Kaas, J. (2004). Evolution of somatosensory and motor cortex in primates. *The Anatomical Record Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology*, 281(1):1148–1156.
- Kandel, E. R., Schwartz, J. H., Jessell, T. M., et al. (2000). *Principles of neural science*, volume 4. McGraw-Hill New York.

- Karnath, H.-O. and Rorden, C. (2012). The anatomy of spatial neglect. *Neuropsychologia*, 50(6):1010–1017.
- Kellenbach, M. L., Brett, M., and Patterson, K. (2001). Large, colorful, or noisy? attribute-and modality-specific activations during retrieval of perceptual attribute knowledge. *Cognitive, Affective, & Behavioral Neuroscience*, 1(3):207–221.
- Kobatake, E. and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of neurophysiology*, 71(3):856–867.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69.
- Kravitz, D. J., Saleem, K. S., Baker, C. I., and Mishkin, M. (2011). A new neural framework for visuospatial processing. *Nature Reviews Neuroscience*, 12(4):217–230.
- Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., and Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in cognitive sciences*, 17(1):26–49.
- Lacquaniti, F., Guigon, E., Bianchi, L., Ferraina, S., and Caminiti, R. (1995). Representing spatial information for limb movement: role of area 5 in the monkey. *Cerebral Cortex*, 5(5):391–409.
- Làdavas, E., Di Pellegrino, G., Farnè, A., and Zeloni, G. (1998). Neuropsychological evidence of an integrated visuotactile representation of peripersonal space in humans. *Journal of Cognitive Neuroscience*, 10(5):581–589.
- Làdavas, E., Farnè, A., Zeloni, G., and di Pellegrino, G. (2000). Seeing or not seeing where your hands are. *Experimental Brain Research*, 131(4):458–467.

- Land, M., Mennie, N., Rusted, J., et al. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception-London*, 28(11):1311–1328.
- Latash, M. L. (2012). The bliss (not the problem) of motor abundance (not redundancy). *Experimental brain research*, 217(1):1–5.
- Lehky, S. R. and Sereno, A. B. (2007). Comparison of shape encoding in primate dorsal and ventral visual pathways. *Journal of neurophysiology*, 97(1):307–319.
- Lennie, P. (2003). Receptive fields. *Current Biology*, 13(6):R216–R219.
- Leoné, F. T., Monaco, S., Henriques, D. Y., Toni, I., and Medendorp, W. P. (2015). Flexible reference frames for grasp planning in human parieto-frontal cortex. *eneuro*, pages ENEURO–0008.
- Lewis, J. W. and Van Essen, D. C. (2000). Mapping of architectonic subdivisions in the macaque monkey, with emphasis on parieto-occipital cortex. *Journal of Comparative Neurology*, 428(1):79–111.
- Li, N. and DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *science*, 321(5895):1502–1507.
- Livingstone, M. S. and Hubel, D. H. (1984). Anatomy and physiology of a color system in the primate visual cortex. *J Neurosci*, 4(1):309–356.
- MacKay, D. J. (2003). *Information theory, inference, and learning algorithms*, volume 7. Citeseer.
- Magosso, E., Zavaglia, M., Serino, A., Di Pellegrino, G., and Ursino, M. (2008). A neural network model of peri-hand space representation and its plastic properties related to tool use. In *BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on*, pages 1–6. IEEE.

- Magosso, E., Zavaglia, M., Serino, A., Di Pellegrino, G., and Ursino, M. (2010). Visuo-tactile representation of peripersonal space: a neural network study. *Neural computation*, 22(1):190–243.
- Makin, T. R., Brozzoli, C., Cardinali, L., Holmes, N. P., and Farnè, A. (2015). Left or right? rapid visuomotor coding of hand laterality during motor decisions. *Cortex*, 64:289–292.
- Makin, T. R., Holmes, N. P., Brozzoli, C., and Farnè, A. (2012). Keeping the world at hand: rapid visuomotor processing for hand–object interactions. *Experimental brain research*, 219(4):421–428.
- Makin, T. R., Holmes, N. P., Brozzoli, C., Rossetti, Y., and Farne, A. (2009). Coding of visual space during motor preparation: approaching objects rapidly modulate corticospinal excitability in hand-centered coordinates. *The Journal of Neuroscience*, 29(38):11841–11851.
- Makin, T. R., Holmes, N. P., and Ehrsson, H. H. (2008). On the other hand: dummy hands and peripersonal space. *Behavioural brain research*, 191(1):1–10.
- Makin, T. R., Holmes, N. P., and Zohary, E. (2007). Is that near my hand? multisensory representation of peripersonal space in human intraparietal sulcus. *The Journal of neuroscience*, 27(4):731–740.
- Maravita, A., Spence, C., Sergent, C., and Driver, J. (2002). Seeing your own touched hands in a mirror modulates cross-modal interactions. *Psychological Science*, 13(4):350–355.
- Mather, J. A. and Fisk, J. D. (1985). Orienting to targets by looking and pointing: parallels and interactions in ocular and manual performance. *The Quarterly Journal of Experimental Psychology*, 37(3):315–338.

- Maunsell, J. H. and Van Essen, D. C. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey. i. selectivity for stimulus direction, speed, and orientation. *Journal of neurophysiology*, 49(5):1127–1147.
- Mazzoni, P., Andersen, R. A., and Jordan, M. I. (1991). A more biologically plausible learning rule for neural networks. *Proceedings of the National Academy of Sciences*, 88(10):4433–4437.
- McGuire, L. M. and Sabes, P. N. (2009). Sensory transformations and the use of multiple reference frames for reach planning. *Nature neuroscience*, 12(8):1056–1061.
- McGuire, L. M. and Sabes, P. N. (2011). Heterogeneous representations in the superior parietal lobule are common across reaches to visual and proprioceptive targets. *The Journal of Neuroscience*, 31(18):6661–6673.
- McIntyre, J., Stratta, F., and Lacquaniti, F. (1998). Short-term memory for reaching to visual targets: psychophysical evidence for body-centered reference frames. *The Journal of Neuroscience*, 18(20):8423–8435.
- Mender, B. M. and Stringer, S. M. (2013). A model of self-organizing head-centered visual responses in primate parietal areas. *PLoS ONE*, 8:81406.
- Mender, B. M. and Stringer, S. M. (2014). Self-organization of head-centered visual responses under ecological training conditions. *Network: Computation in Neural Systems*, 25(3):116–136.
- Merigan, W. H. and Maunsell, J. H. (1993). How parallel are the primate visual pathways? *Annual review of neuroscience*, 16(1):369–402.
- Miall, R., Haggard, P., and Cole, J. (1995). Evidence of a limited visuo-motor memory used in programming wrist movements. *Experimental brain research*, 107(2):267–280.

- Miall, R. C. and Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural networks*, 9(8):1265–1279.
- Milner, A. D. and Goodale, M. A. (1993). Visual pathways to perception and action. *Progress in brain research*, 95:317–317.
- Milner, A. D. and Goodale, M. A. (2008). Two visual systems re-viewed. *Neuropsychologia*, 46(3):774–785.
- Mountcastle, V., Lynch, J., Georgopoulos, A., Sakata, H., and Acuna, C. (1975). Posterior parietal association cortex of the monkey: command functions for operations within extrapersonal space. *Journal of Neurophysiology*, 38(4):871–908.
- 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.
- Newport, R. and Gilpin, H. R. (2011). Multisensory disintegration and the disappearing hand trick. *Current biology*, 21(19):R804–R805.
- O’dhaniel, A., Cohen, Y. E., and Groh, J. M. (2005). Eye-centered, head-centered, and complex coding of visual and auditory targets in the intraparietal sulcus. *Journal of Neurophysiology*, 94(4):2331–2352.
- Oliver, R. T. and Thompson-Schill, S. L. (2003). Dorsal stream activation during retrieval of object size and shape. *Cognitive, Affective, & Behavioral Neuroscience*, 3(4):309–322.
- O’Regan, J. K. (1992). Solving the” real” mysteries of visual perception: the world as an outside memory. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 46(3):461.

- Paillard, J. and Brouchon, M. (1968). Active and passive movements in the calibration of position sense. *The neuropsychology of spatially oriented behavior*, 11:37–55.
- Pandya, D. N. and Seltzer, B. (1982). Intrinsic connections and architectonics of posterior parietal cortex in the rhesus monkey. *Journal of Comparative Neurology*, 204(2):196–210.
- Pasupathy, A. and Connor, C. E. (2001). Shape representation in area v4: position-specific tuning for boundary conformation. *Journal of neurophysiology*, 86(5):2505–2519.
- Pasupathy, A. and Connor, C. E. (2002). Population coding of shape in area v4. *Nature neuroscience*, 5(12):1332–1338.
- Pavani, F., Spence, C., and Driver, J. (2000). Visual capture of touch: Out-of-the-body experiences with rubber gloves. *Psychological science*, 11(5):353–359.
- Pelz, J., Hayhoe, M., and Loeber, R. (2001). The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research*, 139(3):266–277.
- Penfield, W. and Rasmussen, T. (1950). The cerebral cortex of man; a clinical study of localization of function.
- Perry, C. J. and Fallah, M. (2014). Feature integration and object representations along the dorsal stream visual hierarchy. *Frontiers in computational neuroscience*, 8.
- Perry, G., Rolls, E., and Stringer, S. (2010). Continuous transformation learning of translation invariant representations. *Experimental Brain Research*, 204(2):255–270.
- Perry, G., Rolls, E. T., and Stringer, S. M. (2006). Spatial vs temporal continuity in view invariant visual object recognition learning. *Vision Research*, 46(23):3994–4006.
- Pesaran, B., Nelson, M., and Andersen, R. (2006). Dorsal premotor neurons encode the relative position of the hand, eye, and goal during reach planning. *Neuron*, 51(1):125.

- Petreska, B., Adriani, M., Blanke, O., and Billard, A. G. (2007). Apraxia: a review. *Progress in brain research*, 164:61–83.
- Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. *Nat Rev Neurosci*, 1(2):125–132.
- Pouget, A., Deneve, S., and Duhamel, J.-R. (2002). A computational perspective on the neural basis of multisensory spatial representations. *Nature Reviews Neuroscience*, 3(9):741–747.
- Pouget, A. and Driver, J. (2000). Relating unilateral neglect to the neural coding of space. *Current opinion in neurobiology*, 10(2):242–249.
- Pouget, A. and Sejnowski, T. (1997). Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience*, 9(2):222–237.
- Prablanc, C., Pelisson, D., and Goodale, M. (1986). Visual control of reaching movements without vision of the limb. *Experimental Brain Research*, 62(2):293–302.
- Prevosto, V., Graf, W., and Ugolini, G. (2011). Proprioceptive pathways to posterior parietal areas mip and lipv from the dorsal column nuclei and the postcentral somatosensory cortex. *European Journal of Neuroscience*, 33(3):444–460.
- Proske, U. and Gandevia, S. C. (2012). The proprioceptive senses: their roles in signaling body shape, body position and movement, and muscle force. *Physiological reviews*, 92(4):1651–1697.
- Ribot-Ciscar, E., Bergenheim, M., Albert, F., and Roll, J.-P. (2003). Proprioceptive population coding of limb position in humans. *Experimental brain research*, 149(4):512–519.

- Rizzolatti, G., Camarda, R., Fogassi, L., Gentilucci, M., Luppino, G., and Matelli, M. (1988). Functional organization of inferior area 6 in the macaque monkey. *Experimental brain research*, 71(3):491–507.
- Rizzolatti, G. and Matelli, M. (2003). Two different streams form the dorsal visual system: anatomy and functions. *Experimental Brain Research*, 153(2):146–157.
- Rizzolatti, G., Scandolara, C., Matelli, M., and Gentilucci, M. (1981). Afferent properties of periarculate neurons in macaque monkeys. ii. visual responses. *Behavioural brain research*, 2(2):147–163.
- Rolls, E. T. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical areas. *Philosophical Transactions of the Royal Society, London [B]*, 335:11–21.
- Rolls, E. T. (2008). *Memory, Attention, and Decision-Making: A Unifying Computational Neuroscience Approach*. OUP Oxford, 1 edition.
- Rolls, E. T. (2012a). Advantages of dilution in the connectivity of attractor networks in the brain. *Biologically Inspired Cognitive Architectures*, 1:44–54.
- Rolls, E. T. (2012b). Invariant visual object and face recognition: neural and computational bases, and a model, visnet. *Frontiers in Computational Neuroscience*, 6.
- Rolls, E. T. and Milward, T. (2000). A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition and information-based performance measures. *Neural Computation*, 12:2547–2572.
- Rolls, E. T. and Stringer, S. M. (2001). Invariant object recognition in the visual system with error correction and temporal difference learning. *Network: Computation in Neural Systems*, 12(2):111–129.

- Rolls, E. T. and Stringer, S. M. (2007). Invariant global motion recognition in the dorsal visual system: a unifying theory. *Neural computation*, 19(1):139–169.
- Rolls, E. T. and Treves, A. (1990). The relative advantages of sparse versus distributed encoding for associative neuronal networks in the brain. *Network: computation in neural systems*, 1(4):407–421.
- Rolls, E. T. and Treves, A. (1998). *Neural networks and brain function*. Oxford University Press Oxford.
- Rolls, E. T., Treves, A., Tovee, M. J., and Panzeri, S. (1997). Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. 4:309–333.
- Rolls, E. T. and Webb, T. J. (2012). Cortical attractor network dynamics with diluted connectivity. *Brain research*, 1434:212–225.
- Rolls, E. T. and Webb, T. J. (2014). Finding and recognizing objects in natural scenes: complementary computations in the dorsal and ventral visual systems. *Frontiers in computational neuroscience*, 8.
- Rossetti, Y., Desmurget, M., and Prablanc, C. (1995). Vectorial coding of movement: vision, proprioception, or both? *Journal of Neurophysiology*, 74(1):457–463.
- Rumelhart, D. E. and Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive science*, 9(1):75–112.
- Sakata, H., Takaoka, Y., Kawarasaki, A., and Shibutani, H. (1973). Somatosensory properties of neurons in the superior parietal cortex (area 5) of the rhesus monkey. *Brain Research*, 64:85–102.
- Sarlegna, F. and Sainburg, R. (2009). The roles of vision and proprioception in the planning of reaching movements. *Progress in Motor Control*, pages 317–335.

- Saunders, J. and Knill, D. (2003). Humans use continuous visual feedback from the hand to control fast reaching movements. *Experimental Brain Research*, 152(3):341–352.
- Saunders, J. and Knill, D. (2005). Humans use continuous visual feedback from the hand to control both the direction and distance of pointing movements. *Experimental Brain Research*, 162(4):458–473.
- Scott, S. H. (2012). The computational and neural basis of voluntary motor control and planning. *Trends in cognitive sciences*, 16(11):541–549.
- Sereno, A. and Maunsell, J. (1998). Shape selectivity in primate lateral intraparietal cortex. *Nature*, 395(6701):500–503.
- Shadmehr, R., Smith, M. A., and Krakauer, J. W. (2010). Error correction, sensory prediction, and adaptation in motor control. *Annual review of neuroscience*, 33:89–108.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell Technical System Journal*, 27(3):379–423 and 623–656.
- Sherrington, C. (1907). On the proprioceptive system, especially in its reflex aspect. *Brain*, (29):467–482.
- Shipp, S. and Zeki, S. (2002). The functional organization of area v2, ii: the impact of stripes on visual topography. *Visual neuroscience*, 19(02):211–231.
- Snyder, L., Batista, A., and Andersen, R. (2000). Intention-related activity in the posterior parietal cortex: a review. *Vision research*, 40(10-12):1433–1441.
- Snyder, L. H., Grieve, K. L., Brotchie, P., and Andersen, R. A. (1998). Separate body-and world-referenced representations of visual space in parietal cortex. *Nature*, 394(6696):887–891.
- Sober, S. and Sabes, P. (2003). Multisensory integration during motor planning. *The Journal of Neuroscience*, 23(18):6982–6992.

- Sober, S. and Sabes, P. (2005). Flexible strategies for sensory integration during motor planning. *Nature neuroscience*, 8(4):490–497.
- Soechting, J. and Flanders, M. (1992). Moving in three-dimensional space: frames of reference, vectors, and coordinate systems. *Annual review of neuroscience*, 15(1):167–191.
- Spratling, M. W. (2009). Learning posture invariant spatial representations through temporal correlations. *Autonomous Mental Development, IEEE Transactions on*, 1(4):253–263.
- Stringer, S., Perry, G., Rolls, E., and Proske, J. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biological cybernetics*, 94(2):128–142.
- Stringer, S. and Rolls, E. (2000). Position invariant recognition in the visual system with cluttered environments. *Neural Networks*, 13(3):305–315.
- Stringer, S. and Rolls, E. (2006). Self-organizing path integration using a linked continuous attractor and competitive network: Path integration of head direction. *Network: Computation in Neural Systems*, 17(4):419–445.
- Stringer, S. and Rolls, E. (2008). Learning transform invariant object recognition in the visual system with multiple stimuli present during training. *Neural Networks*, 21(7):888–903.
- Stringer, S., Rolls, E., and Tromans, J. (2007). Invariant object recognition with trace learning and multiple stimuli present during training. *Network: Computation in Neural Systems*, 18(2):161–187.
- Takemura, H., Rokem, A., Winawer, J., Yeatman, J. D., Wandell, B. A., and Pestilli, F.

- (2015). A major human white matter pathway between dorsal and ventral visual cortex. *Cerebral Cortex*, page bhv064.
- Thomaschke, R., Hopkins, B., and Miall, R. C. (2012). The planning and control model (pcm) of motorvisual priming: reconciling motorvisual impairment and facilitation effects. *Psychological review*, 119(2):388.
- Tipper, S. P., Lortie, C., and Baylis, G. C. (1992). Selective reaching: evidence for action-centered attention. *Journal of Experimental Psychology: Human Perception and Performance*, 18(4):891.
- Todorov, E. and Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature neuroscience*, 5(11):1226–1235.
- Treves, A. (1991). Dilution and sparse coding in threshold-linear nets. *Journal of Physics A: Mathematical and General*, 24(1):327.
- Tromans, J. (2012). *Computational neuroscience of natural scene processing in the ventral visual pathway*. PhD thesis, University of Oxford.
- Tromans, J., Page, H., and Stringer, S. (2012). Learning separate visual representations of independently rotating objects. *Network: Computation in Neural Systems*, 23(1-2):1–23.
- Tsakiris, M. and Haggard, P. (2005). The rubber hand illusion revisited: visuotactile integration and self-attribution. *Journal of Experimental Psychology: Human Perception and Performance*, 31(1):80.
- Ungerleider, L. G. and Haxby, J. V. (1994). What and wherein the human brain. *Current opinion in neurobiology*, 4(2):157–165.
- Ungerleider, L. G. and Mishkin, M. (1982). Two cortical visual systems. *Analysis of visual behavior*, pages 549–586.

- van Beers, R., Sittig, A. C., and van der Gon, J. J. D. (1998). The precision of proprioceptive position sense. *Experimental Brain Research*, 122(4):367–377.
- van Beers, R. J., Wolpert, D. M., Haggard, P., et al. (2002). When feeling is more important than seeing in sensorimotor adaptation. *Current Biology*, 12(10):834–837.
- Von der Heydt, R., Peterhans, E., and Baumgartner, G. (1990). Neuronal processing of contours. *Physiology*, 5(4):152–155.
- von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14(2):85–100.
- Vuilleumier, P., Valenza, N., Mayer, E., Reverdin, A., and Landis, T. (1998). Near and far visual space in unilateral neglect. *Annals of neurology*, 43(3):406–410.
- Wallis, G. M. and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, 51(2):167–194.
- Wann, J. P. and Ibrahim, S. F. (1992). Does limb proprioception drift? *Experimental Brain Research*, 91(1):162–166.
- Webster, M. J., Bachevalier, J., and Ungerleider, L. G. (1994). Connections of inferior temporal areas teo and te with parietal and frontal cortex in macaque monkeys. *Cerebral cortex*, 4(5):470–483.
- Weiss, P. H., Marshall, J. C., Wunderlich, G., Tellmann, L., Halligan, P. W., Freund, H.-J., Zilles, K., and Fink, G. R. (2000). Neural consequences of acting in near versus far space: a physiological basis for clinical dissociations. *Brain*, 123(12):2531–2541.
- Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences.
- Wheaton, L. A. and Hallett, M. (2007). Ideomotor apraxia: a review. *Journal of the neurological sciences*, 260(1):1–10.

- Widrow, B., Hoff, M. E., et al. (1960). Adaptive switching circuits.
- Wise, S., Boussaoud, D., Johnson, P., and Caminiti, R. (1997). Premotor and parietal cortex: Corticocortical connectivity and combinatorial computations 1. *Annual review of neuroscience*, 20(1):25–42.
- Wolpert, D. M. (1997). Computational approaches to motor control. *Trends in cognitive sciences*, 1(6):209–216.
- Wolpert, D. M. and Ghahramani, Z. (2000). Computational principles of movement neuroscience. *nature neuroscience*, 3:1212–1217.
- Xing, J., Andersen, R., et al. (2000). Models of the posterior parietal cortex which perform multimodal integration and represent space in several coordinate frames. *Cognitive Neuroscience, Journal of*, 12(4):601–614.
- Yarbus, A. L., Haigh, B., and Riggs, L. A. (1967). *Eye movements and vision*, volume 2. Plenum press New York.
- Zeki, S. M. (1973). Colour coding in rhesus monkey prestriate cortex. *Brain research*, 53(2):422–427.
- Zipser, D. and Andersen, R. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, 331(6158):679–684.