Neural Network Modelling of the Primate Ventral Visual Pathway

Akihiro Eguchi

St. Catherine's College
University of Oxford

A thesis submitted for the degree of
DPhil in Experimental Psychology

Hilary Term 2017

Acknowledgements

This doctoral thesis would have not been possible without the valued support of a number of individuals to whom I am sincerely thankful.

Firstly, I would like to thank my supervisor, Dr. Simon Stringer, for his continued encouragement, moral support, and contributions. His boundless optimism and enthusiasm swept aside any self-doubts or apprehensions on countless occasions to keep me running. Also, I would like to express my sincere gratitude to Professor Glyn Humphreys, who was also my supervisor, and unfortunately was deceased on January 14th 2016. He always gave me a great insight into answering various difficult research questions I have tackled. It was undoubtedly a great honour working with him.

Secondly, I must thank my research colleagues who have provided a great source of academic consultations and inspirations. I would like to thank Bedeho M. for the development of the VisNet model that I used in many simulation studies in this thesis; Ben E., for his constant support and advice for the development of the spiking model; James T. for his advice at the beginning of my years as DPhil; James I. and Nasir A. for the development of the GPGPU library Spike! that made my spiking model run much faster; Irina H. for giving insightful advice for the analytic methods of polychronization; Thomas M. and Courtney S. for being enthusiastic students and satisfying to mentor; Daniel W., Juan G., Hector P., Daniel N., Harry J., Jannis B., Hannah D., Mihaela M., and Mickael L. for making my days at the lab cheerful and inspirational through casual discussions about the research in general and many other different topics.

Thirdly, I wish to acknowledge the financial support provided by the Oxford Foundation for Theoretical Neuroscience and Artificial Intelligence (OFTNAI) throughout my DPhil course. Also, I would like to thank Prof. Mark Buckley and Prof. Hannah Smithson for serving on a committee of my transfer of status and my confirmation of status to keep track of the progress of my research, and Prof. Guy Wallis for evaluating my final thesis.

Finally, a special note of thanks to Mirai, who has always been by my side and understanding even when the inevitable frustration from the agony of the repeated failures on the way has fallen upon her shoulders. Last, but certainly not least, I offer immense gratitude to my parents for their unwavering support, not only during my doctoral studies, but along every step of the journey that has led me to this point. There is absolutely no chance I would have written thesis without you.

Short Abstract

The neural representation of object shape in the primate ventral visual system
Akihiro Eguchi
St. Catherine's College
Submitted for the Degree of DPhil in Experimental Psychology
Hilary Term 2017

The aim of this doctoral research is to advance understanding of how the primate brain learns to process the detailed spatial form of natural visual scenes. Neurons in successive stages of the primate ventral visual pathway encode the spatial structure of visual objects and faces. However, it remains a difficult challenge to understand exactly how these neurons develop their response properties through visually guided learning. This thesis approaches this problem through the use of computational modelling. In particular, I first show how the brain may learn to represent the spatial structure of objects and faces through a series of processing stages along the ventral visual pathway. Then I propose how understanding the two complementary unsupervised learning mechanisms of translation invariance may have useful applications in clinical psychology. Next, the potential functional role of top-down (feedback) propagation of visual information in the brain in driving the development of border ownership cells, which are thought to play a role in binding visual features such as boundary edges to their respective objects, is investigated. In particular, the limitations of traditional rate-coded neural networks in modelling these cells are identified. Finally, a general solution to such binding problems with the use of a more biologically realistic spiking neural network is presented. This work is set to make an important contribution towards understanding how the visual system learns to encode the detailed spatial structure of objects and faces within scenes, including representing the binding relations between the visual features that comprise those objects and faces.

Long Abstract

The neural representation of object shape in the primate ventral visual system
Akihiro Eguchi
St. Catherine's College
Submitted for the Degree of DPhil in Experimental Psychology
Hilary Term 2017

The major challenge facing visual neuroscience today is discovering how the brain learns to represent the full complexity of natural visual images. Single unit recording in non-human primates has shed light on how the visual system encodes the spatial structure of objects and faces, showing that representations develop through a hierarchy of neural layers within the ventral visual pathway with visual information represented in a distributed manner across brain areas, which communicate with each other through bottom-up (feedforward) and top-down (feedback) connections. Consequently, understanding how neurons develop their firing properties, and thereby learn to represent visual scenes, will require detailed neural network computer models. This thesis presents a series of computer modelling studies that is aimed at advancing our understanding of how objects and faces are encoded through these linked areas.

The first simulation study explores how representations of simple shape fragments might be developed and are later integrated to represent more complex object shapes with a completely unsupervised learning mechanism and feed-forward processing in a neural network model, VisNet, of the primate ventral visual pathway. In particular, it is demonstrated that when VisNet is trained on many objects with different boundary shapes, the neurons in the higher layers of the network learn to respond to localised boundary contour elements, which are defined by the curvature and location of the boundary element in the frame of reference of the object. Interestingly, our result shows that neurons learn to respond to these boundary elements rather than learning to respond to the whole objects that are actually presented during training. Moreover, the neurons are able to learn to respond with translation invariance as visual objects are shifted across different retinal locations. This is shown to be successful when VisNet is trained with either the artificially constructed visual stimuli or with images of natural visual objects. A population of such neurons, representing many different boundary elements of different curvature and position within the object, could be used to provide a distributed coding of the entire boundary shape of an object. This has been demonstrated with real neurons in primate visual area V4. As such, these neurons are likely to play an important role in how the primate visual system represents the shapes of objects.

The following chapter presents neural network simulations of the visually-guided development of facial representations in VisNet. In particular, the followings are presented: (1) how some neurons along the successive stages of processing learn to represent individual facial features such as the eyes, nose, and mouth even though the visual system is always exposed to whole faces, (2) how some neurons learn to represent particular spatial relationships between facial features, such as the distance between the eyes, with monotonic tuning curves, (3) how some neurons in later stages learn to respond to global attributes such as either a particular identity or expression, and (4) what is the relationship between spatial configurations of facial parts and global representations of face identity and expression. This work represents an important theoretical advance in understanding how the visual system learns to represent the rich spatial structure of the faces, contradicting many current accounts of learning based on the feedback of error signals from higher- to lower-levels of representation.

Then, based on the findings in the previous chapters, a clinical treatment for depression known as Cognitive Bias Modification (CBM) is modelled. This family of treatments seek to eliminate underlying cognitive biases towards negative information and are efficacious in reducing the recurrence of depression. However, the mechanisms behind the bias elimination are not fully understood. The study conducted in this chapter investigates, through computer simula-

tion of neural network models, the neural dynamics underlying the use of CBM in eliminating the negative biases in the way that depressed patients evaluate facial expressions. Two CBM methodologies using biologically plausible synaptic learning mechanisms are investigated: Continuous Transformation (CT) learning and trace learning, which guide learning by exploiting either the spatial or temporal continuity between visual stimuli presented during training. Simulations with a simplified one-layer neural network are describe first, and then simulations in VisNet are described. After training with either CT learning or trace learning, the one-layer neural network eliminated biases in interpreting neutral stimuli as sad. The multi-layer neural network trained with realistic face stimuli was also shown to be able to use CT learning or trace learning in order to reduce biases in the interpretation of neutral facial expressions as sad. The simulation results suggest two biologically plausible synaptic learning mechanisms, CT learning and trace learning, that may subserve CBM. The results are highly informative for the development of experimental protocols to produce optimal CBM training methodologies with human participants.

The following chapter investigates the potential limitation of a traditional rate-coded model such as VisNet when investigating visual feature binding in the form of border ownership representations. Our visual perception tends to assign luminance contrast borders to one or other of the adjacent image regions. Experimental evidence for the neuronal coding of such border-ownership in the primate visual system has been reported in neurophysiology. It is investigated exactly how such neural circuits may develop through visually-guided learning. More specifically, our VisNet model has been modified to include both bottom-up and top-down synaptic connections between successive layers to investigate how top-down connections may play a fundamental role in the development of border ownership representations in the early cortical visual layers V1/V2. The simulations reported in this chapter demonstrate that top-down connections may help to guide competitive learning in lower layers, thus driving the formation of lower level (border ownership) visual representations in V1/V2 that are modulated by higher level (object boundary element) representations in V4. However, additional simulations conducted in this chapter reveal a crucial limitation of such rate-coded models in the more general situation where multiple objects are presented to the network simultaneously. In this more challenging situation, the border ownership cells simulated in the rate-coded model fail to maintain their proper firing properties.

Accordingly, the final two chapters present a potential solution for such limitations by implementing the model with spiking dynamics. In particular, a hierarchical neural network model, in which subpopulations of neurons develop fixed and regularly repeating temporal chains of spikes (polychronization), which respond specifically to randomised Poisson spike trains representing the input images, is presented first. The performance is improved by including top-down and lateral synaptic connections, as well as introducing multiple synaptic contacts between each pair of pre- and postsynaptic neurons with different synaptic contacts having different axonal transmission delays. In the latter case, Spike-Timeing-Dependent Plasticity (STDP) allows the model to select the most effective axonal transmission delay between each pair of pre- and post-synaptic neurons. Furthermore, neurons representing the binding relationships between low-level and high-level visual features emerge through visually-guided learning. This provides a solution to the classic feature binding problem in visual neuroscience and leads to a new hypothesis concerning how information about visual features at every spatial scale may be projected upwards through successive neuronal layers.

Then finally the last chapter of simulations investigates the development of robust representations of border ownership information in the early cortical layers of the spiking model. These last simulations provide computational evidence supporting the hypothesised solution to the binding problem proposed in the previous chapter. More precisely, the limitations of the "superposition catastrophe" in a traditional rate-coded model are overcome within the current spiking neural network model, where the border ownership representations are robust even when

multiple objects are presented to the network simultaneously.

Contents

Chapter 1

Introduction

Over successive stages, the primate ventral visual pathway develops neurons that respond to particular objects or faces independently of their position, size, or orientation (Perrett et al., 1982; Desimone, 1991; Tanaka et al., 1991). The ventral visual pathway is thus thought to be responsible for transform-invariant visual object and face recognition in the brain. However, it remains a difficult challenge to understand exactly how these neurons develop their response properties during learning. The learning processes will depend on how the neurons interact with each other through successive layers of the ventral visual pathway as they are driven by rich visual input from natural scenes. This can be investigated through computer simulations that accurately model the behaviour of individual neurons, how these neurons are linked together in the brain, how the synaptic connections between cells are modified during learning, and the statistical properties of the visual input from the sensory environment. The aim of this doctoral research is to advance understanding of how the primate brain learns to process visual input from natural scenes. This problem goes beyond mere object recognition to understanding how the visual system represents the full complexity of natural visual scenes. This is a profoundly important and timely issue for achieving a more complete understanding of primate vision. Understanding how complex visual scenes are represented in the brain will help to inform clinical treatment of patients with damage to these brain areas, and perhaps lead to a new generation of computer vision systems that are more effective at interpreting natural scenes.

1.1 Visual Processing along the Primate Ventral Visual Pathway

The ability of the brain to analyse and recognize objects under natural viewing conditions is unmatched by today's computer vision systems. In order to achieve this singular ability, the primate brain develops and utilizes a rich tapestry of cells that encode different kinds of visual information. At the retina as the entrance of the visual information to the brain, spatial contrasts that will be useful for the later perception are extracted from two dimensional inputs of brightness and colours detected by photoreceptors. The information about the spatial contrasts at every location around the visual field serves as the important set of features that is to be processed in the later stages. Such feature information is propagated to the cerebral cortex and integrated together along the visual pathway. As a result, over successive stages of processing, the primate ventral visual pathway develops neurons that respond selectively to objects of increasingly complex visual form (Kobatake and Tanaka, 1994), going from simple orientated line segments in area V1 (Hubel and Wiesel, 1962) to whole objects or faces in the inferotemporal cortex (IT) (Perrett et al., 1982; Tsao et al., 2003; Tsunoda et al., 2001). This section gives a review of processing at successive stages of the ventral visual pathway.

1.1.1 Detection of Spatial Contrast

Any kind of visual information is first received by the retina, an inner coat of the eye balls. The lights comes into the eyes through the pupil, and figures in the visual world are projected on the retina. Then, the incoming lights are converted into corresponding electrical signals by the photoreceptors, which sit in the outermost layer of the retina. In the eyes of vertebrates, the generated electrical signals are passed onto ganglion cells via bipolar cells, and then the information is propagated to the brain. In other words, the ganglion cells are seen as the output cells of the retina.

One of the most important roles of the neural circuits in the retina is to detect spatial contrasts in the visual space. More precisely, it is known that ganglion cells do not react to lights that uniformly cover a large area in the visual field, while they become highly reactive when only a part of the visual field is relatively brighter or darker than the surrounding area. A receptive field of a neuron is defined as the area in the visual field in which a stimulus can influence the activity of the particular neuron. It is known that the receptive field of ganglion cells has two types: on-centred and off-centred. The ganglion cells with an on-centred receptive field increase their spike frequencies when the centre of the receptive field is illuminated. The same ganglion cells also increase their spike frequencies when the light in the surrounding area is turned off. On the other hand, the ganglion cells with an off-centred receptive field show the opposite responses, which is increasing the spike frequencies when the light at the centre is turned off or the light in the surrounding area is turned on. As a result, on-centred cells play the role of detecting the points that are brighter than its background, while off-centred cells play the role of detecting the points that are darker than its background. When the entire receptive field of a ganglion cell is illuminated, the effects from the centre and from the surrounding area cancel each other out, thus preventing the ganglion cell from reacting strongly in this condition.

As explained earlier, the photoreceptors make synaptic connections with bipolar cells, and the bipolar cells make synaptic connections with ganglion cells. This pathway normally does not spread horizontally, but rather follows one-to-one vertical topography, constituting the center part of the receptive field of a ganglion cell. On the other hand, there are other kinds of cells in the retina that establish horizontal synaptic connections such as horizontal cells and amacrine cells. These connections constitute the surrounding part of the receptive field.

1.1.2 Edge Detection in V1

The information about spatial contrast within an image is projected to the lateral geniculate nucleus (LGN) via axons from the ganglion cells, and then projected to the striate cortex (V1). Generally, this projection preserves the topography, so it is known that area V1 develops a map of the visual field based on the relative location of the receptive field of each cell in the visual field. In particular, the right-hand side of the visual field is mapped onto the left hemisphere of V1, while the left-hand side of the visual field is mapped onto the right hemisphere of V1.

One of the main characteristics of neurons in area V1 is that they react strongly to a bar of light at particular orientation in a specific retinal position (Hubel and Wiesel, 1962). These neurons are referred to as *simple cells*. The population of such cells at a particular retinal position will cover all stimulus orientations. As explained earlier, the receptive field of an LGN cell has a circular shape, similar to a ganglion cell. However, when the LGN cells establish convergent synaptic connections to V1 cells, the receptive fields of such V1 cells can form elliptical shapes. This can give a rise to the observed selectivity of simple cells in V1 to oriented line segments.

Later visual areas in the cerebral cortex receive information from area V1 via polysynaptic connections to accomplish higher level information processing tasks such object and face recognition. Traditionally, it has been proposed that there are two distinct pathways of visual information propagation from area V1: a dorsal visual pathway towards the parietal lobe and

a ventral visual pathway along the temporal lobe. It is generally thought that the dorsal visual pathway is responsible for representing spatial information to guide bodily actions such as reaching, while the ventral visual pathway is responsible for processing the spatial forms of objects and faces. In this thesis, our focus is mainly on the latter pathway, that is, the ventral visual pathway.

1.1.3 Border Ownership Representations in V1/V2

Cortical area V2 receives inputs from area V1. Zhou et al. (2000) have shown that the responses of simple cells in earlier cortical stages of visual processing such as V1 and V2, which respond preferentially to oriented edges, are also modulated by which side of an object or figure the edge occurs on. This is the case even when the figure/background cues lie well outside the classical receptive field of the neuron, which in area V2 is approximately 3 degrees in size. Such neurons are referred to as *border ownership cells*. Sugihara et al. (2011) later reported that the border ownership signal emerges with a latency of 61 ms, which is about 13 ms later than the onset of orientation selectivity. This suggests that the global image context specifying border ownership modulates the activity of these neurons. In other words, there must be a mechanism that enables the contextual information to be conveyed to these early stage visual neurons in V1 and V2 from higher stages such as V4.

1.1.4 Local Contour Shape Representations in V4

Information about elementary visual features represented in earlier layers such as V1/V2 is integrated in later areas such as V4 and IT (Brincat and Connor, 2004). Experimental studies have shown that neurons in these later stages of the primate ventral visual pathway encode the spatial structure of visual objects and their parts. For example, single unit recording studies carried out by Pasupathy and Connor (2001) have shown that within area V4, an intermediate stage of the ventral visual pathway, there are neurons that respond selectively to the shape of a local boundary element (e.g. concave or convex) at a particular position in the frame of reference of the object (i.e. with respect to the centre of the object). Some of these V4 neurons also maintain their response properties as an object shifts across different locations on the retina; i.e. they have developed translation-invariant representations. Therefore, these cells encode the spatial form of the object boundary by representing the relations between individual boundary contour elements within the object. Many theories suppose that object recognition operates through the computation of intermediate representations which reflect the spatial relations between the parts of objects (Brincat and Connor, 2004; Pasupathy and Connor, 2001; Giersch, 2001). Consistent with this view, a population of such V4 cells will provide a distributed encoding of the complete boundary shape of an object (Pasupathy and Connor, 2002).

1.1.5 Object Representation in Inferotemporal Cortex (IT)

Further experimental studies have shown that neurons in the later stages of the ventral visual pathway, including inferotemporal cortex (IT) and especially posterior IT (TEO), integrate information from multiple boundary contour elements (Brincat and Connor, 2004). This representation of the detailed spatial form of the separate parts of each object may provide a necessary foundation for the subsequent recognition of whole objects. That is, object selective cells at the end of the ventral visual pathway may learn to respond to unique distributed representations of object shape in earlier areas (Booth and Rolls, 1998). In higher layers of the ventral pathway such as the anterior IT (TE), the responses of neurons to objects and faces show invariance to retinal location, size, and orientation (Tanaka et al., 1991; Rolls et al., 1992; Rolls, 2000; Perrett and Oram, 1993; Rolls and Deco, 2002).

Similarly, functional magnetic resonance imaging (fMRI) studies in humans have revealed several cortical regions within the temporal lobe, which are exclusively dedicated to face processing (Perrett et al., 1992; Kanwisher et al., 1997; Pitcher et al., 2011; Zhang et al., 2012). Also, there is evidence for hierarchical processing. For example, an early stage of processing, the occipital face area (OFA) in the inferior occipital gyrus, has been found to contribute to face perception by responding to individual facial features such as the eyes, nose, and mouth (Pitcher et al., 2011). On the other hand, a later stage of processing, the fusiform face area (FFA) in the lateral fusiform gyrus, has been found to integrate such information by responding more strongly to intact rather than scrambled faces (Kanwisher et al., 1997; Zhang et al., 2012). Recently, it has also been reported that the face areas may also exhibit "faciotopy" where different cortical patches represent different face features, and the cortical distances between the feature patches reflect the physical distance between the features in a face (Henriksson et al., 2015).

Consistent with these findings, also in macaques, several face sensitive areas have been identified in the temporal lobe, which are known as *face patches* (Gross et al., 1972). It has been argued that the homologue of the FFA in macaques is the *middle face patch* (Tsao et al., 2006). Some of those neurons are found to respond to the presence of facial features such as the eyes, nose, or mouth, while other neurons encode the many spatial relationships between these facial features (Freiwald et al., 2009). Some neurons also encode global properties such as facial identity or expression (Morin et al., 2014; Hasselmo et al., 1989a). These experimental findings indicate that our ability to process and recognise faces utilises this rich tapestry of different kinds of visual information.

1.2 Computational Modelling Study

Inspired by the known hierarchical architecture of the primate vision, various models of visual processing incorporate a convergent hierarchy of neurons organized into layers. One of the earliest and well-known models, 'neocognitron', was developed by Fukushima (1980). In this model, the network consists of an input layer of simulated photoreceptors followed by two types of neuronal layers in turn for multiple times: a layer of simple-cells (S-layer) to enhance feature integrations with modifiable synaptic connections and a layer of complex-cells (C-layer) to achieve transform invariance. This model set a foundation of the current approach of hierarchical modelling study of primate visual brains (Riesenhuber and Poggio, 1999; Wallis and Rolls, 1997).

While many modelling studies have investigated various kinds of visual processing in the primate visual system (Eliasmith et al., 2012; Serre et al., 2005), most of these investigations have not employed biologically accurate models or were not concerned with uncovering the synaptic learning mechanisms by which the visual representations develop in the first instance. However, there is a large body of experimental evidence for learning of visual form recognition within the temporal lobes (Wallis, 2013). For example, Baker et al. (2002) showed that exposure to abstract shapes formed by combing multiple parts enhanced both parts-level and holistic shape tuning of neurons in the Inferior temporal cortex (IT). Studies with fMRI have also reported large-scale alteration of the organization and selectivity of temporal lobe cortex in humans after training with visual stimuli (Beeck et al., 2006; Gillebert et al., 2008). Additionally, although the discrimination of non-face objects is known to be more difficult than for faces, training on non-face objects improves the discrimination of these stimuli to nearly that of faces (Yue et al., 2006). Thus, how visual representations develop through a biologically plausible process of visually-guided learning is a key question that needs to be addressed by theoreticians, and is a fundamental aspect of the model simulations presented in this thesis.

1.2.1 V4/TEO cells

A number of modelling studies have tried to reproduce the observed shape selective and translation invariant firing properties of neurons in area V4 and TEO (Rodrguez-Snchez and Tsotsos, 2012; Cadieu et al., 2007). However, these past models have not utilised biologically plausible learning mechanisms to guide the development of cell firing properties. In particular, previous models have not used plausible, local learning rules, which use pre- and post-synaptic cell quantities to drive modification of the synaptic connections during visually-guided learning. Therefore, it still remains a challenge to understand exactly how those V4 and TEO neurons develop their shape selective response properties through learning.

1.2.2 Facial Representations

It is also the case that the mechanism by which the visual system separates the representations of global facial attributes, which are always seen together, into different brain areas has been a long-standing question in this field. Based on such functional and anatomical specialization, Haxby et al. (2000) hypothesized the existence of a 'core system' that is dedicated for visual analysis in the temporal lobe. The core system includes the OFA that detects the simpler features of faces, the Superior temporal sulcus (STS) that processes changeable attributes of faces such as expressions, and the FFA that processes invariant attributes of faces such as identity. However, these previous theoretical and experimental studies do not explain the precise learning mechanisms by which the neuronal representations of global attributes, such as identity and expression, may become mapped onto separate processing areas in the later stages of the visual system.

Lades et al. (1993) presented the first self-organising neural model that developed representations of the spatial relationships between facial features. Their model employed a feature based approach to face recognition via active dynamic linking of features (von der Malsburg, 1981; von der Malsburg and Schneider, 1986). The model uses an input representation by which each face is convolved with a set of Gabor filters across the visual field. The output layer of the network constructs a graph representation of the face, with each node in the graph representing a particular facial feature, and each link representing the feature relation (Lades et al., 1993). The model was proposed to be biologically realistic because the development of the output face representations is unsupervised. Nevertheless, it is not clear how a population of neurons in the brain may store facial representations in the form of graphs.

A more biologically plausible approach to modelling how transform (e.g. location or view) invariant representations of faces and non-face objects may develop through unsupervised, associative learning mechanisms in the higher stages of the ventral visual pathway was carried out by Wallis and Rolls (1997). The network architecture is shown in Figure 1.1. The inputs are represented as columns of V1-like spatial filter activation values similar to the model proposed by Lades et al. (1993). The architecture consists of four layers of competitive neural networks representing successive visual areas V2, V4, TEO (posterior IT), and TE (anterior IT). During training with visual images of faces and other objects, the feedforward synaptic connections between successive layers were modified by a biologically plausible, local, associative learning rule. The study showed that competitive learning allows neurons in the intermediate layers of the model to learn to respond to particular combinations of simple visual features present in faces and non-face objects. By building on these intermediate layer representations, the higher layers were then able to develop transform invariant representations of whole faces. A detailed description of the model is provided in the Section 1.3.

Tromans et al. (2011) demonstrated how physically separated representations of facial identity and expression may develop through a biologically plausible process of unsupervised competitive learning in this model, VisNet. Nonetheless, this modelling study used highly idealised cartoon faces, in which these two global attributes were artificially encoded by different facial

features. More recently, Wallis (2013) has started exploring various aspects of recognition which are generally regarded as unique to faces such as holistic processing (Tanaka and Farah, 1993), configural processing (Leder and Bruce, 1998), sensitivity to inversion (Yin, 1969; Maurer et al., 2002), the other-race effects (Chance et al., 1982). Additionally, the development of face representations within a more biologically accurate spiking neural network model with spike-timing dependent plasticity (STDP) has been presented by Masquelier and Thorpe (2007). However, again, these previous studies have not yet fully explained how these visual representations, which correspond to those observed in single unit recording neurophysiology studies, develop through successive layers of the model.

1.2.3 Spike Dynamics and Polychronization

Furthermore, many early neural network models of brain function assumed that neurons transmit information exclusively through modulation of their mean firing rates. These 'rate-coded' models represented only the current average firing rate of each neuron, and did not explicitly represent the timings of the action potentials or 'spikes' emitted by cells. However, in modern literature the precise timing of spikes has been proposed to strongly contribute to encoding in the brain (Fujii et al., 1996; Akolkar et al., 2015; Nikoli et al., 2013). Consistent with this view, there is growing evidence from neurophysiological studies supporting the importance of spike-timing dynamics in the brain (Softky, 1995; Lindsey et al., 1997; Prut et al., 1998; Mao et al., 2001).

In the brain, neurons represent information and communicate with each other by pulses in their somatic membrane potential, called action potentials or 'spikes'. The activity of a somatic spike propagates down the axon of the neuron, causing neurotransmitters to be released from multiple presynaptic axon terminals into their corresponding synaptic clefts. Binding of the neurotransmitters to the receptors of the postsynaptic dendrites causes a change in the electrical activity of the postsynaptic neurons, constituting a communication of information from the presynaptic neuron to the postsynaptic neuron. This neuron also spikes if the excitation of the postsynaptic neuron from its afferent synapses increases the membrane potential above its firing threshold potential. Raising the membrane potential of the postsynaptic neuron above the firing threshold generally requires the activation of afferent synapses within a brief temporal window, as the membrane potential naturally decays quickly back to a resting potential without further afferent excitatory activation.

The relative timings of the spikes emitted by a pair of pre- and postsynaptic neurons has also been shown to affect learning through spike time dependent changes in synaptic efficacy (Bi and Poo, 1998; Markram et al., 1997), and hence how information and representations are stored and propagated in the network. If a presynaptic neuron fires in a short time period (up to tens of ms) prior to the postsynaptic neuron firing, the synaptic efficacy increases. An increase in synaptic efficacy is known as Long Term Potentiation (LTP). If the presynaptic neuron instead fires in a short period of time following the firing of a postsynaptic neuron, the efficacy of the synapse is reduced. This reduction in synaptic efficacy is known as Long Term Depression (LTD). These forms of LTP and LTD, which depend on the relative timings of the pre- and postsynaptic neurons, are known as Spike-Timing-Dependent Plasticity (STDP). Compared to firing rate based synaptic learning rules employed in rate-coded models, an STDP learning rule can result in very different self-organisation of the synaptic connectivity in the network when trained on visual scenes containing multiple objects (Evans and Stringer, 2012, 2013).

Simulation studies have shown that if the synaptic connections within a large population of neurons have axonal transmission delays that are drawn from a random distribution of variable magnitudes, from say a few milliseconds to several tens of milliseconds, then groups of cells that detect the coincidence of incoming volleys of spikes emerge through STDP (Izhikevich et al., 2004). Furthermore, the network develops repeating temporal chains of spiking activity distributed across subgroups of coincidence detecting neurons, i.e. neurons firing in a well defined

temporal sequence. This is referred to as 'polychronization' (Izhikevich, 2006). Each subgroup of coincidence detecting neurons that comes together to form a regularly repeating temporal chain of activity is known as a 'polychronous group' (PG). It has been hypothesised that each PG could represent a particular sensory (e.g. visual) stimulus such as an object or perhaps episodic memory (Izhikevich, 2006).

In theory, polychronization in spiking networks can offer a dramatic increase in representational capacity compared to rate-coded models that do not exploit the timings of spikes (Izhikevich, 2006). Paugam-Moisy et al. (2008) have recently examined how PGs selectively respond to artificial input patterns after training with STDP and shed light on the potential of utilising PGs for real-life machine learning tasks such as handwritten digit recognition. However, the study carried out by these authors did not address three key issues as follows. Firstly, the study carried out by Paugam-Moisy et al. (2008) used carefully ordered spike trains to represent input images, which is not biologically plausible. What would happen if the input spike trains contained much more random variation as would be expected in the brain? Secondly, their model did not incorporate multiple synaptic connections with different randomised axonal transmission delays between each pair of pre- and postsynaptic neurons. This meant that the axonal transmission delay between any pair of neurons was fixed to a single value, and could not be effectively selected from a number of alternatives by STDP learning. Consequently, the set of possible PGs that a neuron could participate in was limited before learning. Thirdly, and perhaps most importantly, the study of Paugam-Moisy et al. (2008) did not investigate how feature binding representations, which explicitly encode the binding relations between low and high-level features, might develop through polychronization within a hierarchical model of visual processing.

1.2.4 Border Ownership Representation and Feature Binding

One example of which the spike dynamics may play an important role is the border ownership representation in the lower visual layers (corresponding to cortical areas V1 or V2) that respond selectively to a vertical straight edge on either the left or right boundary of an object presented at the neuron's preferred retinal location. Some theoreticians have suggested that the context integration required for border ownership representations in V1 and V2 can be achieved via lateral propagation of signals within a layer via horizontal fibres (Zhaoping, 2005; Baek and Sajda, 2005; Nishimura and Sakai, 2004). However, Sugihara et al. (2011) have argued that the conduction velocity of horizontal fibres is too slow (most of them being between 0.1 and 0.4 m/s (Angelucci and Bullier, 2003)) to produce the border ownership signals within the short latency observed in neurophysiology studies. Furthermore, Sugihara et al. (2011) showed that varying the distance between the target border and the visual features that carry contextual information about the 'owner' of the border does not in fact influence the latency before the border ownership signals arise. Therefore, they concluded that context influence by horizontal signal propagation alone is highly unlikely.

On the other hand, the feedforward (bottom-up) and feedback (top-down) connections between successive visual stages have fast-conducting axons, with conduction velocities of between 2 and 6 m/s, which is about ten times faster than cortical horizontal fibres (Angelucci and Bullier, 2003). Accordingly, both Craft et al. (2007) and Jehee et al. (2007) have proposed models that involve hypothetical 'grouping circuits' within a higher cortical layer that capture the contextual information about local boundary elements, and these contextual signals are then relayed down through feedback connections to modulate responses in an earlier layer. They proposed that the larger receptive fields in the higher layer allow the network to employ 'grouping circuits' without having to rely on slow lateral propagation of signals.

However, within the typical modelling studies with rate-coded neurons, there is the well-known problem of "superposition catastrophe" (von der Malsburg, 1999) that occurs within a rate-coded model when multiple objects are presented simultaneously, so the exact learning

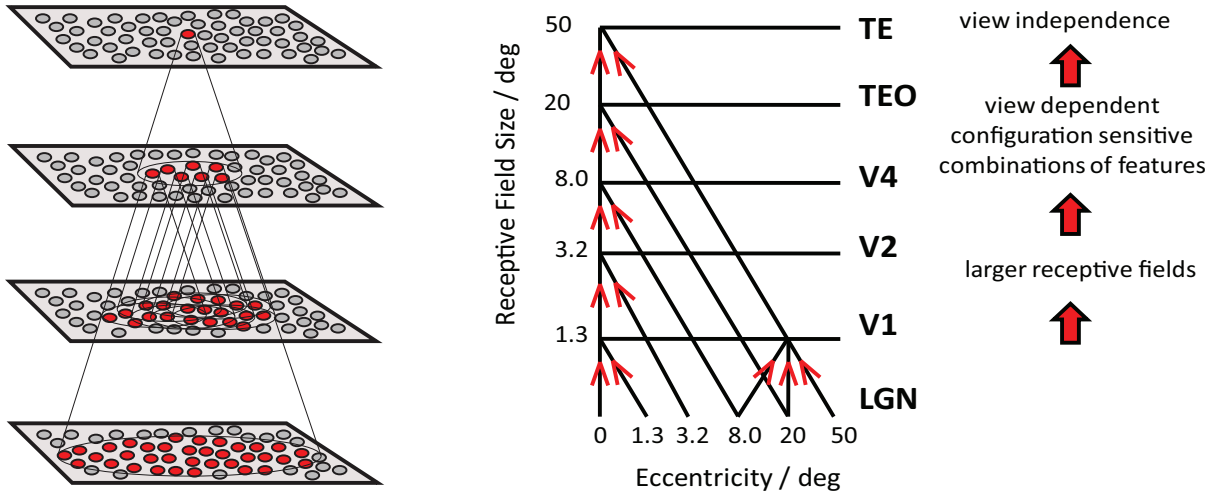

Figure 1.1: Left: Stylised image of the four layer VisNet architecture. Convergence through the network is designed to provide fourth layer neurons with information from across the input retina. Right: Convergence in the visual system V1: visual cortex area V1;TEO posterior inferior temporal cortex, TE anterior inferior temporal cortex. The diagonal lines show the convergence of feedforward connections through successive layers of the ventral visual system leading to an increase in receptive field size from 1.3° in V1 to 50° in TE.

mechanism of such neuronal circuits are not yet uncovered. One potential solution for this limitation would be the temporal coding which could be achieved with the spiking neural networks.

1.3 VisNet

In this section, the detailed implementation of the hierarchical neural network model of the primate ventral visual pathway, VisNet, which was used for a number of simulation studies conducted in this thesis (Chapter 2, 3, and 4) is described.

VisNet is a rate-coded neural network model of the primate ventral visual pathway, which was originally developed by Wallis and Rolls (1997). The standard network architecture is shown in Figure 1.1. It is based on the following: (i) A series of hierarchical competitive layers with local graded lateral inhibition. (ii) Convergent connections to each neuron from a topologically corresponding region of the preceding layer, leading to an increase in the receptive field size of neurons through the visual processing areas. (iii) Synaptic plasticity based on a biologically-plausible local learning rule such as the Hebb rule (1.7) or trace rule (1.8) and (1.9), which are explained below.

In past work, the hierarchical series of 4 neuronal layers of VisNet have been related to the following successive stages of processing in the ventral visual pathway: V2, V4, the posterior IT (TEO), and the anterior IT (TE).

In VisNet, 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 contained approximately 67% of the connections from the preceding layer. Typical values employed in the current studies are given in Table 1.1. The gradual increase in the receptive field of cells in successive layers reflects the known physiology of the primate ventral visual pathway (Freeman and Simoncelli, 2011; Pasupathy, 2006; Pettet and Gilbert, 1992).

In many single-cell studies, translation invariance is measured by moving the stimulus within the receptive field, not across the full retina. Accordingly, the size of the receptive fields in the 4th (output) layer neurons is limited by the degree of divergence in the topological feedforward connections between successive layers of the model. Usually, the connectivity is set to permit neurons in the output layer of VisNet to receive visual signals from any part of the

Table 1.1: VisNet parameters

Layer	Dimensions	number of connections	radius
Layer 4	128 × 128	100	12
Layer 3	128 × 128	100	9
Layer 2	128 × 128	100	6
Layer 1	128 × 128	201	6
Retina	256 × 256 × 16		

retina simulated. However, this does not automatically mean that the receptive fields of output neurons will cover the entire retina.

The VisNet architecture is feed-forward with lateral interactions within layers. There are many engineering approaches to efficiently solve similar problems extensively rely their architectures on top-down information flows, mainly for their supervised learning. However, our objectives are rather different. Our aim is to pin down the simplest form of core-mechanisms in intermediate vision that are sufficient to explain a specific brain function. In fact, in other hierarchical neural network modelling studies, such top-down information transfer is often excluded (Olshausen et al., 1993; Riesenhuber and Poggio, 1999; Serre et al., 2005, 2007; Wallis, 2013).

The researchers involved in these last publications acknowledge the extensive presence of such back projections in the visual cortex; however, they also think the exact roles of these projections still remain a matter of debate. For example, it has been proposed that the role of these feedback pathways is to relay the interpretations of higher cortical areas to lower cortical areas in order to verify the high-level interpretation of a scene (Mumford, 1992) or to refine the tuning characteristics of lower-level cortical cells based upon the interpretations made in higher cortical areas (Tsotsos, 1993). On the other hand, numerous physiological studies have also reported that only short time spans are required for various selective responses to appear in monkey IT cells, which imply that feedback processes may not be critical for coarse, rapid recognition (Perrett et al., 1992; Hung et al., 2005; Vanrullen, 2007).

I also take a similar point of view, and learning mechanisms implemented in the simulation studies conducted in this thesis are a direct extension of previous papers in the field. The underlying learning mechanisms within the competitive networks are well understood (Rumelhart and Zipser, 1985). In this thesis, these established learning mechanisms were applied to the important new problem of how the primate ventral visual system learns to represent various kinds of visual inputs.

1.3.1 Pre-processing of the visual input by Gabor filters

Before the visual images are presented to VisNet's input layer 1, they are pre-processed by a set of input filters that accord with the general tuning profiles of simple cells in V1. The filters provide a unique pattern of filter outputs for each transform of each visual object, which is passed through to the first layer of VisNet. In the VisNet studies conducted in this thesis, the input filters matched the firing properties of V1 simple cells, which respond to local oriented bars and edges within the visual field (Jones and Palmer, 1987; Cumming and Parker, 1999). The input filters used are computed by the following equations:

$$g(x, y, \lambda, \theta, \psi, b, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{x'}{\lambda} + \psi\right) \tag{1.1}$$

with the following definitions:

$$\begin{aligned} x' &= x\ \cos\theta + y\ \sin\theta \\ y' &= -x\ \sin\theta + y\ \cos\theta \\ \sigma &= \frac{\lambda(2^b+1)}{\pi(2^b-1)}\sqrt{\frac{\ln 2}{2}} \end{aligned} \tag{1.2}$$

where x and y specify the position of a light impulse in the visual field (Petkov and Kruizinga, 1997). The parameter λ is the wavelength ($1/\lambda$ is the spatial frequency), σ controls number of such periods inside the Gaussian window based on λ and spatial bandwidth b, θ defines the orientation of the feature, ψ defines the phase, and γ sets the aspect ratio that determines the shape of the receptive field. In the experiments in this thesis, an array of Gabor filters is generated at each pixel in the input images with the parameters given in Table 1.2.

Table 1.2: Parameters for Gabor input filters

Parameter (Symbol)	Value
Wavelength(λ)	2, 4, 8, and 16
Spatial bandwidth (b)	1.5 octaves
Orientation(θ)	$0, \pi/4, \pi/2, 3\pi/4$
Phase shift (ψ)	0: white on black bar π: black on white bar
Aspect ratio (γ)	0.5

The outputs of the Gabor filters are passed to the neurons in layer 1 of VisNet according to the synaptic connectivity given in Table 1.1. Each layer 1 neuron received connections from 201 randomly chosen Gabor filters localised within a topologically corresponding region of the retina. In the original VisNet model (Wallis and Rolls, 1997), the input filters were tuned to the four different spatial wavelengths 2, 4, 8, and 16 pixels. The shortest wavelength filters provided the highest resolution information about the image. The neurons in the first layer of VisNet were thus assigned most of their afferent inputs from the shortest wavelength filters. In some of the simulation simulations reported in this thesis, the model used inputs from only the shortest wavelength filters, which was found to be sufficient to represent the simple visual objects. However, for consistency with past VisNet simulations, each neuron in the first layer of VisNet always received afferent connections from 201 of the short wavelength filters.

1.3.2 Calculation of cell activations within the network

Within each of the neural layers 1 to 4 of the network, the activation h_i of each neuron i was set equal to a linear sum of the inputs r_j from afferent neurons j in the preceding layer weighted by the synaptic weights w_{ij}. That is,

$$h_i = \sum_j w_{ij} r_j \tag{1.3}$$

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

1.3.3 Lateral inhibition and excitation between neurons within each layer

In the simulations reported below, the lateral inhibition between the neurons within each neuronal layer was implemented in one of two different ways. The simplest approach was to implement a competitive network architecture (Rolls and Treves, 1998), in which neurons inhibited all of their neighbours. However, in some simulations a more complex Self-Organising Map (SOM) architecture (Kohonen, 1982), which included both short range excitation and longer range inhibition between neurons (i.e., a 'Mexican hat' connectivity), was also implemented. A SOM architecture leads to a map-like arrangement of neuronal response characteristics across a layer after training, with nearby cells responding to similar inputs.

VisNet is implemented with a wrap-around organization of the cells in each layer. This means that the upper neighbour of the cell at index (0,0) is the cell at index (0, y_{max}), and the left neighbour of the cell at index (0,0) is the cell at index (x_{max}, 0). With this way, any imbalanced computation due to the boundary could be ignored.

1.3.3.1 Competitive network architecture

The original VisNet model implemented a competitive network within each layer. Within each layer, competition was graded rather than winner-take-all. To implement lateral competition, the activations h_i of neurons within a layer were convolved with a spatial filter, I_{ab}, where δ controlled the contrast and σ controlled the width, and a and b indexed the distance away from the centre of the filter:

$$I_{a,b} = \begin{cases} -\delta \exp\left(-\frac{a^2+b^2}{\sigma^2}\right) & a \neq 0 \,\text{or}\, \text{b} \neq 0 \\ 1 - \sum\limits_{a \neq 0, b \neq 0} I_{a,b} & a = 0 \,\text{and}\, \text{b} = 0 \end{cases} \tag{1.4}$$

1.3.3.2 Self-organising map

In this thesis, I also ran simulations with a self-organising map (SOM) (von der Malsburg, 1973; Kohonen, 1982) implemented within each layer. In the case of the SOM architecture, short-range excitation and long-range inhibition are combined to form a Mexican-hat spatial profile and is constructed as a difference of two Gaussians as follows:

$$I_{a,b} = -\delta_I \exp\left(-\frac{a^2+b^2}{\sigma_I^2}\right) + \delta_E \exp\left(-\frac{a^2+b^2}{\sigma_E^2}\right) \tag{1.5}$$

To implement the SOM, the activations h_i of neurons within a layer were convolved with a spatial filter, I_{ab}, where δ_I controlled the inhibitory contrast and δ_E controlled the excitatory contrast. The width of the inhibitory radius was controlled by σ_I and the width of the excitatory radius by σ_E. The parameters a and b indexed the distance away from the centre of the filter.

1.3.4 Contrast enhancement of neuronal firing rates within each layer

Next, the contrast between the activities of neurons with each layer was enhanced by passing the activations of the neurons through a sigmoid transfer function (Rolls and Treves, 1998) as follows:

$$r = f^{sigmoid}(h') = \frac{1}{1 + \exp\left(-2\beta(h' - \alpha)\right)} \tag{1.6}$$

where h' is the activation after applying the lateral competition or SOM filter, r 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 within each layer of neurons to control the sparseness of the firing rates. For example, to set the sparseness to 5%, the threshold is set to the value of the 95th percentile point of the activations within the layer.

1.3.5 Training the network: visually-guided learning of synaptic weights

During training, visual images were presented to the network after pre-processing by the Gabor input filters. The outputs of the Gabor filters were passed to layer 1 of VisNet. Activity was then propagated sequentially through layers 2 to 4 using the same mechanisms at each layer.

During training with visual objects, the strengths of the feed-forward synaptic connections between successive neuronal layers are modified by biologically plausible local learning rules, where the change in the strength of a synapse depends on the current or recent activities of the pre- and post-synaptic neurons. Two such learning rules were implemented with different learning properties.

1.3.5.1 The Hebb learning rule

One simple well known learning rule is the Hebb rule:

$$\delta w_{ij} = k r_i^\tau r_j^\tau \tag{1.7}$$

where δw_{ij} is the change of synaptic weight w_{ij} from pre-synaptic neuron j to post-synaptic neuron i, r_i^τ is the firing rate of post-synaptic neuron i at timestep τ, r_j^τ is the firing rate of pre-synaptic neuron j at timestep τ, and k is the learning rate constant. The relatively simple Hebb learning rule is standardly used in competitive neural networks. It is employed in this thesis for simulations that aim to explore how neurons may learn to respond to individual visual stimuli.

1.3.5.2 The trace learning rule

Alternatively, a trace learning rule may be implemented for the simulation studies that require neurons to learn different transforms of objects based on those temporal continuity. As described in more detail in Section 1.4.2, Trace learning is a biologically plausible mechanism to achieve such temporal association by incorporating a memory trace of recent neuronal activity into the learning rule used to modify the feedforward synaptic connections (Foldiak, 1991; Wallis and Rolls, 1997):

$$\delta w_{ij} = k \overline{r}_i^{\tau-1} r_j^\tau \tag{1.8}$$

where $\overline{r}_i^\tau$ is the trace value of the firing rate of post-synaptic neuron i at timestep τ. The trace term is updated at each timestep according to

$$\overline{r}_i^\tau = (1-\eta) r_i^\tau + \eta \overline{r}_i^{\tau-1} \tag{1.9}$$

where η may be set anywhere in the interval $[0, 1]$, and for the simulations described below, η was set to 0.8. The effect of this learning rule is to encourage neurons to learn to respond to visual input patterns that tend to occur close together in time. If the eyes shift about a visual scene containing a static object, then the trace learning rule will tend to bind together successive images corresponding to that object in different retinal locations. Consequently, such a learning rule has been previously used to model the development of translation invariant neuronal responses in the primate ventral visual pathway (Foldiak, 1991; Wallis and Rolls, 1997).

In a typical simulation study, natural eye movements are simulated implicitly during training by shifting each visual object in turn across a number of retinal locations. That is, to simulate natural rapid eye movements during visual inspection of each object, the visual object itself is shifted across the retina. After an object is shifted through all of the retinal locations, the next object is presented across the same locations. These image presentation statistics combine with the trace learning rule to encourage neurons to respond selectively to particular visual features within the objects over all retinal locations.

The trace rule provides a biologically plausible mechanism whereby translation invariant neuronal responses may emerge in the brain without the need to artificially prewire synaptic connections as has been implemented in some other modelling approaches.

1.3.5.3 Weight Normalization

To prevent the same few neurons always winning the competition, the synaptic weight vector $\mathbf{w}_i$ of each neuron i is renormalised to unit length after each learning update for each training pattern by setting

$$\mathbf{w}_i = \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|} \tag{1.10}$$

where $\|\mathbf{w}_i\|$ is the length of the vector $\mathbf{w}_i$ given by

$$\|\mathbf{w}_i\| = \sqrt{\sum_j w_{ij}^2} \tag{1.11}$$

Neurophysiological evidence for synaptic weight normalization is provided by Royer and Paré (2003).

1.4 Translation Invariant Learning Mechanisms

In vision, it is important to correctly identify an object in the environment as being the same despite changes in the retinal image. As introduced in Section 1.1, over successive stages in the visual system, neurons develop response properties that are invariant to the size, position, and view of an object (Rolls et al., 1992; Rolls, 2000; Rolls and Deco, 2002; Desimone, 1991; Tanaka et al., 1991). Cells in IT that show invariance to the translation (Op de Beeck and Vogels, 2000; Kobatake and Tanaka, 1994; Ito et al., 1995; Tovee et al., 1994), size (Rolls and Baylis, 1986; Ito et al., 1995), contrast (Rolls and Baylis, 1986), lighting (Vogels and Biederman, 2002), spatial frequency (Rolls et al., 1985, 1987), and view (Hasselmo et al., 1989a; Booth and Rolls, 1998) of objects have been reported.

Developing invariant recognition of objects involves associating together representations of the same object under different conditions. In the particular case of translation invariance, this would mean developing associations between the neural representations of an object in different spatial locations on the retina. In order to develop these associations, the visual system can exploit constraints placed upon object translation by the environment. For example, when an object translates from one point to another, it does so in a manner that is continuous in both space and time. These same constraints can be exploited for the development of view invariance, as different views of an object also appear in a spatially and temporally continuous manner.

1.4.1 Continuous Transformation (CT) Learning

One method for developing translation and view invariance, known as *Continuous Transformation (CT) learning*, depends on the spatial continuity of object transformation (Stringer et al., 2006). It has been reported that people learn to associate visually similar images together.

Human infants can categorize stimuli into human vs. nonhuman categories during the initial months of life, learn to categorise human faces into male vs. female in the next several months, and distinguish broad age groups of the faces in the further next several months (Quinn, 2010; Quinn et al., 2002; Damon et al., 2016). In other words, various holistic processing of the faces should gradually develop affected by experience. In an experimental study, Preminger et al. (2007) trained subjects to classify faces into two categories: friends (F) and non-friends (NF). Upon reaching good performance, subjects were then trained with a sequence of morphed images from F to NF. The subjects were tested on how they classified the morphed images. Initially, the first half of the morphed image sequence was classified as F, while the second half of the morphed sequence was classified as NF. However, as training progressed, the separation threshold moved towards NF; that is, an increasing number of frames were classified as F. Eventually, all morphed frames were classified as F.

CT learning is an invariance learning mechanism that may provide an insight into the mechanism of such memory reconstruction via ordinary Hebbian learning 1.7 at the neuronal level (Stringer et al., 2006). It associatively remaps the feedforward connections between successive neural layers while keeping the same initial set of output neurons activated as the input patterns are gradually changed. Consider a set of stimuli that can be arranged into a continuum, in which each successive stimulus in the continuum has a degree of overlap – a number of features in common – with the previous stimulus in the continuum. CT learning can exploit this feature

overlap between successive stimuli to form a single percept of all, or at least a large subset, of the stimuli in the stimulus set.

Specifically, when an output neuron responds to one of the input patterns, the feedforward connections from the active input neurons to the active output neuron are strengthened by associative (Hebbian) learning. Then, when the next similar (overlapping) input pattern is presented, the same output neuron is again activated due to the previously strengthened connections. Now the second input pattern is associated with the same output neuron through further associative learning. This process can continue to map a sequence of many gradually transforming input patterns, where each input pattern has a degree of spatial overlap with its neighbours, onto the same output neuron.

1.4.2 Trace Learning

A second method for developing translation invariant representations utilises the temporally continuous nature of object translation. More specifically, stimuli that are experienced close together in time are likely to be strongly related; for instance, successive stimuli could be different views of the same object. If a mechanism exists to associate together stimuli that tend to occur close together in time, then a network will learn that those stimuli form a single percept. Neurophysiological evidence suggests that the brain might use this type of information to develop translation invariant representations of objects (Li and DiCarlo, 2008). As breaking temporal continuity causes neurons to lose their selective responses to different objects. Different approaches have been developed in order to understand how the brain might exploit this temporally continuity, such as using inputs representing temporal context to guide learning (Becker, 1999), learning high probability sequences of visual input in order to infer the object being presented (George and Hawkins, 2005), and extracting slowly changing features in the visual inputs to analyse the transform invariant representations (Berkes and Wiskott, 2005; Wiskott and Sejnowski, 2002).

Trace learning provides one such mechanism by incorporating a temporal memory trace of postsynaptic cell activity $\bar{r}_i$ into a standard Hebbian learning rule as described in Equation (1.8) and (1.9) (Foldiak, 1991; Rolls et al., 1992; Wallis and Rolls, 1997). This encourages neurons to respond to stimulus image transforms that occur close together in time.

The advantage of this approach is that it can arise naturally out of biophysically realistic spiking neural networks when longer time constants for synaptic conductance are introduced (Evans and Stringer, 2012). Increasing this time constant keeps the neuron active for longer as it lengthens the time period over which current leaks into the postsynaptic neuron, thus allowing temporal trace learning to occur. Therefore, it is feasible that this type of learning could occur in the brain without requiring a specific architecture to operate.

1.5 Analysis Methods

In this section, the typical analysis techniques used across the simulation studies conducted in this thesis are described.

1.5.1 Testing the network

After the synaptic weights were established by training the network on a set of visual objects, the learned response properties of neurons through successive layers were tested. This was done by presenting a testing set of visual objects. A number of tests are applied to the recorded neuronal responses, including information theory, which are described below. The learned response properties of an output cell were also analysed by plotting the subset of input Gabor filters with the strongest feed-forward connections to that output cell after training.

1.5.2 Information Analysis

To quantify the performance in transformation invariance learning with VisNet, the techniques of Shannon's information theory have previously been used (Rolls and Treves, 1998), which is based on the Kullback-Leibler (KL) divergence of the conditional response distribution from the unconditional distribution. Information theory can be used to quantify how selective neurons are for particular stimuli, each of which may translate across different locations on the retina. If the responses r of a neuron carry a high level of information about the presence of a particular stimulus s, then this implies that the neuron will respond selectively to the presence of that stimulus regardless of where the stimulus is presented on the retina. In this way, information theory can provide a direct measure of both the selectivity of a neuron for a particular stimulus, as well as how translation-invariant the neuronal responses are as the stimulus is shifted across the retina.

Typically, two information measures were used to assess the ability of the network to develop neurons that are selective to the presence of stimuli but also invariant to their occurrence in different retinal locations (see Rolls et al. (1997); Rolls and Milward (2000)). These two measures use the responses from either individual neurons (single-cell information analysis) or small ensembles of neurons (multiple-cell information analysis), each of which will be discussed in turn.

The following exposition provides a theoretical account of the two information measures used in this thesis. However, in order to keep the notation consistent with past publications (Rolls et al., 1997; Rolls and Milward, 2000), the neuronal firing rates were denoted by r.

1.5.2.1 Single-cell information

A single cell information measure was applied to individual cells to measure how much information is available from the responses of a single cell about which stimulus input is present. The amount of stimulus specific information that a certain cell transmits is calculated from the following formula with details given by Rolls and Milward (2000).:

$$I(s, \vec{R}) = \sum_{r \in \vec{R}} P(r|s) log_2 \frac{P(r|s)}{P(r)} \tag{1.12}$$

Here s is a particular stimulus and $\vec{R}$ is the set of responses of a cell to the set of stimuli. The maximum information that an ideally developed cell could carry is given by the formula:

$$\text{Maximum cell information} = log_2(n) \text{ bits} \tag{1.13}$$

where n is a number of different stimuli. This maximum single cell information measure is achieved when a neuron responds selectively to all transforms of a particular category of objects, but does not respond to any other object.

1.5.2.2 Multiple-cell information

While useful in assessing the tuning properties of a particular neuron, the single-cell information measure cannot give a complete assessment of VisNet's performance with respect to recognition of each category of visual stimuli. If all cells learned to respond to the same stimulus category (according to the single-cell measure) then there would be relatively little information available about the whole set of stimulus categories $\vec{S}$. To address this issue, a multiple-cell information measure, which assesses the amount of information that is available about the whole set of categories of visual stimuli from a *population* of neurons, is also calculated. This measure quantifies the network's ability to determine which stimulus is currently presented to the network based on the set of responses, $\vec{R}$, of a sub-population of cells. Here the procedures for calculating the

multiple-cell information measure as described by Rolls and Treves (1998); Rolls and Milward (2000) are adopted.

In brief, we would like to calculate the mutual information between the stimuli and the responses – the average amount of information obtained (across all stimuli) from the responses of the ensemble, about which stimulus was present after a single presentation of a stimulus. However, due to the difficulty in adequately sampling this high dimensional neural response space, it is very hard to construct accurate probability distributions for directly calculating the mutual information. Instead, a decoding procedure is used to estimate which stimulus s' gave rise to the particular firing rate response vector on each trial. In other words, the predicted stimulus is simply a function of the response, as determined by the algorithm considered. A probability table is then constructed between the real stimuli, s and the decoded stimuli, s'. From this probability table, the multiple-cell information is then calculated as follows.

$$I_{\vec{C}}(S, S') = \sum_{s,s'} P(s, s') log_2 \frac{P(s, s')}{P(s)P(s')} \tag{1.14}$$

$$P(s') = \sum_{s \in S} P(s'|R_{\vec{C}}(s)) \times P(R_{\vec{C}}(s))) \tag{1.15}$$

$$P(s, s') = P(s'|R_{\vec{C}}(s)) \times P(R_{\vec{C}}(s)) \tag{1.16}$$

Here, S represents the set of the stimuli presented to the network, and $\vec{C}$ defines the set of cells used in the analysis, which had as single cells the most information about which stimulus was present. From the set of cells $\vec{C}$, the firing responses $R_{\vec{C}}$ $(R = r(c)|c \in \vec{C})$ to each stimulus in S are used as the basis for the Bayesian decoding procedure as follows:

$$P(s'|R_{\vec{C}}) = \frac{P(s') \prod_{c \in \vec{C}} P(R_c(s')|s')}{\sum_{s'' \in S} P(s'') \prod_{c \in \vec{C}} P(R_c(s'')|s'')} \tag{1.17}$$

$$P(R_c(s)|s') = \frac{\sum_{t=1}^{nTrans} pdf(R_c(s,t), \bar{R}_c(s'), SD_c(s'))}{nTrans} \tag{1.18}$$

where $nTrans$ defines the number of possible transforms, and *pdf* computes the probability density function of the firing response from a subset of cells when exposed to a stimulus s in the t^{th} transform using the normal distribution with their mean and standard deviation.

For a given set of cells, the probabilities generated by the decoding procedure are factored into a confusion matrix, that matches up the actual input stimuli in $\vec{S}$ with the predicted stimuli in $\vec{S}'$. Here, $P(s_i{}')$ represents the probability that the predicted stimulus $s_i{}'$ is actually the stimulus s_i that is currently presented to the network. A higher value of $P(s, s')$ relative to $P(s)P(s')$ indicates a stronger relationship between s and s'; this information provides the basis for calculating the multiple-cell information analysis.

1.5.3 Estimation of Input Gabor Filters

The feedforward synaptic connections between successive layers are traced back to the retina in order to determine the specific features of a stimulus that drive categorical discrimination among the trained neurons of the output layer. Starting from the target output cell, we select the connections from the previous layer that have the highest weights, repeating this process through successive layers until the connections reach the Gabor filters in the retina. This then allows us to plot the pattern of Gabor input filters which the target output neuron has become tuned to.

1.6 Overview of research conducted

1.6.1 The Neural Basis of Object Shape Representations

As explained earlier, neurons in successive stages of the primate ventral visual pathway encode the spatial structure of visual objects. For example, neurons in V4 encode the conformation of boundary contour elements at a particular position within an object regardless of the location of the object on the retina, while neurons in TEO integrate information from multiple boundary contour elements. The first set of studies conducted in Chapter 2 investigates through computer simulation how these cell firing properties may develop through unsupervised visually-guided learning.

Our model of the primate ventral visual pathway, VisNet, is trained by presenting many different object shapes to the network while the strengths of the synaptic connections between successive layers are modified by a local learning rule. Individual neurons in the model are shown to exploit statistical regularity and temporal continuity of the visual inputs during training to learn firing properties that are similar to neurons in V4 and TEO.

These computer simulations represent an important step towards understanding how the visual system learns to encode the detailed spatial structure of objects within natural scenes. This representation goes beyond mere object recognition, in which neurons simply respond to the presence of a whole object, but provides an essential foundation from which the brain is subsequently able to recognise the whole object. This work has been published in Frontiers in Computational Neuroscience (Eguchi et al., 2015).

1.6.2 The Neural Basis of Face Representations

Experimental studies have shown that neurons at an intermediate stage of the primate ventral visual pathway, occipital face area, encode individual facial parts such as eyes and nose while neurons in the later stages, middle face patches, are selective to the full face by encoding the spatial relations between facial features. In this second set of simulation studies conducted in Chapter 3, a computer modelling study was performed to investigate how these cell firing properties may develop through unsupervised visually-guided learning.

VisNet is trained by presenting many randomly generated faces to the network while a local learning rule modifies the strengths of the synaptic connections between neurons in successive layers. After training, the model is found to have developed the experimentally observed cell firing properties. In particular, I showed how the visual system forms separate representations of facial features such as the eyes, nose, and mouth as well as monotonically tuned representations of the spatial relationships between these facial features. I also demonstrated how the primate brain learns to represent facial expression independently of facial identity.

Furthermore, based on the simulation results, I propose that neurons encoding different global attributes simply represent different spatial relationships between local features with monotonic tuning curves or particular combinations of these spatial relations. This work has been published in Psychological Review (Eguchi et al., 2016).

1.6.3 Neural Basis of Cognitive Bias Modification (CBM) as a Clinical Treatment for Depression

Following the results reported in previous chapters, the study in Chapter 4 tried to investigate the clinical application of the finding in the previous study. Many mental health problems are linked to cognitive biases towards emotionally negative information. For example, depressed patients have a greater tendency to interpret faces as sad, and are less able to detect mildly happy expressions. Recently, interest has grown in a new class of psychological treatments for depression, anxiety, and addictive disorders known as Cognitive Bias Modification (CBM), which can eliminate these underlying negative cognitive biases.

It is thought that the elimination of negative cognitive biases may help to shift the depressed mood state of a patient. The study reported in this chapter uses a computer simulation to investigate the neural and synaptic dynamics underlying two forms of CBM, which may be able to eliminate the negative biases in the way that depressed patients evaluate facial expressions. The new CBM methodologies utilise two previously established biologically plausible synaptic learning mechanisms, continuous transformation (CT) learning and trace learning. These learning mechanisms are able to guide visual development by exploiting either the spatial continuity or temporal continuity between visual stimuli presented during training.

Our simulation results show that both of these learning mechanisms, when combined with carefully designed sequences of transforming face images presented to the model, will eliminate negative biases in the interpretation of facial expression. That is, a sub-population of 'sad' output neurons that initially responds to both sad and neutral faces before learning will only respond to the sad faces after CBM training. Simulations with a simplified one-layer neural network architecture is described first in order to test the two hypothesised CBM learning mechanisms in a highly controlled manner. Then simulation results in which realistic face stimuli are used to train VisNet is presented. This work has been published in Journal of Consulting and Clinical Psychology (Eguchi et al., 2017b).

1.6.4 The Neural Basis of Border Ownership Representations

As Rubin's famous vase demonstrates, our visual perception tends to assign luminance contrast borders to one or other of the adjacent image regions. Experimental evidence for the neuronal coding of such border-ownership in the primate visual system has been reported in neurophysiology studies.

Chapter 5 investigated exactly how such neural circuits may develop through visually-guided learning. More specifically, it was investigated through computer simulation how top-down connections may play a fundamental role in the development of border ownership representations in the early cortical visual layers V1/V2. Our model consists of a hierarchy of competitive neuronal layers, with both bottom-up and top-down synaptic connections between successive layers, and the synaptic connections are self-organised by a biologically plausible, temporal trace learning rule during training on differently shaped visual objects.

The simulations reported in this chapter have demonstrated that top-down connections may help to guide competitive learning in lower layers, thus driving the formation of lower level (border ownership) visual representations in V1/V2 that are modulated by higher level (object boundary element) representations in V4. Lastly the limitations of our model in the more general situation where multiple objects are presented to the network simultaneously were investigated. This work has been published in Neurobiology of Learning and Memory (Eguchi and Stringer, 2016).

1.6.5 Polychronization and Feature Binding in a Spiking Neural Network Model

Chapter 6 investigates the behaviour of a biologically realistic hierarchical neural network model of the primate ventral visual system. It is shown that, even when the input images are represented by randomised Poisson spike trains during training, the network model develops stimulus representations in the form of fixed and regularly repeating temporal chains of spikes emitted by subpopulations of neurons ('polychronization'). It was found that the inclusion of top-down and lateral synaptic connections in the network architecture results in an increase in the number and length of such temporal spiking patterns compared to a purely bottom-up architecture.

The performance of the model could be further improved by including multiple synaptic contacts between each pair of pre- and postsynaptic neurons, with different synaptic contacts having axonal delays of different durations. In this case, STDP enables the network to select

which of the synaptic connections between two neurons to strengthen in order to set the effective axonal transmission delay between the cells. This in turn helps to drive the development of polychronous groups (PGs) of neurons with precise temporal patterns of spiking activity, and hence results in a greater representational capacity for the network. Finally, it was found that the PGs that emerged in the network during visually-guided learning contained a type of neuron, which we have named a 'binding neuron', which represents the binding relationship between low-level and high-level visual features.

These binding neurons provide a solution to the classic feature binding problem in visual neuroscience. Our simulation results suggest that binding is a much richer phenomenon than traditionally described by visual psychologists. Indeed, the binding mechanism proposed here is potentially so rich that it would be difficult to describe the process at a high psychological level; it requires a description at the neuronal level as presented in this chapter. The proposed mechanism for the development of binding neurons leads directly to a new hypothesis concerning how information about visual features at every spatial scale may be projected upwards through successive neuronal layers to the highest (output) layer of the network, which we have termed the *holographic principle*. This might be a useful operation if the subsequent behavioural systems of the brain are limited to reading out visual information from only the later stages of the visual system.

1.6.6 The Neural Basis of Object Shape Representations in a Spiking Neural Network Model

The final study reported in Chapter 7 demonstrates the development of border ownership representations in a hierarchical spiking neural network model. Consistent with the studies in Chapter 6, the visual objects presented to the network are represented as randomised Poisson spike trains in the input layer. During visual training, V4 like neurons develop in the highest (output) layer that encode the conformation of boundary contour elements at a particular position within an object regardless of the location of the object on the retina. At the same time, border ownership cells develop in the lower layers, which respond preferentially to oriented edges and are also modulated by which side of an object or figure the edge occurs on. Importantly, the limitations of the "superposition catastrophe" (von der Malsburg, 1999) in a traditional rate-coded model are overcome within the current spiking neural network model, and the border ownership representations are robust even when multiple objects are presented to the network simultaneously.

Chapter 2

The Neural Basis of Object Shape Representations

Experimental studies have shown that neurons in successive stages of the primate ventral visual pathway encode the spatial structure of visual objects. For example, neurons at an intermediate stage, area V4, encode the conformation of boundary contour elements at a particular position within an object, regardless of the location of the object on the retina. On the other hand, neurons in the later stages TEO and TE integrate information from multiple boundary contour elements. In this chapter, I investigate through computer simulation how these cell firing properties may develop through visually-guided learning, and thus how the primate ventral visual pathway learns to represent the spatial structure of objects.

A biologically plausible neural network model, VisNet, of the primate ventral visual pathway, which consists of a hierarchical series of competitive layers corresponding to successive stages of the ventral visual pathway, is used in this study (See Section 1.3). The model is trained by presenting many different object shapes to the network while the strengths of the synaptic connections between successive layers are modified by a local learning rule that depends on the activities of the pre-synaptic and post-synaptic neurons. After training, the model is found to have developed the experimentally observed cell firing properties found in V4, TEO, and TE.

Two key mechanisms drive the development of the cell firing properties during learning. First, the network architecture is able to exploit the statistical decoupling that exists between different boundary contour elements over a large population of different shapes in order to produce separate neural representations of different boundary contour elements. Secondly, the same neurons can learn to respond with translation invariance as objects shift across the retina through the use of a temporal *trace learning* rule to set up the synaptic connection strengths, which encourages post-synaptic neurons to respond to input patterns that occur in close temporal proximity as described in Section 1.4.2.

These computer simulations represent an important step towards understanding how the visual system learns to encode the detailed spatial structure of objects within natural scenes. This goes beyond mere object recognition, in which neurons may respond to the presence of an undifferentiated object. Instead, the neurons in our model learn to respond to the conformation of localised boundary contour elements at a particular position within an object, irrespective of the object's retinal location. This representation of the spatial form of each object may provide an essential foundation from which the brain is subsequently able to recognise the whole object.

2.1 Introduction

2.1.1 Hierarchical representations in the primate ventral visual pathway

Over successive stages of processing, the primate ventral visual pathway develops neurons that respond selectively to objects of increasingly complex visual form (Kobatake and Tanaka, 1994), going from simple orientated line segments in area V1 (Hubel and Wiesel, 1962) to whole objects or faces in the inferotemporal cortex (IT) (Perrett et al., 1982; Tsao et al., 2003; Tsunoda et al., 2001). In addition, in higher layers of the ventral pathway, the responses of neurons to objects and faces show invariance to retinal location, size, and orientation (Tanaka et al., 1991; Rolls et al., 1992; Rolls, 2000; Perrett and Oram, 1993; Rolls and Deco, 2002). These later stages of processing carry out object recognition by integrating information from more elementary visual features represented in earlier layers (Brincat and Connor, 2004). Thus, in order to understand visual object recognition in the primate brain, we need also to understand the encoding of more elementary features in the early and middle stages of the ventral visual pathway. In particular, many theories suppose that object recognition operates through the computation of intermediate representations which reflect the spatial relations between the parts of objects (Brincat and Connor, 2004; Pasupathy and Connor, 2001; Giersch, 2001).

Experimental studies have shown that neurons in successive stages of the primate ventral visual pathway encode the spatial structure of visual objects and their parts. For example, single unit recording studies carried out by Pasupathy and Connor (2001) have shown that, within an intermediate stage of the ventral visual pathway, area V4, there are neurons that respond selectively to the shape of a local boundary element (e.g., concave or convex) at a particular position in the frame of reference of the object (i.e., with respect to the centre of the object). Some of these V4 neurons also maintain their response properties as an object shifts across different locations on the retina; i.e. they have learned translation-invariant representations. Therefore, these cells encode the spatial form of the object boundary by representing the relations between individual boundary contour elements within the object. A population of such cells will provide a distributed encoding of the complete boundary shape of the object (Pasupathy and Connor, 2002). Further experimental studies have shown that neurons in the later stages of the ventral visual pathway, anterior IT (TEO) and posterior IT (TE), integrate information from multiple boundary contour elements (Brincat and Connor, 2004). This representation of the detailed spatial form of the separate parts of each object may provide a necessary foundation for the subsequent recognition of whole objects. That is, object selective cells at the end of the ventral visual pathway may learn to respond to unique distributed representations of object shape in earlier areas (Booth and Rolls, 1998).

2.1.2 Computer Modelling Study

A number of modelling studies have tried to reproduce the observed shape selective and translation invariant firing properties of neurons in area V4 (Rodrguez-Snchez and Tsotsos, 2012; Cadieu et al., 2007). However, these past models have not utilised biologically plausible learning mechanisms to guide the development of cell firing properties. In particular, previous models have not used plausible, local learning rules, which use pre- and post-synaptic cell quantities to drive modification of the synaptic connections during visually-guided learning. Therefore, it still remains a challenge to understand exactly how V4 neurons develop their shape selective response properties through learning. The purpose of this chapter is to provide the first biologically plausible theory of this learning process. More generally, I investigate through computer simulation how the cell firing properties reported in visual areas V4 and TEO may develop through visually-guided learning, and thus how the primate ventral visual pathway learns to represent the spatial structure of objects.

The simulation studies presented below are conducted with an established hierarchical neural

network model of the primate ventral visual pathway, VisNet (Wallis and Rolls, 1997), described in detail in Section 1.3. The standard network architecture consists of a hierarchy of four competitive neural layers (Rumelhart and Zipser, 1985) corresponding to successive stages of the ventral visual pathway. Within a competitive layer, neurons 'compete' with each other to respond to the current visual input stimulus. Competition between neurons in the brain is mediated by inhibitory interneurons. In VisNet, competition between neurons is effected by local inhibitory filters.

2.2 Hypothesis

In this chapter, I consider how biologically plausible neuronal and synaptic learning mechanisms may be applied to the challenge of explaining (i) how neurons in V4 learn to respond selectively to the shape and location of localised boundary contour elements in the frame of reference of the object, (ii) how neurons in area TEO learn to respond to localised combinations of boundary contour elements, and (iii) how these neurons learn to respond with translation invariance as the object is shifted through different retinal locations. In particular, I hypothesise that a biologically plausible solution may be provided by combining the statistical decoupling (Stringer et al., 2007; Stringer and Rolls, 2008) that will occur between different forms of boundary contour element over a large population of different object shapes, with the use of a temporal trace learning rule (see Section 1.4.2) to modify synaptic weights as objects shift across different retinal locations (Wallis and Rolls, 1997; Rolls, 2000).

2.2.1 Neurons learn to respond to individual boundary contour elements by exploiting statistical decoupling

Previously, Stringer et al. (2007) and Stringer and Rolls (2008) investigated how VisNet may learn transform invariant representations of individual objects if the network is always presented with multiple objects simultaneously during training. This is an important problem to address in order to understand how the visual system learns to recognise individual objects if they are seen in natural scenes with other objects present during training. They have found that if VisNet is trained on different combinations of objects on different occasions and as long as there are enough objects in the total pool of objects, this will result in statistical decoupling between any two objects. This statistical decoupling forces neurons in the higher competitive layers of VisNet to learn to respond to the individual objects, rather than the combinations of objects on which the network is actually trained.

This is because a competitive neural network has a capacity limit in terms of the number of object categories that can be represented in a non-overlapping manner in the output layer. For example, in the simplest situation of winner-take-all competition, the output layer can develop non-overlapping representations of at most n object categories where n is the number of neurons within the layer. With soft competition, in which a small number of output neurons may remain active at a time, the number of non-overlapping output representations possible is further reduced. If the network is trained on more object shapes than can be represented by non-overlapping subsets of output neurons, then the output representations must start to overlap. In this case, the competitive network will start to represent the objects in a distributed, overlapping manner.

Figure 2.1 provides some further insight into the learning mechanisms driving the formation of neurons encoding individual object. Consider the highly simplified situation where, a winner-take-all competitive network with $64 \times 64 = 4096$ output neurons is presented with n different objects, which are presented in pairs to VisNet during training. Will the neurons in the output layer learn to represent the individual objects or the pairs of objects that the network is actually trained on? The governing factor is the capacity limit of the competitive network.

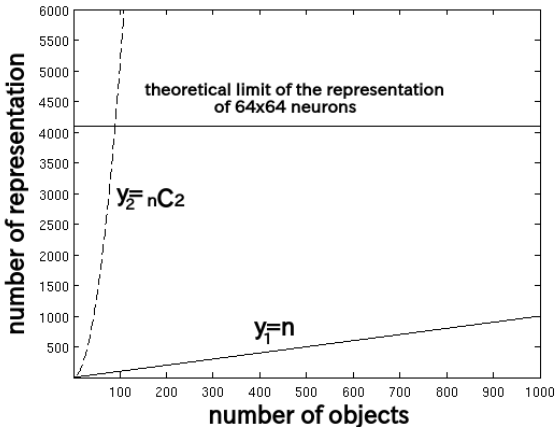

Figure 2.1: How the capacity limit of a competitive neural network forces individual output neurons to switch from representing object shapes to representing the boundary elements as the number of object shapes on which the network is trained increases. Consider a competitive network consisting of $64 \times 64 = 4096$ neurons, which is trained on all possible objects constructed from pairings of n boundary elements. In the simplest case of winner-take-all competition, in which only a single neuron remains active at a time, the network can represent at most 4096 object categories in a non-overlapping manner. The two graphs show how the number of boundary elements, $y1$, and the number of objects comprised of pairs of boundary elements, $y2$, rise with n. The number of objects, $y2$, rises quadratically with n, and thus reaches the capacity limit of the network much more quickly at $n = 91$ boundary elements ($91C2 = 4095$). At this point, the network is forced to switch from representing the objects to representing individual boundary elements.

With winner-take-all competition, the network is able to develop 4096 non-overlapping output representations. Figure 2.1 shows how the number of individual objects, $y_1 = n$, and the number of possible objects comprised of pairs of objects $y_2 =_n C_2 = n(n-1)/2$, rise quadratically with increasing n. Because of this, y_2 reaches the capacity limit of the network much more quickly than y_1. Therefore, for n greater than 91, individual output neurons are forced to switch from representing the objects to representing the individual objects. At the same time, of course, the output layer as a whole will still provide unique representations of the pairs of objects, themselves, but in a distributed, overlapping manner. This effect has been reported by Stringer and Rolls (2008), who found that for small numbers of objects n, the output neurons still represented the paired-stimulus input patterns. However, for large enough n, the output neurons began to learn to respond to the individual objects instead of the multi-object input patterns used during training. This work was extended by Stringer et al. (2007), who showed that the same effect occurred when the network was trained on input patterns composed of three objects.

I now propose that a similar learning mechanism may operate to enable the network to learn to represent the individual boundary contour elements within objects. For example, consider the simplified case shown in Figure 2.2. This figure shows a set of four sided shapes, where each side has one of three possible conformations: concave, straight, or convex. Therefore, there are 4 sides $\times$ 3 side types = 12 different boundary contour elements (each defined by a unique combination of position and shape), which may be used to construct a total of $3^4 = 81$ different whole objects. It is demonstrated that, when VisNet is trained on such a large population of different object shapes constructed from different combinations of boundary contour elements, there is statistical decoupling between any two boundary contour elements. That is, any two boundary contour elements are seen together very infrequently across the entire population of objects. This forces neurons in the intermediate and higher layers of VisNet to learn to respond to the individual boundary contour elements rather than to particular combinations of contour elements representing whole objects.

Figure 2.3 provides an illustration of how the capacity limit forces output neurons to learn to represent individual boundary elements. Figure 2.3(left) shows two different object shapes

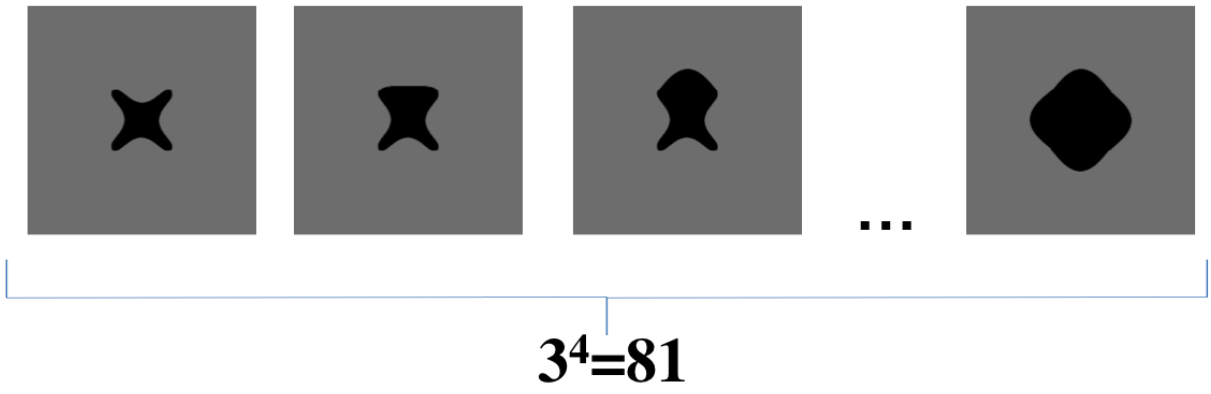

Figure 2.2: Form of visual objects used to train and test VisNet for Study 1. The objects are constructed from a pool of boundary contour elements. Specifically, each object has a fixed number of sides (n), each of which has a fixed number of possible boundary conformations (p). In the figure, examples of objects with 4 sides, each of which has three possible conformations (concave, straight and convex) are shown. This gives a total number of $3^4 = 81$ objects.

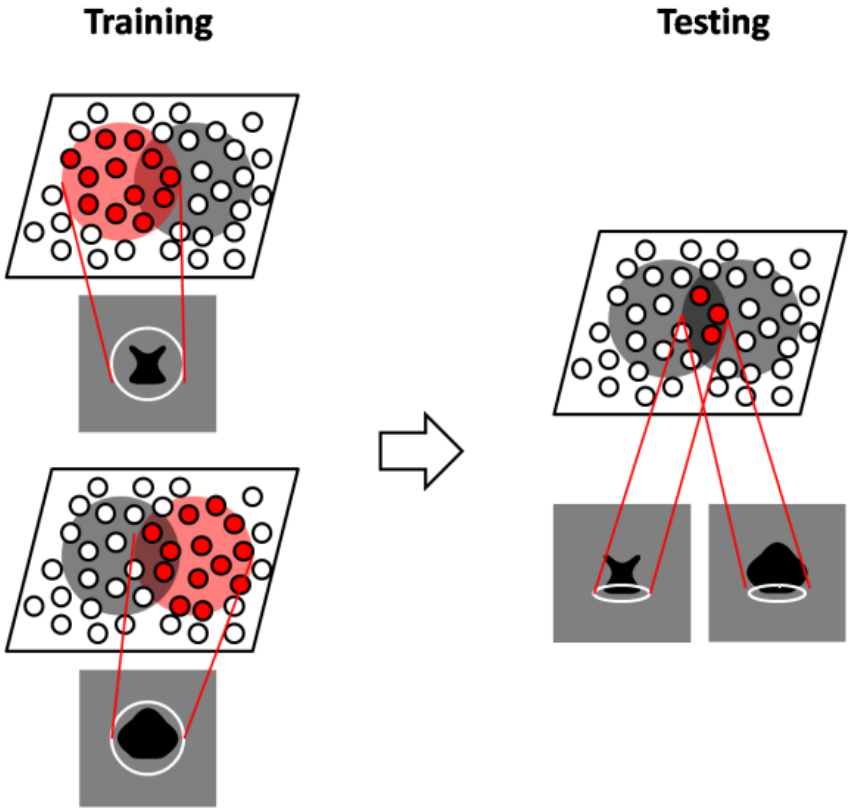

Figure 2.3: Illustration of how the network model develops neurons that have learned to respond to individual boundary elements of 2D object shapes. Left: during training, the network is presented with many different 2D object shapes, where each shape is defined by a unique combination of boundary elements of different curvatures. Two such shapes are shown here. Each of these shapes stimulates a different subset of neurons in the output layer of the network. The two shapes shown have one boundary element in common, which is located at the bottom of the two shapes. This boundary element becomes especially strongly connected, through associative learning in the feed-forward synaptic connections, with the intersection of the two subsets of output neurons shown. This intersecting subset of neurons will come to represent the boundary element present at the bottom of the two shapes. Right: during testing, whenever the particular boundary element is part of a 2D object shape, the same intersecting subset of output neurons will be activated. A similar learning process will drive the development of many other subsets of output neurons representing different localised boundary elements.

presented to the network during training. Each of the two objects stimulates a subset of output neurons. The number of output neurons activated by each shape will depend on the level of competition between neurons in the layer, which in the brain is mediated by inhibitory interneurons. In the VisNet model used in this chapter, the level of competition is implicitly modelled by specifying the sparseness of the firing rates within each layer, which is itself controlled by adjusting the sigmoid threshold α. For the simplified case of neurons with binarised firing rates $= 0/1$, the sparseness is defined as the proportion $\in [0, 1]$ of neurons that are active. For example, to set the sparseness to 5%, the sigmoid threshold α is set to the value of the 95th percentile point of the activations within the layer. In this case, the network can represent up to twenty objects with non-overlapping sets of output neurons.

However, if the network is trained on more than twenty objects, then the network is forced to develop overlapping distributed representations of the object shapes. This situation is shown in Figure 2.1(left). Here we see that the two objects stimulate overlapping subsets of output neurons. Thus, the key point is that if there are more object shapes than can be represented by non-overlapping subsets of output neurons, then the output representations must start to overlap. What, then, will be the nature of these distributed representations? What object components will individual neurons in the output layer learn to represent? In Figure 2.3(left) the two object shapes share a boundary element at the bottom of the two shapes. This boundary element becomes especially strongly associated, through associative learning in the feed-forward synaptic connections, with the subset of output neurons at the intersection of the two object shape representations. Figure 2.3(right) shows that during testing this intersecting subset of output neurons will respond whenever the network is presented with an object shape containing the given boundary element. In this way, it can be seen how output neurons in fact learn to respond to the individual boundary elements of objects. In this manner, without any top-down information transfer, the network should be able to develop representations of localised boundary elements. This kind of the distributed coding of 2D object shape, utilising an alphabet of localised boundary elements, may be used to represent the shape of any object. This will potentially explain an unsupervised processing of the input images that leads to the efficient code based on shared features similar to the fruit of the algorithm reported in Torralba et al. (2007) by reducing the computational and sample complexity.

2.2.2 Neurons develop translation invariant responses through trace learning (temporal association)

Another key property of the neurons reported by Pasupathy and Connor (2001) in area V4 and neurons reported by Brincat and Connor (2004) in area TEO is that they respond with translation invariance as an object shifts across different locations over the receptive field. The question is how these neurons might learn to respond in such a translation invariant manner?

One possible explanation is that the brain uses temporal associative learning to develop such transformation invariant representations as described in Section 1.4.2. The theory assumes that, every now and then, a primate will make a series of fixations at different points on the same visual object before moving onto another object; much experimental work has studied the statistics of saccades and fixations across natural visual scenes (Findlay and Gilchrist, 2003). Of particular relevance is how the eyes saccade around natural visual scenes containing multiple objects. Seminal psychophysical studies of how human subjects move their gaze around pictures of natural scenes were carried out by Yarbus (1967). It was indeed evident from this work that there was a tendency for observers to shift their fixation to a number of different points on a salient object, such as a person, before moving onto the next object.

It is still unclear that whether such learning can occur across shifts in fixation. Therefore, in the simulation study, the eye movements are assumed to be sufficiently small so that the same object is always projected within the simulated receptive field when learning it. I believe this constraint is reasonable for simulating recent physiological findings. For example, Li and

DiCarlo (2008) conducted a study where monkeys are trained to track an object on a screen where an object with identity A is originally placed on one of two possible retinal positions (+3 or −3 degrees) and later shifted to the centre (0 degree). In the experimental condition, the identity of the object is swapped from A to B when it is shifted to the centre, and the eyes saccade to it. As a result, individual neurons in primate IT that are originally selective to object A start to respond also to object B at the central retinal location. This finding does not exclude the possibility of the temporal association learning which may occur at larger eye movement; however, it provided a reasonable evidence for the translation invariance learning mechanism within IT (Isik et al., 2012).

Accordingly, our proposed solution is *temporal trace learning* (Foldiak, 1991; Wallis and Rolls, 1997; Rolls and Milward, 2000). An example of such a learning rule is given in equations (1.8) and (1.9) in Section 1.4.2. This learning rule incorporates a memory trace, $\overline{r}_i^{\tau-1}$, of recent neuronal activity. The effect of trace learning in VisNet is to encourage neurons to learn to respond to visual input patterns that tend to occur close together in time. If the eyes shift about a visual scene more rapidly than the objects change within the scene, then the images of an object in different locations, scale, or orientations on the retina will tend to be clustered together in time. In this case, a trace learning rule will encourage neurons in higher layers to learn to respond with transform invariance to specific objects or features.

This rule is biologically plausible in terms of the way it utilises only locally available biological quantities, that is, the present and recent activities of the pre- and post-synaptic neurons respectively. Also, it has been shown that this type of temporal associative learning has been shown to arises naturally within biophysically realistic spiking neural networks when longer time constants for synaptic conductance are introduced (Evans and Stringer, 2012), of which I also investigated in Chapter 7.

Some of the past research have shown that this trace learning rule may be combined with the mechanism of statistical decoupling described above to produce translation invariant representations of statistically independent visual objects (Stringer et al., 2007; Stringer and Rolls, 2008). It can be thus hypothesised that the same trace learning rule could encourage neurons representing boundary contour elements to respond with translation invariance across different retinal locations.

2.2.3 Overview of simulation studies carried out in this chapter

Study 1 provides a proof-of-principle analysis. VisNet was trained on artificial visual objects similar to those shown in Figure 2.2. For each simulation, the visual objects had a fixed number of sides (n), each of which had a fixed number of different possible boundary conformations (p). These carefully constructed objects allowed us to explore how the statistical decoupling between different boundary contour elements influences the neuronal firing properties that develop during learning. I also showed how the capacity of the network to represent many different boundary contour conformations can be increased by introducing a Self-Organising Map (SOM) architecture within each layer (Section 1.3.3.2). Finally, the same artificial visual stimuli was used to confirm that trace learning can produce neurons that respond to individual boundary contour elements with translation invariance across different retinal locations.

In Study 2, the sets of visual stimuli presented to VisNet during training and testing were similar to those used in the original physiological experiments of Pasupathy and Connor (2001). Examples are shown in Figure 2.4. This allowed for a direct comparison between the performance of the VisNet model and real neurons recorded in area V4 of the primate ventral visual pathway. It was demonstrated that, using the computational principles discussed above, VisNet developed neurons during visually-guided training that have firing properties similar to those reported by Pasupathy and Connor (2001). That is, the neurons learned to respond to local boundary contour elements defined by a specific combination of their curvature and position within the whole stimulus.

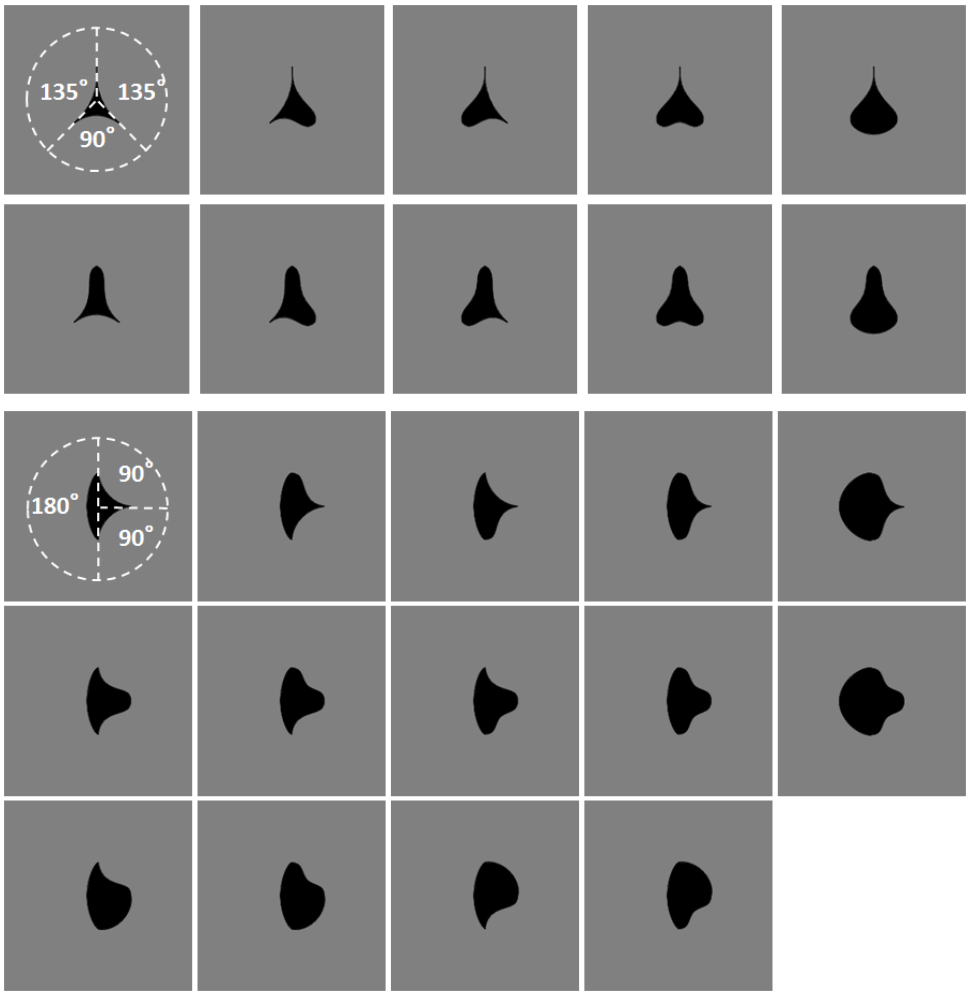

Figure 2.4: Form of visual objects used to train and test VisNet in Study 2. These visual stimuli are similar to those used in the original neurophysiological experiments of Pasupathy and Connor (2001). This permitted direct comparison between the performance of the VisNet model and real neurons recorded in area V4 of the primate ventral visual pathway. The stimuli were created by systematically combining convex and concave boundary elements. Each shape is defined by the configuration of three convex projections. The convex projections ended in either sharp angles or medium curves. The angular separations between the convex projections were either 135°/135°/90° (top) or 180°/90°/90° (bottom). The convex projections were connected by concave and convex circular arcs. As an example, the object shown in the top left plot had three sharp convex projections separated by angles 135°/135°/90°, which are connected by concave circular arcs.

In Study 3, VisNet was trained on a large number of realistic visual objects with different boundary shapes. A sample of these objects is shown in Figure 2.5. This generated a more realistic and demanding test of the underlying theory. Across a large set of different object shapes, there should be statistical decoupling between localised boundary contour elements of different conformation. After training on real objects, the network was tested using visual stimuli similar to those originally used by Pasupathy and Connor (2001) and shown in Figure 2.4. In support of the theory, it was found that neurons again learned to respond to local boundary contours defined by their curvature and position with respect to the whole stimulus.

The computer simulations reported in this chapter are designed to test the feasibility and efficacy of the computational principles described above, as well as build an intuitive understanding of these mechanisms. The simulations thus make an important contribution towards understanding how the visual system may learn to represent the detailed spatial structure of objects. These visual representations go beyond the mere recognition of objects, in which neurons simply respond if a particular object is present. Instead, the neurons found in area V4 by Pasupathy and Connor (2001) and reproduced in our model learn to respond to the conformation

Figure 2.5: Examples of some of the realistic visual objects used for training VisNet in Study 3. In total, there were 177 realistic objects used for training VisNet in this study.

of localised boundary contour elements at a particular position within an object, irrespective of where the object is on the receptive field. This representation of the spatial form of individual objects may provide a foundation of intermediate-level representations of object parts, which the brain subsequently utilises to recognise whole objects within natural scenes.

2.3 Materials & Methods

2.3.1 VisNet

VisNet is a hierarchical neural network model of the primate ventral visual pathway, which was originally developed by Wallis and Rolls (1997) (see Section 1.3). The standard network architecture is shown in Figure 1.1. It is based on the following: (i) A series of hierarchical competitive layers with local graded lateral inhibition. (ii) Convergent connections to each neuron from a topologically corresponding region of the preceding layer, leading to an increase in the receptive field size of neurons through the visual processing areas. (iii) Synaptic plasticity based on a biologically-plausible local learning rule such as the Hebb rule (1.7) or trace rule (1.8) and (1.9), which are explained in Section 1.3.5.

In past work, the hierarchical series of 4 neuronal layers of VisNet have been related to the following successive stages of processing in the ventral visual pathway: V2, V4, the posterior IT (TEO), and the anterior IT (TE). In this chapter, neuronal response properties observed within a series of intermediate layers, V4 and TEO, of the ventral pathway were modelled for the first time. Due to the relatively course-grained 4-layer architecture of VisNet, I do not wish to emphasise a specific correspondence between the layers of VisNet and particular stages of the ventral pathway. However, as our main focus was on the neuronal properties reported in V4 and TEO, the focus is mostly set on the first three layers of VisNet.

In VisNet, 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 contained approximately 67% of the connections from the preceding layer. Typical values employed in the current studies are given in Table 2.1, which have been proposed to be realistic in Wallis and Rolls (1997). These were the values used unless otherwise stated. The gradual increase in the receptive field of cells in successive layers reflects the known physiology of the primate ventral visual pathway (Freeman and Simoncelli, 2011; Pasupathy, 2006; Pettet and Gilbert, 1992).

Table 2.1: VisNet parameters

Layer	Dimensions	number of connections	radius
Layer 4	128×128	100	12
Layer 3	128×128	100	9
Layer 2	128×128	100	6
Layer 1	128×128	201	6
Retina	$256 \times 256 \times 16$		

The parameters for the sigmoid activation function are shown in Table 2.2. These are the standard parameter values that have been used in past VisNet studies (Stringer et al., 2006, 2007; Stringer and Rolls, 2008).

Table 2.2: Parameters for Sigmoid activation function

Layer	1	2	3	4
Percentile	99.2	98	88	91
Slope (β)	190	40	75	26

In each experiment, an array of Gabor filters is generated at each of 256×256 retinal locations with the parameters given in Table 2.3 based on the Equations described in Section 1.3.1. The outputs of the Gabor filters are passed to the neurons in layer 1 of VisNet according

Table 2.3: Parameters for Gabor input filters

Parameter (Symbol)	Value
Wavelength(λ)	2
Spatial bandwidth (σ)	1.5 octaves
Orientation(θ)	$0, \pi/4, \pi/2, 3\pi/4$
Phase shift (ψ)	0: white on black bar π: black on white bar
Aspect ratio (γ)	0.5

to the synaptic connectivity given in Table 2.1. Each layer 1 neuron received connections from 201 randomly chosen Gabor filters localised within a topologically corresponding region of the retina. In the current simulations reported here, the model used inputs from only the shortest wavelength filters, which was found to be sufficient to represent the simple visual objects. For consistency with past VisNet simulations, each neuron in the first layer of VisNet received afferent connections from 201 of the short wavelength filters.

2.3.1.1 Lateral inhibition/excitation between neurons within each layer

In the simulations reported below, the lateral inhibition between the neurons within each neuronal layer was implemented in one of two different ways. The simplest approach was to implement a competitive network architecture (Rolls and Treves, 1998), in which neurons inhibited all or some of their neighbours (see Section 1.3.3.1). However, in some simulations a more complex Self-Organising Map (SOM) architecture (Kohonen, 1982), which included both short range excitation and longer range inhibition between neurons (i.e., a 'Mexican hat' connectivity), was implemented. A SOM architecture leads to a map-like arrangement of neuronal response characteristics across a layer after training, with nearby cells responding to similar inputs (see Section 1.3.3.2). In particular, I investigated the hypothesis that the SOM architecture could increase the capacity of the network by enabling neurons in the higher layers to discriminate between more boundary contour shapes. Parameters shown in Table 2.4 and 2.5 were selected based on those that previously optimised performance (Tromans et al., 2011; Rolls and Milward, 2000). The lateral inhibition parameters for the competitive network architecture are given in Table 2.4.

Table 2.4: Lateral inhibition parameters for the competitive network architecture

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

The lateral inhibition and excitation parameters used in the SOM architecture are given in Table 2.5.

In the simulations reported in Section 2.4, each of the artificial or natural visual objects were constructed from a large pool of local boundary contour elements. During training, many objects constructed from different combinations of boundary elements were presented to the network. Due to the effects of statistical decoupling between the boundary elements, neurons in the higher layers of the model learn to respond to individual boundary elements, or small localised clusters of boundary elements.

By using a combination of trace learning and training on many different object shapes, the model is thus able to simulate and illuminate the learning mechanisms underpinning visually-guided development in neurons that encode the local boundary conformation of objects as reported by Pasupathy and Connor (2001), in area V4, and by Brincat and Connor (2004), in area TEO.

Table 2.5: SOM parameters

Layer	1	2	3	4
Excitatory Radius (σ_E)	1.4	1.1	0.8	1.2
Excitatory Contrast (δ_E)	5.35	33.15	117.57	120.12
Inhibitory Radius (σ_I)	2.76	5.4	8.0	12.0
Inhibitory Contrast (δ_I)	1.5	1.5	1.6	1.4

2.3.2 Modification of Information Analysis

To quantify the performance in transformation invariance learning with VisNet, the techniques of Shannon's information theory have previously been used (Rolls and Treves, 1998), which is based on the Kullback-Leibler divergence of the conditional response distribution from the unconditional distribution. As described in Section 1.5.2, information theory can be used to quantify how selective neurons are for particular boundary elements, each of which occurs within a subset of objects, which may translate across different locations on the retina. If the responses r of a neuron carry a high level of information about the presence of a particular boundary element stimulus s within an object, then this implies that the neuron will respond selectively to the presence of that boundary element across the complete subset of objects that feature it. Furthermore, for simulations in which the objects are shifted across different locations on the retina, a high level of information would imply that the neuron responds to the presence of that boundary element within an object regardless of where the object is presented on the retina. In this way, information theory can provide a direct measure of both the selectivity of a neuron for a particular boundary element, as well as how translation-invariant the neuronal responses are as the object is shifted across the retina.

Two information measures were used to assess the ability of the network to develop neurons that are selective to the presence of individual boundary contour stimuli but also invariant to their occurrence within different objects and in different retinal locations. These two measure use the responses from either individual neurons (single-cell information analysis) or small ensembles of neurons (multiple-cell information analysis), each of which will be discussed in turn. The standard implementation of the analysis is described in Chapter 1.5.2.

2.3.2.1 Single-cell information analysis

A single cell information measure was first applied to analyse the responses of individual cells. To be informative in the context of this study, the responses of a given neuron (r) should be specific to a particular contour that appears at a particular side (s), and independent of the remaining global form of the object or retinal location. The amount of stimulus-specific information that a certain cell transmits is calculated from the formula (1.12) described in Section 1.5.2.

Table 2.6 shows an example of a cell that is ideally developed to respond to such a stimulus across all the objects containing it. In this experiment, the number of sides (n) is three and the number of contours on a side (p) is two. In past research with VisNet, this single-cell information analysis was used when only one object was presented to the network at a time. Therefore, the maximum information that an ideally developed cell could carry was log_2(number of stimuli). However, in this study, the complete object shape (composed of n contours) is presented. Therefore, this would have been conceptually equivalent to always presenting n stimuli simultaneously, thus altering the maximum attainable value of the single-cell information.

Suppose in an experiment where $n = 3$ and $p = 2$, there is a cell that is ideally developed to respond to any object containing a convex curve at the top. In this case, a single cell's response to just one shape cannot provide sufficient information to determine whether the cell is responding to the convex curve on the top or any other contours on the other two sides. However, it is still apparent that such an ideal cell will not respond to any object where its preferred contour is absent. Therefore, even though the maximum amount of information conveyed cannot reach

Table 2.6: Simple example of a calculation of single cell information

	Firing rates of a cell X				$P(s_i)$
	$r_1[0-0.25)$	$r_2[0.25-0.5)$	$r_3[0.5-0.75)$	$r_4[0.75-1]$	
contour 1	0	0	0	4	4/24
contour 2	4	0	0	0	4/24
contour 3	2	0	0	2	4/24
contour 4	2	0	0	2	4/24
contour 5	2	0	0	2	4/24
contour 6	2	0	0	2	4/24
$P(r_i)$	12/24	0/24	0/24	12/24	

$$\begin{aligned} I(s_1, \vec{R}) &= P(r_1|s_1)\log_2(P(r_1|s_1)/P(r_1)) + ... + P(r_4|s_1)\log_2(P(r_4|s_1)/P(r_4)) \\ &= 0 + 0 + 0 + (4/4)\cdot log_2((4/4)/(12/24)) \\ &= \log_2(2) = 1bit \end{aligned}$$

$log_2(n \times p)$ bits, the cell still carries up to $log_2(p)$ bits of information, as the cell responses are independent between the p contours on the same side.

In accordance with this, the single-cell information measure in this chapter is modified to calculate the conditional probabilities, not from the individual responses of cells to particular stimuli, but from their average responses across the set of objects with one particular feature held constant.

2.3.2.2 Multiple-cell information analysis

While useful in assessing the tuning properties of a particular neuron, the single-cell information measure cannot give a complete assessment of VisNet's performance with respect to recognition of the set of boundary contour elements as explained in Section 1.5.2.2. We, therefore, also calculated a multiple-cell information, which assesses the amount of mutual information that is available about the whole set of boundary elements from a *population* of neurons. This measure quantifies the network's ability to tell which stimulus is currently exposed to the network based on the set of responses, $\vec{R}$, of a sub-population of cells. Here the procedures for calculating the multiple-cell information measure as described by Rolls and Treves (1998); Rolls and Milward (2000) is adopted.

A decoding procedure is used to estimate which boundary element stimulus s' gave rise to the particular firing rate response vector on each trial. A probability table is then constructed between the real boundary element stimuli, s and the decoded stimuli, s'. From this probability table, the multiple-cell information is then calculated with the equations (1.14). For further details of this decoding, see the procedure explained in Section 1.5.2.2.

In past experiments, the confusion matrix was constructed based on the firing responses to each individual input stimulus. However, as discussed in the single-cell information theory section, it is impossible to determine the specific contour that elicited a particular neural response because n contours are always presented simultaneously. Therefore, instead of using the neuron's firing responses to individual stimuli, the responses are averaged across multiple objects which all feature a common contour. It is these averaged responses which are then used to estimate the required probability distributions.

2.4 Simulation Studies

2.4.1 Study 1: VisNet simulations with artificial visual objects constructed from multiple boundary elements

In Study 1, VisNet was trained on artificial visual objects similar to those shown in Figure 2.2. For each simulation, these visual objects had a fixed number of sides (n), and the curvature of each side was selected from a fixed number of different boundary conformations or elements (p) and were projected on 256×256 pixels of simulated retina. Therefore, for each simulation

there were p^n complete objects constructed from all combinations of the $n \times p$ contour elements. These artificially constructed objects allowed us to investigate how the learned neuronal response properties are affected by the number of object sides and number of possible boundary conformations at each side. The development of translation invariance as objects are shifted by 10 pixels at a time over a grid of 4 different locations on the retina by utilising the trace learning mechanism discussed above was then investigated . In the final simulations of the study, the effects of rotating the objects on the retina by 2 degrees at a time was investigated.

2.4.1.1 Development of neurons that respond to localised boundary conformation

I began by demonstrating how neurons in the output layer learn to respond to individual boundary contour elements when VisNet, implemented with competitive network, is trained on whole objects comprised of a number of such boundary elements. During training, the feedforward synaptic connections were modified using the Hebb learning rule (1.7).

VisNet was first trained on a set of stimuli with $n = 3$ sides: top, left and right. Each side has two possible boundary conformations: concave and convex. This gave a total of $2^3 = 8$ objects. As conceptually the third layer of VisNet may represent TEO, the VisNet architecture consisted of three competitive network layers in this simulation.

Figure 2.6 shows the learned responses y, given by equation (1.6), of a typical output cell in layer 3 of VisNet, which developed selectivity to a concave contour situated at the top of each object after training; the criteria of the selectivity is whether the cell responds with a firing rate, r, approximately equal to 1 ($1.00000 \geq r \geq 0.99995$) across a set of whole objects containing a concave contour on the top while the cell responds with a firing rate approximately equal to 0 ($0.00005 > r \geq 0.00000$) across a set of whole objects not containing a concave contour on the top (i.e., cells with nearly maximum single cell information).

Figure 2.6 (top) shows a histogram of the average firing rate responses of the neuron to 6 (overlapping) subsets of objects, where each subset contains all those objects that incorporate a particular one of the 6 contour elements. Figure 2.6 (bottom) shows the actual subsets of objects that correspond to the 6 data points shown in the histogram. The results confirm that the neuron responds selectively.

Figure 2.7 shows the input Gabor filters that the same output cell in layer 3 has learned to respond to after training. Specifically, Figure 2.7 plots the Gabor filters with the strongest connectivity to the output cell, where each Gabor filter is weighted by the strengths of the connections from that filter through successive layers to the output neuron (see Section 1.5.3). In this case, the neuron receives the strongest inputs from a subset of Gabor filters that represent a concave contour on the top of each object.

Figure 2.8 shows results for six neurons in layer 2 that have learned to respond selectively to different boundary elements. For each neuron, the input Gabor filters that have the strongest connectivity through successive layers to that neuron are shown. These plots show directly the boundary contours that each of the neurons has learned to respond to. In this case, the six neurons learned to respond to the following boundary elements: (a) right/concave, (b) right/convex, (c) left/concave, (d) left/convex, (e) top/concave, (f) top/convex. Figure 2.8 also shows the firing rate responses of the neurons to the eight objects that are constructed from these boundary elements. After training, the six neurons learned to respond maximally to the four objects that contain their preferred boundary element, but do not respond to the other four objects. Such cells for all six of the contours in both layer 2 and layer 3, which are roughly corresponding to V4 and TEO of ventral visual pathway were found.

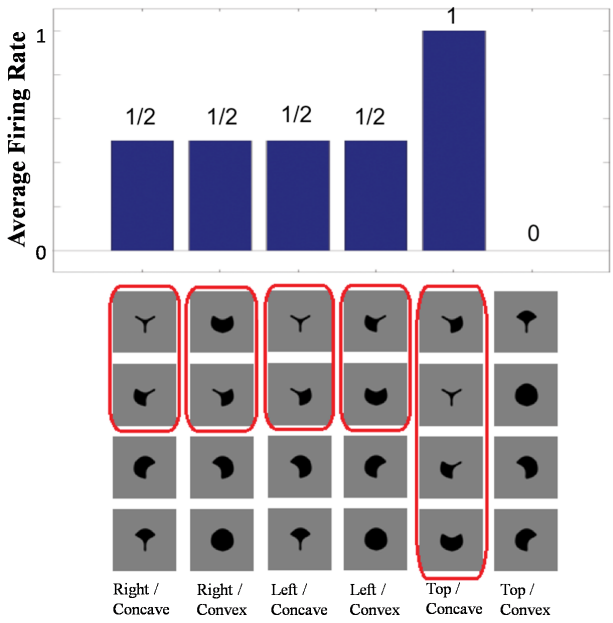

Figure 2.6: The responses of a cell in the output (3rd) layer of VisNet that developed selectivity to the concave contour on the top of each object. The simulation was performed with a competitive network architecture incorporated within each layer of the model. Top: a histogram of the average firing rate responses of the neuron to 6 (overlapping) subsets of objects, where each subset contains all those objects that incorporate a particular one of the 6 contour elements. For example, the first datum plotted shows the average response of the neuron to all those objects that contained a concave contour on the right, which is designated by 'right/concave'. The next 5 data points represent the average response of the neuron to subsets of objects incorporating the following contour elements: right/convex, left/concave, left/convex, top/concave, top/convex. Bottom: the actual subsets of objects that correspond to the 6 data points shown in the histogram. The objects that the cell responds to are ringed in red. Since the firing rates of neurons were effectively binarised (0/1) by the steep slopes β used to parameterise the sigmoid transfer functions, a cell was categorised as responsive if its firing rate was approximately 1. It can be seen that the neuron responds with an average firing rate of 1 to the subset of objects that contain a concave contour on the top. Thus, the neuron always fires maximally to any object containing this particular contour. The same neuron has an average firing rate of zero to the subset of objects with a convex contour on the top. Thus, the neuron fails to respond if the object does not have a concave contour on the top. The neuron has an average firing rate of 1/2 to the other four subsets of objects, right/concave, right/convex, left/concave, left/convex, because the neuron responds maximally to the half of the objects in these subsets that contain a concave contour on the top.

2.4.1.2 How the responses of neurons to their preferred boundary elements depend on the position of the boundary element in the frame of reference of the object

Additional simulations investigated how the responses of neurons to their preferred boundary element depended on the position of the boundary element with respect to the object. In these simulations, VisNet, implemented with competitive networks, was trained on objects constructed with $n = 4$ sides: top, bottom, left, and right. Each side had $p = 3$ possible boundary conformations: concave, straight, and convex. During training, the feedforward synaptic connections were modified using the Hebb learning rule (1.7).

After training, the network was tested to find an output neuron that had learned to respond to a straight contour on a particular position of each object. In the case of neurons that become selective to a straight vertical contour specifically on the right of each object, it would be expected that such a neuron to receive a strong input from a vertical straight contour, with somewhat weaker inputs from other boundary contours on the left of the vertical straight contour. In order for the output neuron to respond selectively to the vertical straight contour on the right of all possible objects, the neuron must receive inputs from all possible boundary contours in all of the other locations on the left, but which will be somewhat weaker than the inputs from the vertical straight contour on the right. This will make the output neuron sensitive to the local context in which the vertical straight contour occurs; i.e. the vertical straight contour

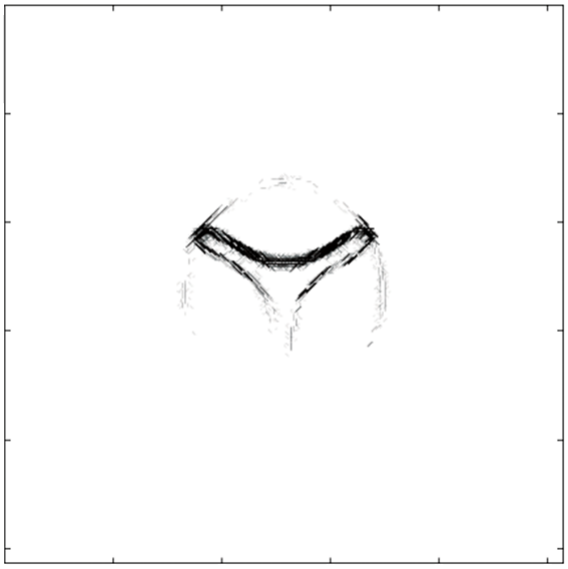

Figure 2.7: The input Gabor filters that an output cell in layer 3 has learned to respond to after training. This is the same neuron whose firing rate responses are plotted in Figure 2.6. Specifically, the Gabor filters with the strongest connectivity to the output cell are plotted, where each Gabor filter is weighted by the strengths of the feed-forward connections from that filter through successive layers to the output neuron in layer 3. It can be seen that this neuron receives the strongest inputs from a subset of Gabor filters that represent a concave contour on the top of each object.

must be situated on the right of a collection of other boundary contours defining the left part of the object.

Figure 2.9 shows the input Gabor filters that had strong connectivity through the layers to such a neuron. The plot is dominated by a strong vertical straight bar on the right hand side. This shows that the neuron has learned to respond to a straight contour on the right of each object. However, the activity of the neuron will also be influenced by other less strong filters shown in the plot. These additional filters extend furthest to the left of the dominating vertical straight bar. In particular, the strong input filters to the left of the vertical straight bar represent boundary contour features that could co-occur within an object with the vertical straight contour on the right. The same is not true for the curve on the right of the vertical straight bar, which joins the same two vertices linked by the vertical straight bar and so would have to be an alternative contour element to the vertical straight bar. The effect of this pattern of additional input filters is that the neuron may require the presence of additional object contours to the left of the vertical straight contour in order for the neuron to respond. That is, the neuron will only respond to a vertical straight contour when that particular contour shape is on the right hand side of an object rather than the left of the object.

This was confirmed in further simulations in which the responses of the neuron were recorded as VisNet was tested with two sets of objects. The first set contained those 4-sided objects from the original training set that had at least one straight contour element, either on the right, bottom, left, or top. The second set contained mirror images of the first set of objects. The mirror images were constructed by reflecting the original trained objects around the retinal location of the vertical straight contour on the right of the training objects. An example of an object and its mirror image is shown in Figure 2.10(a). It can be seen that the vertical straight contours on the right and left of the two objects are aligned on the retina. This procedure ensured that the right and left vertical straight contours from the first and second sets of objects, respectively, overlapped on the retina. I wondered whether the neuron would not only respond to the mirrored object but also to the mirror image objects with a vertical straight contour on the left. This should not happen if the neuron has also learned about the

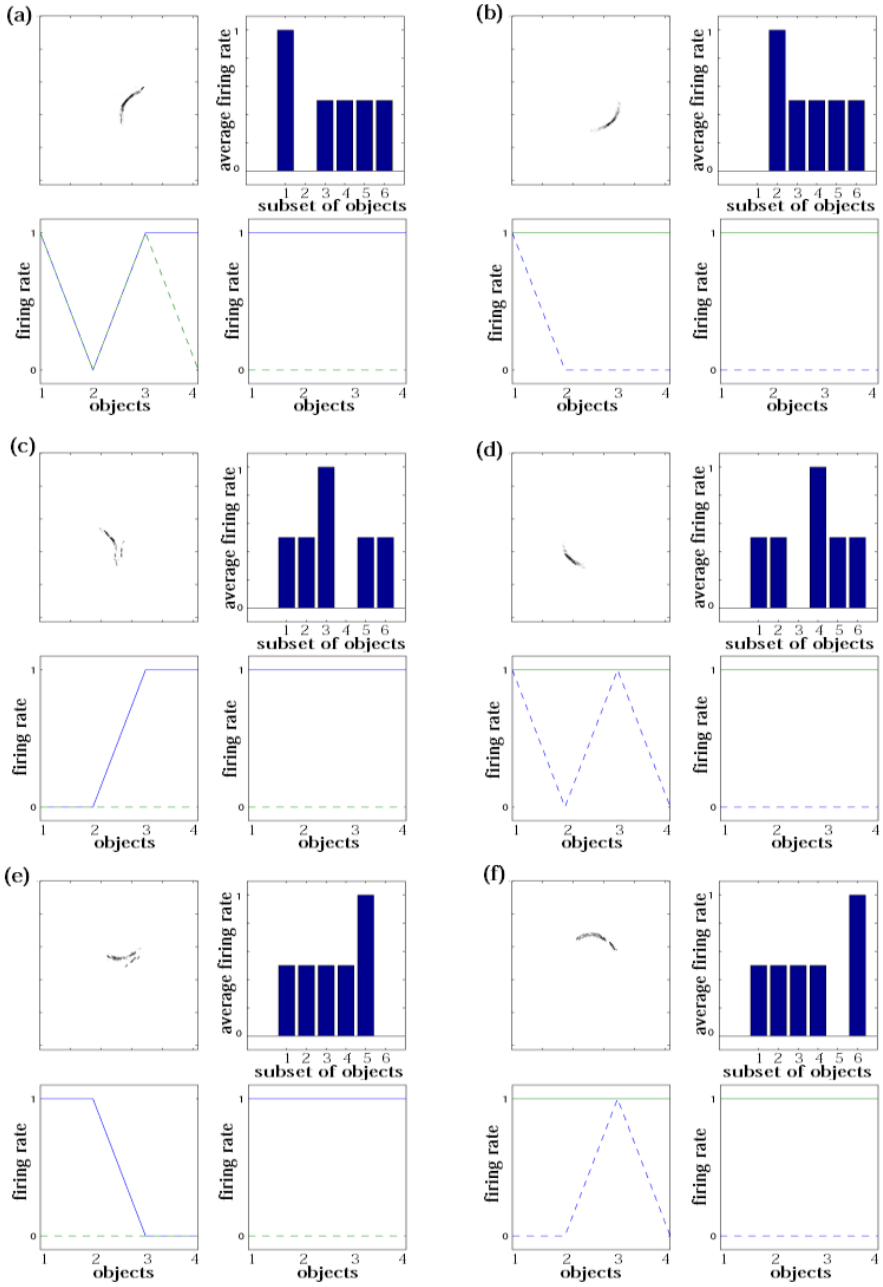

Figure 2.8: Results for six neurons (a)-(f) in layer 2 that learned to respond selectively to different boundary elements. The simulations were performed with a competitive network architecture implemented within each layer. For each neuron, the following four subplots were shown. Top left: input Gabor filters that have the strongest connectivity through successive layers to the neuron. The input Gabor filters immediately show which boundary element each of the six cells has learned to respond to. From these plots, it can be seen that the six neurons have learned to respond to the following boundary elements: (a) right/concave, (b) right/convex, (c) left/concave, (d) left/convex, (e) top/concave, (f) top/convex. Top right: histogram showing average firing rate response of the neuron to the six subsets of objects that contain one of the different boundary elements. That is, each of the data points (1-6) represents the average firing rate of the neuron across the four objects containing the following boundary elements: (1) right/concave, (2) right/convex, (3) left/concave, (4) left/convex, (5) top/concave, (6) top/convex. Bottom left: the firing rate responses of the neuron to all eight objects before training. Results are shown separately for the four objects that contain the neuron's preferred boundary element (solid line) and the four objects that do not contain the neuron's preferred boundary element (dashed line). Bottom right: the firing rate responses of the neuron after training. It can be seen that after training, the six neurons have learned to respond maximally to the four objects that contain their preferred boundary element, but do not respond to the four objects that do not contain that boundary element.

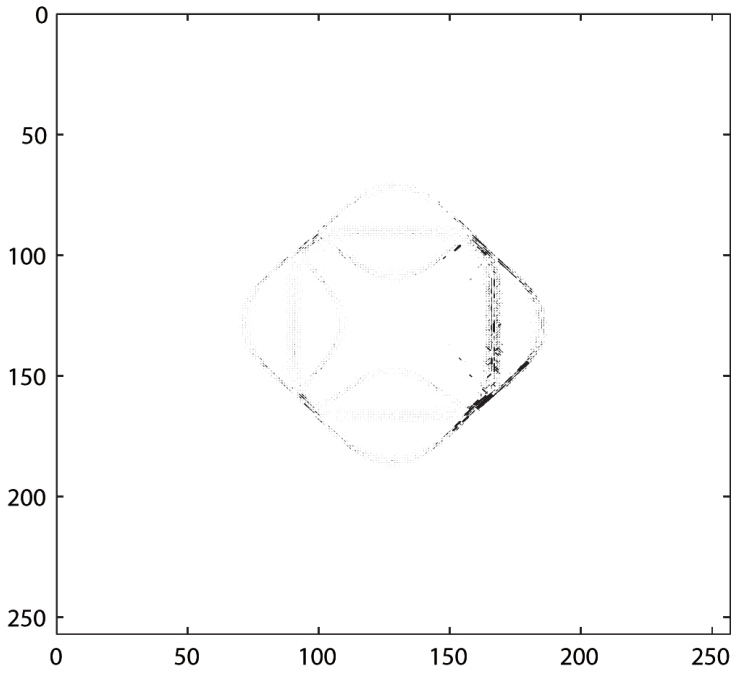

Figure 2.9: Demonstration of neuronal response tuning in an object centred frame of reference. For these simulations, a competitive network architecture was implemented within each layer of VisNet. VisNet is trained on objects with four sides: top, bottom, left and right. Each side has three possible boundary elements: concave, straight and convex. The plot shows the input Gabor filters that had strong connectivity through the layers to a neuron that had learned to respond to a straight contour on the right of each object. There is a dominating vertical straight bar on the right hand side of the plot, which shows that the neuron had learned to respond to a straight contour on the right of the objects. However, there are some other less strong input filters present in the plot, which will also influence the responses of the same output neuron. This pattern of additional input filters indicates that this neuron may require the presence of additional object contours to the left of the vertical straight contour in order for the neuron to be activated. That is, the neuron will only respond to a vertical straight contour when that contour is on the right but not the left of an object.

local image context represented by nearby input filers, as shown in Figure 2.9, and has thereby learned to respond selectively only when the bulk of the object is on the left so that the vertical straight contour is on the right. This effect is confirmed in Figure 2.10(b) and 2.10(c). Figure 2.10(b) shows a histogram of the average firing rate response of the neuron to the four subsets of trained objects that contain a straight contour at one of the sides: right, bottom, left and top (conventions as in Figure 2.6). The histogram confirms that the neuron has learned to respond to a vertical straight contour on the right of each of the trained objects. Figure 2.10(c) shows similar results for the mirror image objects. Here it can be seen that the neuron fails to respond to any of the mirror image objects, including those mirror image objects with a vertical straight contour on the left.

The neuron is thus selective for those members of the first set of trained objects that have a vertical straight contour on the right. The neuron does not respond to objects from the second mirror image set with a vertical straight contour on the left, even if these objects are shifted to ensure the vertical straight contours are overlapping between the two sets of objects. In other words, these neuronal responses seem to encode a 'border ownership' information of the vertical straight contour presented. If so, it may be said that the connections through successive layers to the output neuron learn about both the shape of the boundary contour element and the local image context such as the relative positions of other object features.

However, it is important to note that this result illustrates only a simple example of neural responses that seems to code the 'border ownership context. For example, the figures used in this simulation are always black. Therefore, in order to conclude whether the strict border ownership response selectivity is achieved through the learned feed-forward connectivity set up

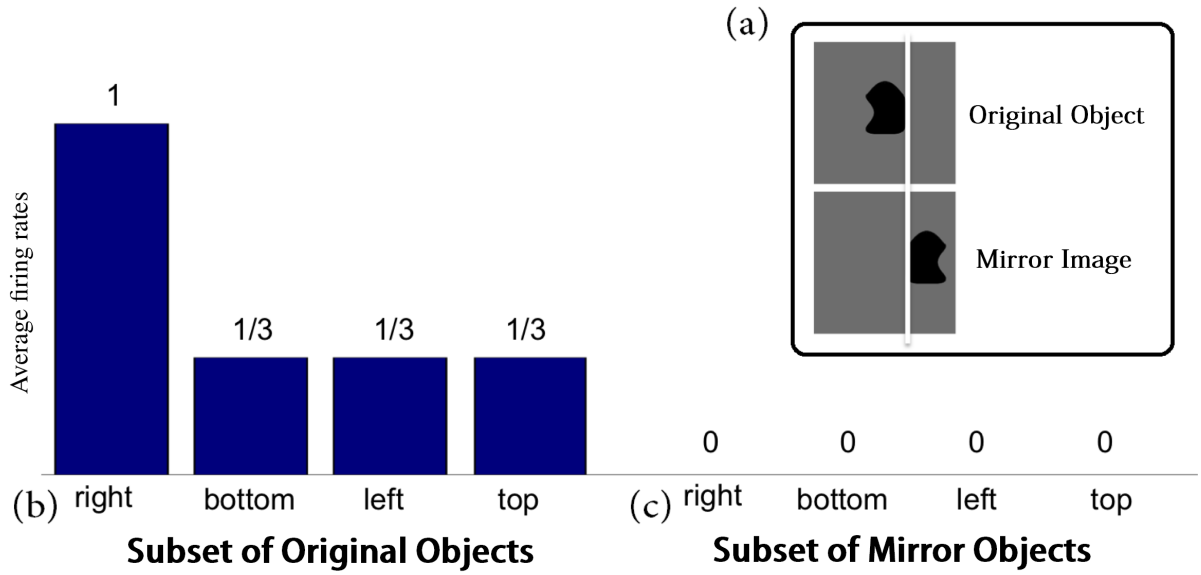

Figure 2.10: Demonstration of neuronal response tuning in an object centred frame of reference. The responses of the neuron analysed in Figure 8 were recorded as VisNet was tested with two sets of objects. The first set contained 4-sided objects from an original training set that had at least one straight contour element. The second set of objects were mirror images of the first set of objects. The mirror images were constructed by reflecting the original trained objects around the retinal location of the vertical straight contour on the right of the trained objects. An example of an object and its mirror image is shown in part (a). It can be seen that the vertical straight contours on the right and left of the two objects are aligned on the retina. Part (b) shows a histogram of the average firing rate response of the neuron to the four subsets of trained objects that contain a straight contour at one of the sides: right, bottom, left and top. The histogram confirms that the neuron has learned to respond to a vertical straight contour on the right of each of the trained objects. Part (c) shows similar results for the mirror image objects. Here it can be seen that the neuron fails to respond to any of the mirror image objects, including those mirror image objects with a vertical straight contour on the left. The neuron is thus selective for those members of the first set of trained objects that have a vertical straight contour on the right. The neuron does not respond to objects from the second mirror image set with a vertical straight contour on the left, even if these objects are shifted to ensure the vertical straight contours overlap across the two sets of objects.

during training, more detailed investigations are conducted in a later chapter in Chapter 5.

2.4.1.3 How the number of object sides (n) and the number of possible boundary elements at each side (p) affect the learned neuronal response properties

I investigated how the neuronal firing properties that develop in the network depend on the number of object sides (n) and the number of possible boundary contour elements (p) at each side. Each simulation was run with a fixed value of n and p. Across simulations, the number of sides, n, was varied from 3 to 8, while the number of possible boundary elements, p, was varied from 2 to 4. For each simulation, the network was trained on the full set of objects that could be constructed given the fixed values of n and p for that simulation. This means that the total number of distinct visual stimulus presented to the network during the simulations for each condition. Therefore, for practical reasons, simulations with $p^n > 1,000$ were omitted. Alternatively, the number of visual stimulus across different conditions could have been matched. Nevertheless, as well as practical reasons, because even with the smallest number of visual stimulus ($p^n = 8$), the network had settled as reported in Section 2.4.1.1, I did not choose the option. During training, the feedforward synaptic connections were modified using the Hebb learning rule (1.7) within VisNet implemented with competitive networks.

For each combination of n and p, Figure 2.11(top) gives the number of neurons that learned to respond selectively to all objects that contained one particular type of boundary contour element, but not to objects that did not contain that boundary element.

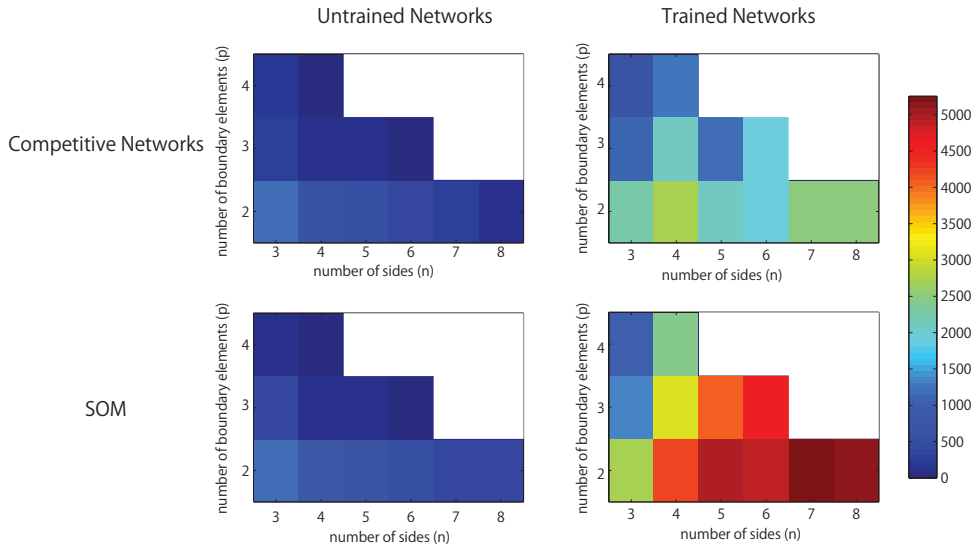

Figure 2.11: Results of simulations in which VisNet is trained and tested on objects constructed with a fixed number of sides (n) and number of possible boundary elements at each side, p. These simulations were performed with a competitive network architecture (top) or SOM (bottom) incorporated into each layer of VisNet. Due to the limited capacity of the VisNet programme, only the results for simulations where the total number of objects, p^n, is less than 1,000 are given. For each simulation, the table records the number of neurons (in the 3rd layer) that learned to respond selectively to all objects that contained one particular type of boundary contour element, but not to objects that did not contain that boundary element. Results are given before training (left) and after training (right). For all combinations of n and p, it can be seen that training VisNet on all possible p^n objects has lead to many neurons learning to respond selectively to objects containing a particular boundary contour element.

It was found that the last layer of the untrained network already contained a small number of cells that were selective for objects that contained one type of boundary element. This was because this simulation task was relatively easy in that it did not require the output neurons to respond invariantly as objects were translated across different retinal locations. In simulations reported later, the output neurons were tested with the objects presented in different retinal locations. In these simulations, training was indeed required to produce any neurons that responded selectively to objects containing one kind of boundary element.

In the trained network, it can be seen that all simulations produced large numbers of neurons that were selective for objects that contained one particular type of boundary element. Secondly, the number of object sides, n, did not have a significant systematic effect on the performance of the network. In contrast, as the number of possible boundary elements at each side, p, increased, the number of neurons that learned to respond selectively to objects containing one type of boundary element declined.

Here, it is important to understand the exact effects of varying n and p. While varying n should affects the size and displacement of the features, changing p should affect dimensionality and similarity simultaneously. Therefore, I hypothesise that the increased difficulty of neurons in the higher layers to develop separate representations of contour elements at each side is due to the effective increase in the density of the boundary contour elements, which results in more similar boundary conformations.

In particular, an invariance learning mechanism known as Continuous Transformation (CT) learning (Stringer et al., 2006) may cause neurons in higher layers to learn to respond to a number of similar boundary conformations at each side; CT learning is able to bind smoothly varying input patterns, such as a continuum of different possible boundary conformations at one of the object sides, onto the same postsynaptic neuron (see Section 1.4.1). In this way, CT learning may dramatically reduce the selectivity of neurons for particular boundary conformations.

Typical network behaviour for a relatively large value of p is shown in Figure 2.12. In

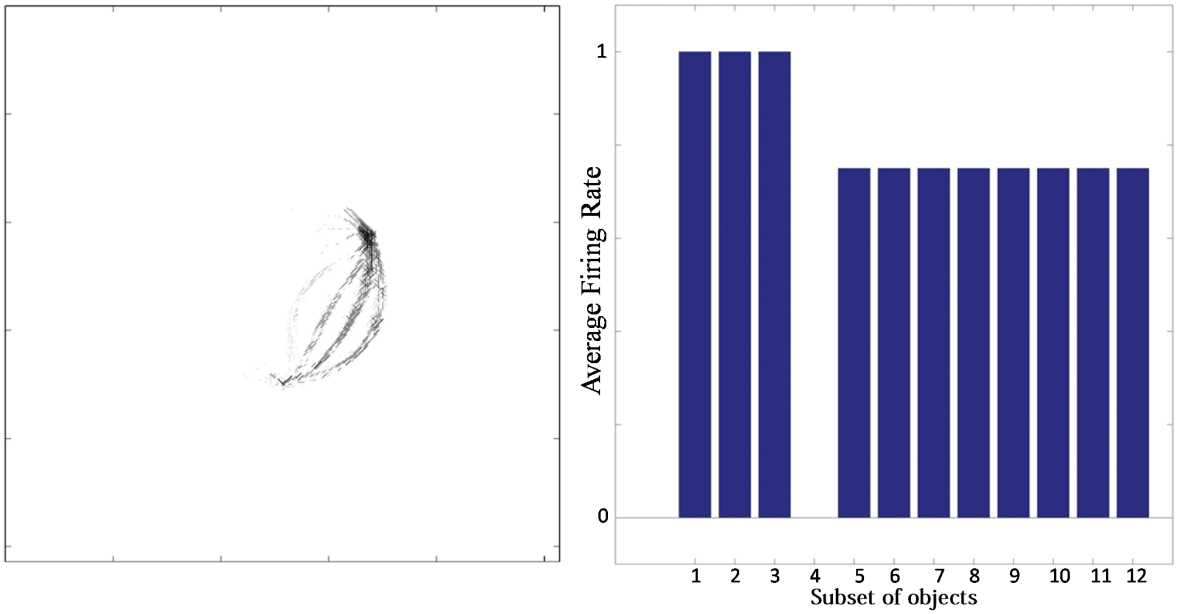

Figure 2.12: Simulation showing the failure of an output neuron to discriminate between a relatively large number of boundary contours at one object side. The simulation was performed with a competitive network architecture implemented within each layer of VisNet. The network is trained on objects with $n = 3$ sides, where each side has a relatively large number of $p = 4$ possible boundary elements. The figure shows results after training for a typical output neuron that failed to learn to respond selectively to one particular boundary element. Left: The input Gabor filters that had strong connectivity through the layers to the output neuron. It can be seen that the neuron has strong connections from three similar boundary elements on the lower right. Right: histogram showing average firing rate response of the neuron to the 12 subsets of objects that contain one of the different boundary elements. That is, each of the data points (1-12) represents the average firing rate of the neuron across the 16 objects containing the following boundary elements: (1) right/sharp-convex, (2) right/convex, (3) right/concave, (4) right/sharp-concave, (5) left/sharp-convex, (6) left/convex, (7) left/concave, (8) left/sharp-concave, (9) top/sharp-convex, (10) top/convex, (11) top/concave, (12) top/sharp-concave. It can be seen that the neuron responds maximally to the first three subsets of objects, which contain the three boundary elements that are strongly represented in the left plot. Thus, the neuron has learned to respond to these three boundary elements and is unable to distinguish between them. This effect is typical when the number (density) of boundary contour elements at each side is increased.

this example, the network was trained on objects with $n = 3$ sides, each of which had $p = 4$ possible boundary elements. The figure shows results for a typical output cell that failed to learn to respond selectively to objects containing one particular type of boundary contour. Figure 2.12(left) shows the input Gabor filters that had strong connectivity through the layers to the neuron. The neuron has strong connections from three similar boundary elements on the lower right. Figure 2.12(right) shows the average firing rate response of the neuron to the 12 subsets of objects that contain one of the different boundary elements. The neuron responds maximally to the first three subsets of objects, which contain the three boundary elements that are strongly represented in the left plot. Thus, the neuron has learned to respond equally strongly to all of these three boundary elements and is unable to distinguish between them. This observed behaviour is typical when the number (density) of boundary contour elements at each side is increased. Investigation into the responses of neurons across the output layer after the training the network on objects where each side had a relatively high number of possible boundary element contours, p, showed that many cells were unable to distinguish between differently shaped contours on the same sides.

The simulations at this juncture show that a biologically plausible neural network can learn to code relative position information for visual elements, but has limited capacity. In the next section, I show how introducing a Self-Organising Map (SOM) architecture within each layer of VisNet can enhance the selectivity of neurons for individual boundary elements when the number of boundary elements at each side, p, is large, overcoming the capacity limitation.

2.4.1.4 The effect of a Self-Organising Map (SOM) architecture on learned neural selectivity for boundary contour elements

I compared the performance of the standard competitive network architecture in each layer with performance when a Self-Organising Map (SOM) was introduced. It hypothesised that the SOM architecture could increase the capacity of the network to represent and distinguish between a larger number of finer variations in boundary contour curvature.

As discussed in the previous section, a competitive network may have difficulty in forming separate output representations of similar input patterns. In particular, Continuous Transformation learning introduced in Section 1.4.2 (Stringer et al., 2006) may encourage the same output neurons to learn to respond to similar input patterns representing boundary contour elements of slightly different shape, or even bind together a continuum of input patterns covering the space of all possible boundary shapes at a particular object-centred boundary location.

The SOM architecture is specifically designed to encourage the output neurons to develop a fine-scaled representation of a continuum of smoothly varying input patterns (Kohonen, 1982). A SOM has additional short range lateral excitatory connections between neurons within each layer. These connections encourage nearby output neurons to learn to respond to similar input patterns, which in turn leads to a map-like arrangement of neuronal response characteristics across the layer after training. In particular, slightly different input patterns will be distributed across different output neurons. Thus, the effect of these additional short range excitatory connections is to influence learning in the network to spread the representations of a continuum of overlapping input patterns over a map of output neurons. This should allow the network to develop a more fine-grained representation of the space of possible boundary contour shapes.

I therefore hypothesised that the introduction of a SOM architecture within each layer of VisNet would spread out the representations of many different boundary contour curvatures (p) at a particular side of the object over a map of output neurons. This would help to produce distinct neural representations of a large number of different boundary contour elements in the output layer, and effectively increase the capacity of the network to represent finer variations in boundary contour curvature.

During training, the feedforward synaptic connections were, again modified using the Hebb learning rule (1.7), and the simulation results with the Self-Organising Map (SOM) architecture implemented within each layer are presented in Figure 2.11(bottom). The network was tested on objects constructed with a fixed number of sides, n, and different numbers of possible boundary elements at each side, p. For each simulation, the heatmap shows the number of neurons that learned to respond selectively to all objects that contained one particular type of boundary contour element, but not to other objects. These results should be compared with Figure 2.11(top), which gives the corresponding results with a competitive network architecture implemented within each layer. As hypothesised, the introduction of the short range excitatory connections of the SOM architecture within each layer led to many more neurons learning to respond selectively to objects containing a particular boundary contour element. This effect is particularly pronounced for larger numbers of n and p.

These effects can also be seen by examining the amount of information carried by neurons about the presence of particular types of boundary elements within the objects presented to VisNet. We have previously used information theoretic measures to assess the amount of information carried by neurons about the presence of whole object stimuli within a scene, where the objects may be presented under different transforms such as changes in retinal position or orientation (Wallis and Rolls, 1997; Rolls and Milward, 2000; Stringer et al., 2007; Stringer and Rolls, 2008). A neuron that responds selectively to one particular stimulus across a large number of transforms will carry a high level of information about the presence of that object within a scene. In this chapter, I was instead interested in the amount of information carried by neurons about the presence of particular boundary elements within an object.

Two information measures, single-cell information analysis and multiple-cell information

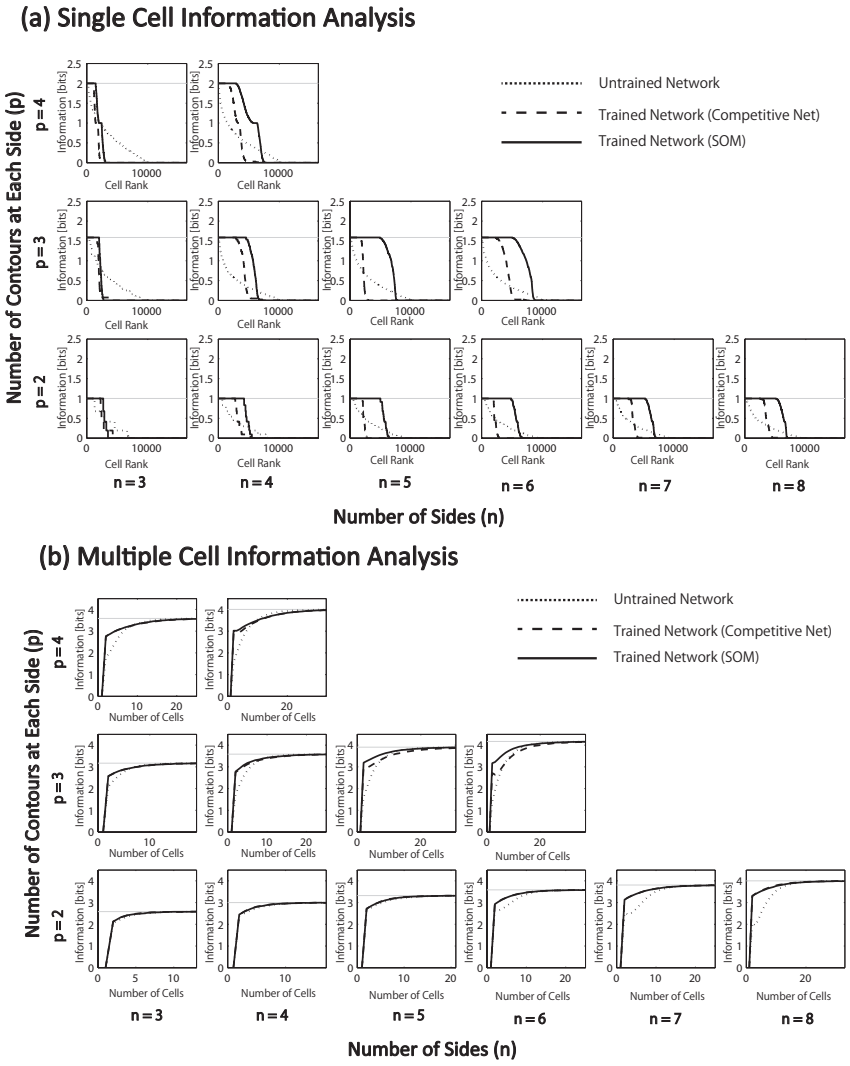

Figure 2.13: (a) Single cell information analysis results and (b) multiple cell information analysis results for simulations in which VisNet is trained and tested on objects with different numbers of sides, n, and numbers of possible boundary elements at each side, p. For each simulation, the single cell information measures for all output (3rd) layer neurons are plotted in rank order according to how much information they carry, and the multiple cell information measures are plotted according to the number of cells used to construct the probability table for the analysis. For each simulation, results are presented before training (dotted line) and after training with competitive network (broken dashed line) and with SOM (solid line). In all simulations, training the network on the set of p^n whole objects led to many top layer neurons attaining the maximal level of single cell information of $log_2(p)$ bits. However, consistent with our hypothesis, there was a clear trend that the SOM architecture increased the number of neurons that attained the maximal level of single cell information. Furthermore, for all simulations, the multiple cell information measures asymptoted to maximal values of $log_2(n \times p)$ bits, which indicated that all of the boundary contour elements were represented by a subset of different neurons.

analysis, were used to assess the ability of the network to develop neurons that are able to respond selectively to the presence of individual boundary contour stimuli. Single-cell information analysis indexes, for each cell, how much information is available about which boundary element is present at a particular side location within an object. On the other hand, multiple-cell information analysis indexes how many cells are needed in order to perfectly discriminate the entire set of input stimuli based on the ensemble of those cells' activations. The detailed procedures are provided in Section 2.3.2.

Figure 2.13 present (a) the single cell information analysis results and (b) the multiple cell information analysis results for simulations in which VisNet was tested on objects with different numbers of sides, n, and numbers of possible boundary elements at each side, p. For

each simulation, results are presented before training (dotted line), after training with the competitive network architecture (broken dashed line) and with the SOM architecture (solid line).

The single cell information measures for all output layer neurons are plotted in rank order according to how much information they carry. In all simulations, training the network on the set of p^n whole objects led to many top layer neurons attaining the maximal level of single cell information of $log_2(p)$ bits. These results imply that training the network on the whole objects led to many output neurons learning to respond selectively to all of the objects that contained a particular one of the boundary contour elements, but not to objects that do not contain that boundary element. That is, these neurons had learned to respond to the presence of that particular boundary contour element within any object. In all simulations, many top layer neurons attained the maximal level of single cell information of $log_2(p)$ bits. However, consistent with our hypothesis, the incorporation of a SOM architecture typically led to a significant increase in the number of neurons that attained the maximal level of single cell information.

Furthermore, for all simulations, the multiple cell information measure asymptoted to the maximal possible value of $log_2(n \times p)$ bits, demonstrating that the whole set of boundary contour elements was successfully represented by a set of different neurons. In particular, for simulations with relatively large n or p, in which it is less likely for untrained neurons to respond selectively to a particular boundary contour shape across all objects by chance, there is a noticeable improvement in the multiple cell information computed with just a few (e.g. 5) neurons after training. These few neurons that convey high levels of multiple cell information between them are those neurons that have learned to respond selectively to individual boundary contour elements across all objects.

For multiple cell information plots, there was little difference between the SOM and the competitive network. This was because both architectures produced a large enough number of neurons with maximal single cell information to ensure that only a small subset of these neurons was needed to give high levels of multiple cell information in either case. However, for relatively large values of n and p, such as $p = 3$ and $n = 6$, the SOM architecture gave rise to larger multiple cell information measures with fewer neurons. This provided further evidence that when the network needs to produce a more fine-grained representation of larger numbers of similar boundary contours, the SOM is able to outperform the competitive network architecture.

Furthermore, different sub-populations of cells that carry maximum single-cell information about each contour element were mapped onto the corresponding locations within the layer. This extended analysis has revealed that using a SOM led to a feature map as shown in Figure 2.14.

The lateral excitation utilized in the SOM architecture was initially used to produce self-organising systems that mimicked mappings found in early visual areas (Kohonen, 1982). In addition, this result shows that the architecture also results in developing the clustering of neural selectivity across visual stimuli. This result was consistent with various physiological findings that indicate the topographic organization within ventral visual pathway (Larsson and Heeger, 2006; Hansen et al., 2007; Silver and Kastner, 2009). This type of clustering of neural selectivity across faces throughout the temporal lobe is called 'face patch' (Wang et al., 1998; Tsao et al., 2006), which is simulated in the later in Chapter 3.

2.4.1.5 Response properties of neurons through successive layers of VisNet

I subsequently investigated how the response properties of neurons vary through successive layers of VisNet, which is implemented with either the competitive network or Self-Organising Map (SOM), before and after training. For all of the simulations performed, the feedforward synaptic connections were modified using the Hebb learning rule (1.7).

(a) Information map (n=3 and p=2) (b) Information map (n=3 and p=3)

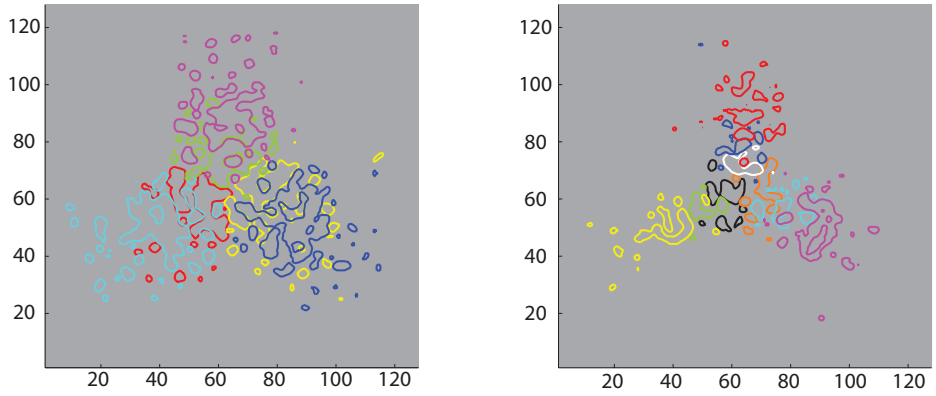

Figure 2.14: Simulation results demonstrating that the SOM architecture leads to a feature map in the output layer. (a) Left: contour plots showing the amount of single cell information carried by the 128 x 128 layer of output neurons for six boundary elements after training on objects with n=3 and p=2. The different coloured contour plots correspond to the following boundary elements: top/convex (pink), top/concave (light green), right/convex (blue), right/concave (yellow), left/convex (light blue), and left/concave (red). (b) Right: similar results for the case n=3 and p=3. For both simulations, it can be seen that the output neurons have self-organised into feature maps, in which nearby neurons in the layer tend to be tuned to the same kind of boundary element.

Figure 2.15 presents simulation results comparing the performance of the competitive network (left) and SOM architectures (right). In these simulations, VisNet was presented with objects with $n = 4$ sides, where each side has $p = 4$ possible boundary elements with varying degrees of curvature. The table presents the results for layers 1, 2 and 3. Each cell within the table gives the total number of neurons that learned to respond selectively to objects that contained a particular boundary element defined by side position and curvature. In all three layers, the SOM architecture outperformed the competitive network architecture in terms of producing more neurons that had learned to respond selectively to individual boundary elements.

The results also illustrated that the proportion of neurons representing more highly curved boundary elements increases in the higher layers. This is another physiological property replicated in our simulations (Kayaert et al., 2005). For both the competitive network and SOM architectures, the proportion of neurons representing more highly curved boundary elements increased through successive layers of VisNet as shown in Figure 2.15. Contour specific cells start to appear in layer 1, which itself is consistent with the experimental observation that curvature processing begins in visual area V2, an early stage of the ventral visual pathway (Ito and Komatsu, 2004). There was only a slight bias towards contour elements of higher curvature in layer 1. In layer 2, there is a much stronger bias towards the representation of boundary elements with higher curvature. This bias reached a maximum in layer 3, where the vast majority of cells are tuned to higher curvature contour elements.

Table 2.7 presents simulation results showing the responses of neurons through layers 1 to 3 with the Self-Organising Map (SOM) architecture. The results are presented for two different simulations: objects with $n = 4$ sides and $p = 2$ contour elements per side; and objects with $n = 5$ sides and $p = 3$ contour elements per side. For each simulation, results before and after training are compared. Each sub-table gives the number of neurons that responded selectively to either objects containing a single boundary element, objects containing a combination of two boundary elements, or a single whole object. For both simulations, it can be seen that, in all three layers, training the network led to a substantial increase in the number of neurons that responded to objects containing a single boundary element. The numbers of neurons that learned to respond to individual boundary elements increased through successive layers of VisNet.

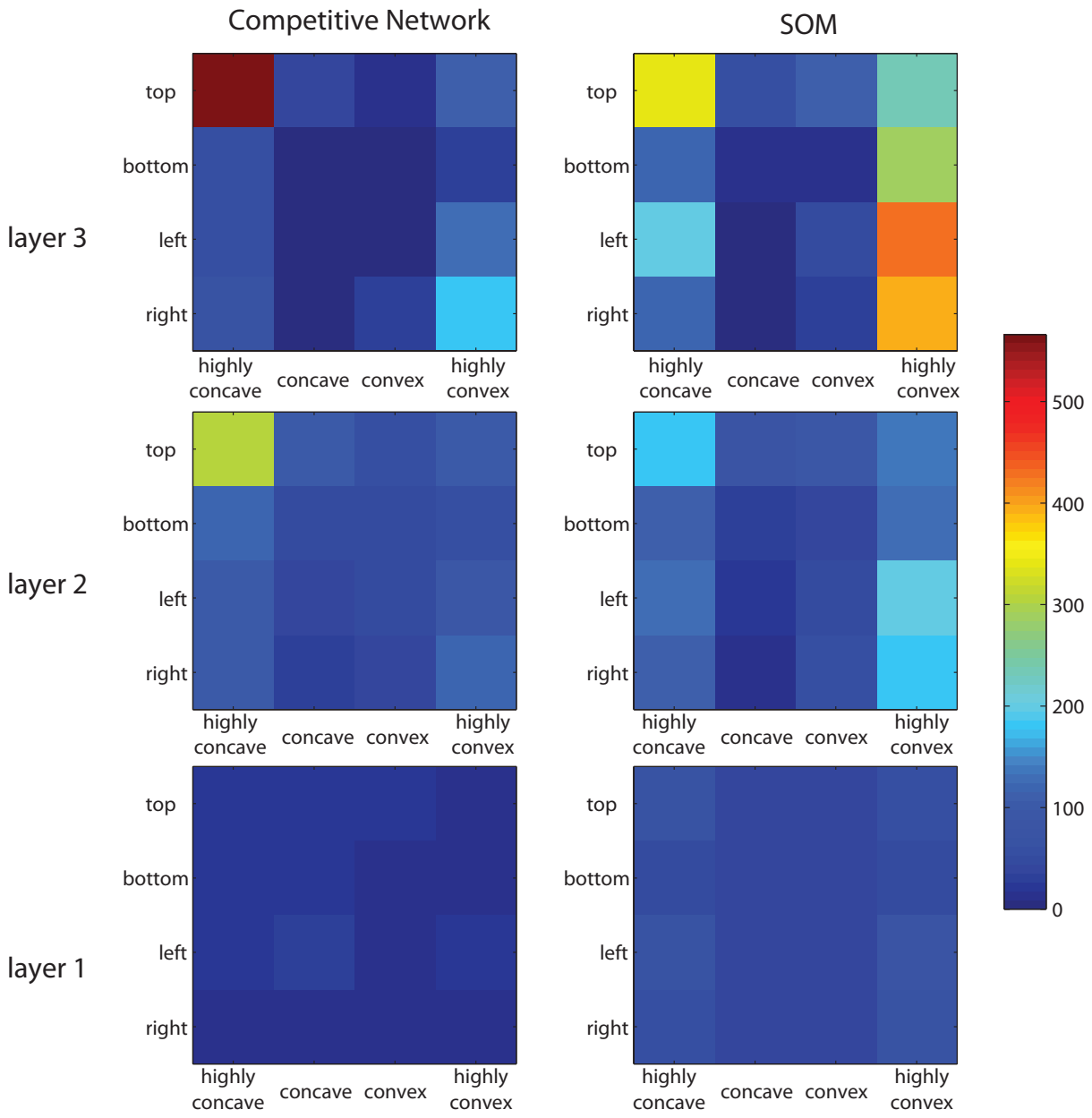

Figure 2.15: Simulation results comparing the performance of the competitive network architecture (left) and the Self-Organising Map (SOM) architecture (right) through successive layers of the network. In these simulations, VisNet is trained and tested on objects with n=4 sides, where each side has p=4 possible boundary curvatures: highly concave, moderately concave, moderately convex and highly convex. Each of the two subplots shows the number of neurons that learned to respond selectively to objects containing each kind of contour curvature regardless of side position. It can be seen that, for both the competitive network and SOM architectures, the proportion of neurons representing more highly curved boundary elements increases in the higher layers of VisNet.

Table 2.7: Simulation results showing the responses of neurons through layers 1 to 3 with the Self-Organising Map (SOM) architecture.

N4P2 experiment (SOM)

Untrained Network			
Layer	1 contour	2 contours	object
3	856	270	538
2	418	104	82
1	293	0	0

Trained Network			
Layer	1 contour	2 contours	object
3	4216	92	89
2	2440	10	10
1	540	8	0

N5P3 experiment (SOM)

Untrained Network			
Layer	1 contour	2 contours	object
3	131	9	143
2	487	8	10
1	287	0	0

Trained Network			
Layer	1 contour	2 contours	object
3	4028	1	34
2	2507	0	17
1	754	0	0

On the other hand, the training did not lead to a similarly large increase in the number of neurons that responded to a particular combination of two boundary elements and in the number of neurons that responded to a particular whole object (Table 2.7).

In terms of physiology, (Brincat and Connor, 2004) reported that neurons in the later stages of the ventral visual pathway, TEO, integrate information from multiple boundary contour elements to represent increasingly complex object shape. This means that in order for the whole shape representations to develop in the simulation, it should require a large number of representation of local contour elements in earlier stage of processing.

The question is why the numbers of cells that seemingly exhibit selectivity to a particular combination of two boundary elements and whole shapes are larger before training than those after training (Table 2.7). In order to investigate how exactly these neurons before training might have exhibited such response properties, the input Gabor filters that had a strong connectivity to an example cell that exhibits selectivity to a particular combination of two contour elements and to a single whole object are examined (Figure 2.16).

It turned out that most of the untrained cells that seemingly exhibit selectivity to a particular combination of multiple local contour elements in fact neither encoded any shape nor performed integrations of the local shapes, but rather simply responded based on more abstract location of the stimulation on the input layer. On the other hand, many trained cells actually learned to respond to a specific combinations of particular local contour shapes.

Figure 2.16 compares the response properties of trained and untrained neurons in simulations with the Self-Organising Map (SOM) architecture. The network is presented with objects containing $n = 4$ sides, where each side has $p = 2$ possible boundary elements. Results are shown for four neurons. For each neuron, the input Gabor filters that had strong connectivity through the layers to the neuron (left), and a histogram showing average firing rate response of the neuron to the objects that contain one of the 8 boundary elements (right) are shown. The four neurons shown in the figure had the following characteristics. (a) A trained neuron that has learned to respond to a combination of two adjacent boundary contour elements: top convex and right convex. The Gabor filter plot shows that the feed-forward synaptic weights have been strengthened selectively from the two boundary elements only. (b) A trained neuron that has learned to respond to a whole object. The preferred object is comprised of two concave on top and right and two convex on bottom and left. The Gabor filter plot shows that the neuron has learned to respond to the complete set of boundary elements comprising the preferred object. (c) An untrained neuron that happens to respond selectively during testing to two adjacent boundary elements. However, the Gabor filter plot shows that a random collection of Gabor filters have strong feed-forward connections to the neuron. This means that across a richer diversity of test images, this neuron would not maintain such a strict selectivity, and would in fact be most effectively stimulated by the random constellation of Gabor filters shown. (d) An untrained neuron that responds selectively to a whole object. The Gabor filter plot shows that the neuron receives strong connections from a random collection of Gabor filters. This

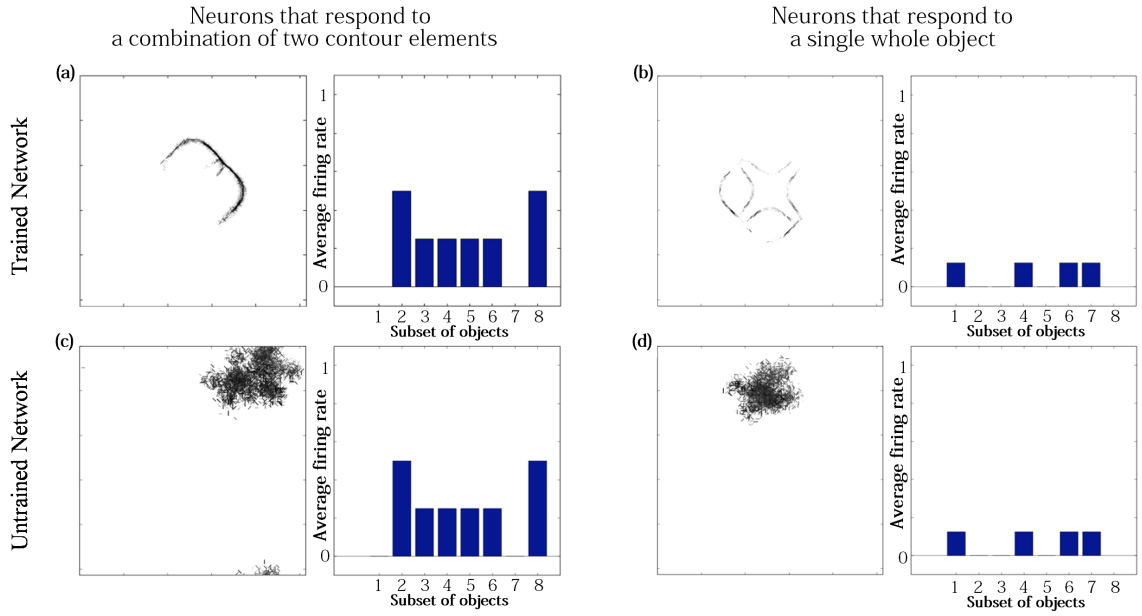

Figure 2.16: Comparison of response properties of trained and untrained neurons in simulations with the Self-Organising Map (SOM) architecture. The network is presented with objects containing $n = 4$ sides, where each side has $p = 2$ possible boundary elements. The figure shows results for four typical neurons. For each neuron, two plots are shown. Left: The input Gabor filters that had strong connectivity through the layers to the neuron. Right: histogram showing average firing rate response of the neuron to the 8 subsets of objects that contain one of the different boundary elements. That is, each of the data points (1-8) represents the average firing rate of the neuron across the 8 objects containing the following boundary elements: (1) right/concave, (2) right/convex, (3) bottom/concave, (4) bottom/convex, (5) left/concave, (6) left/convex, (7) top/concave, (8) top/convex. The four neurons have the following response properties: (a) a trained neuron that learned to respond to a combination of two adjacent boundary contour elements, (b) a trained neuron that learned to respond to a whole object, (c) an untrained neuron that responds selectively to two adjacent boundary elements, and (d) an untrained neuron that responds selectively to a whole four-sided object.

neuron would not maintain a strict selectivity to the object when tested on a greater diversity of images.

The conclusion of the results shown in Figure 2.16 is that although Table 2.7 appeared not to show an increase during training in the numbers of neurons that responded to combinations of two boundary elements or a whole object, in fact training did lead to an increase in the numbers of neurons that had specifically learned to respond to whole stimuli. However, in Table 2.7, this effect had been masked by the existence of many untrained cells that already responded by chance to combinations of two boundary elements or a whole object, but which in fact had random inputs from a large randomised collection of Gabor filters. Such untrained neurons are unlikely to be selective for combinations of two boundary elements or a particular object if the network were tested on a richer diversity of images. In particular, these untrained neurons would respond more selectively for images corresponding to the random constellations of Gabor filters shown in Figure 2.16 for cells (c) and (d). In contrast, the trained neurons (a) and (b) have strengthened connections specifically from combinations of two boundary elements or a whole object, and would therefore maintain their selectivity more robustly across a greater variety of test images.

This juncture was tested in more detail subsequently. I also found that output neurons in layer 3 learned to respond to whole objects by combining inputs from neurons in the preceding layer that responded to the individual boundary elements. This can be seen by examining the strengths of the synaptic connections from neurons in layer 2 to output neurons in layer 3 after training. Output neurons that had learned to respond to a particular object received the strongest synaptic connections from neurons in layer 2 that represented the constituent boundary elements of that object. Figure 2.17(a) shows an output neuron that learned to

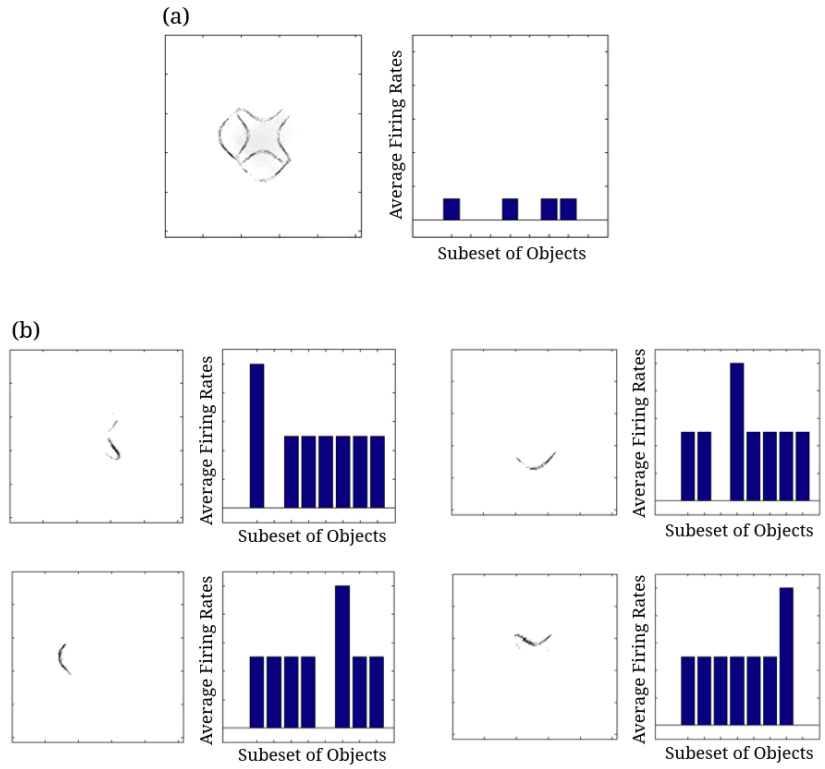

Figure 2.17: Simulation results showing how output neurons may learn to represent whole objects by combining inputs from multiple neurons in the preceding layer that represent individual boundary contours. (a) An example output neuron in the third layer that learned to respond to a whole object containing concave contours at the top and right sides and convex contours at the bottom and left sides. The left plot shows the input Gabor filters that had strong connectivity through the layers to the output neuron. The right plot presents a histogram of the average firing rate response of the output neuron to the 8 subsets of objects that contain one of the different boundary elements. That is, each of the data points (1-8) represents the average firing rate of the neuron across the 8 objects containing the following boundary elements: (1) right/concave, (2) right/convex, (3) bottom/concave, (4) bottom/convex, (5) left/concave, (6) left/convex, (7) top/concave, (8) top/convex. Although the Gabor filter plot also shows concave contours at the bottom and left sides, the average firing rate responses indicate that the neuron did not respond to these two contours. (b) Four neurons in layer 2 with strong synaptic connections to the output neuron shown in (a). For each of these four neurons in layer 2, the input Gabor filters that had strong connectivity through the layers to that neuron (left) and a histogram showing average firing rate response of the neuron to the 8 subsets of objects that contain one of the different boundary elements are shown. It can be seen that each of the four neurons in layer 2 had learned to respond to a different one of the boundary elements which were contained in the object that the output neuron had learned to respond to.

respond to a whole object containing concave contours at the top and right sides and convex contours at the bottom and left sides. Figure 2.17(b) shows four neurons in layer 2 with strong synaptic connections to the output neuron shown in part (a). Each of the four neurons in layer 2 had learned to respond to a different one of the boundary elements which were contained in the object that the output neuron had learned to respond to. This example shows that neurons in the later stages of the model are able to integrate information from multiple boundary contour elements, as consistent with neurophysiological results for area TEO of the primate ventral visual pathway (Brincat and Connor, 2004).

2.4.1.6 Translation invariance of neuronal responses as objects are shifted across different locations on the retina

The neurons reported by Pasupathy and Connor (2001) in area V4, and neurons reported by Brincat and Connor (2004) in area TEO, respond with translation invariance as an object is shifted across different retinal locations. In this section I show how these translation invariant neuronal responses may be set up by training the network with the trace learning rule (1.8) and

(1.9).

The trace learning rule encourages individual postsynaptic neurons to learn to respond to subsets of input patterns that tend to occur close together in time. Therefore, in the simulation described below, during training each object is selected in turn and presented in a number of different retinal locations before moving on to the next object. Thus, for each object, the presentations at different retinal locations were clustered together in time. In this case, the trace learning rule will encourage output neurons to learn to respond to a particular object or boundary contour with translation invariance on the retina.

For this simulation, VisNet had 4 layers with a SOM architecture implemented within each layer. The visual objects had $n = 4$ sides, where each side has $p = 3$ possible boundary elements. Each of the visual objects was presented in a 2×2 grid of 4 different retinal locations, which were separated by horizontal and vertical shifts of 10 pixels. This means that many stimuli contains overlapping region with a differently shaped stimulus at a different retinal location, which makes the problem more difficult.

Although it is not shown in this report, prior to the current simulation study with trace learning rule implemented, the network was trained with the same set of stimuli with ordinal Hebbian learning rule. The result was that none of the output cells learned to exhibit translation invariant responses. With the trace learning rule implemented in the current simulation study, we expected that the network can combine the retinal location specific representations of the shapes reported in earlier sections to learn to develop the translation invariant representations.

Figure 2.18 shows the results after training for a typical output neuron in layer 4. Figure 2.18(top) shows the input Gabor filters that had strong connectivity through the layers to the output neuron. It can be seen that the neuron has strong connections from a convex boundary element on the left of an object. The separate contours that can be seen in the plot correspond to the different retinal locations in which the objects are trained. Figure 2.18(bottom) shows a histogram presenting the average firing rate response of the output neuron to the 12 subsets of objects that contain one of the boundary contour elements. The neuron responds maximally to the subset of objects containing a convex boundary element on the left. Notably, the neuron responds maximally to this subset of objects over all 4 retinal locations. Thus, the neuron has learned to respond to objects containing the convex boundary element on the left regardless of where the object is presented on the retina. These translation invariant neuronal responses are a result of training the network with the trace learning rule.

Fig 2.19 shows findings from the single cell information analysis (top) and the multiple cell information analysis (bottom). The results are presented before training (broken line) and after training (solid line). Training the network on the set of p^n whole objects over the 4 retinal locations led to many top layer neurons attaining the maximal level of single cell information of $log_2(p)$ bits. Neurons carrying maximal single cell information responded selectively to a subset of objects containing one particular type of boundary element, and with translation invariance as the objects also were shifted over all 4 retinal locations. Furthermore, after training, the multiple cell information measures asymptoted to the maximal level of $log_2(n \times p)$ bits. This indicates that all of the different boundary contour elements are represented with translation invariance by different output neurons after training. In these simulations with translation invariance, the multiple cell information is dramatically increased after training. This is because it is very unlikely for untrained neurons to both respond selectively to a single boundary contour element across all objects, and be able to respond with translation invariance as these objects are shifted across the retina. Therefore, training will lead to a much more significant difference between the performances of the untrained and trained networks.

Figure 2.20 shows the numbers of neurons that responded selectively to all objects that contained one particular type of boundary element, and with translation invariance as the objects shifted over all 4 retinal locations. The results are presented before training (left) and after training (right). Before training, there were no neurons in any of the layers that

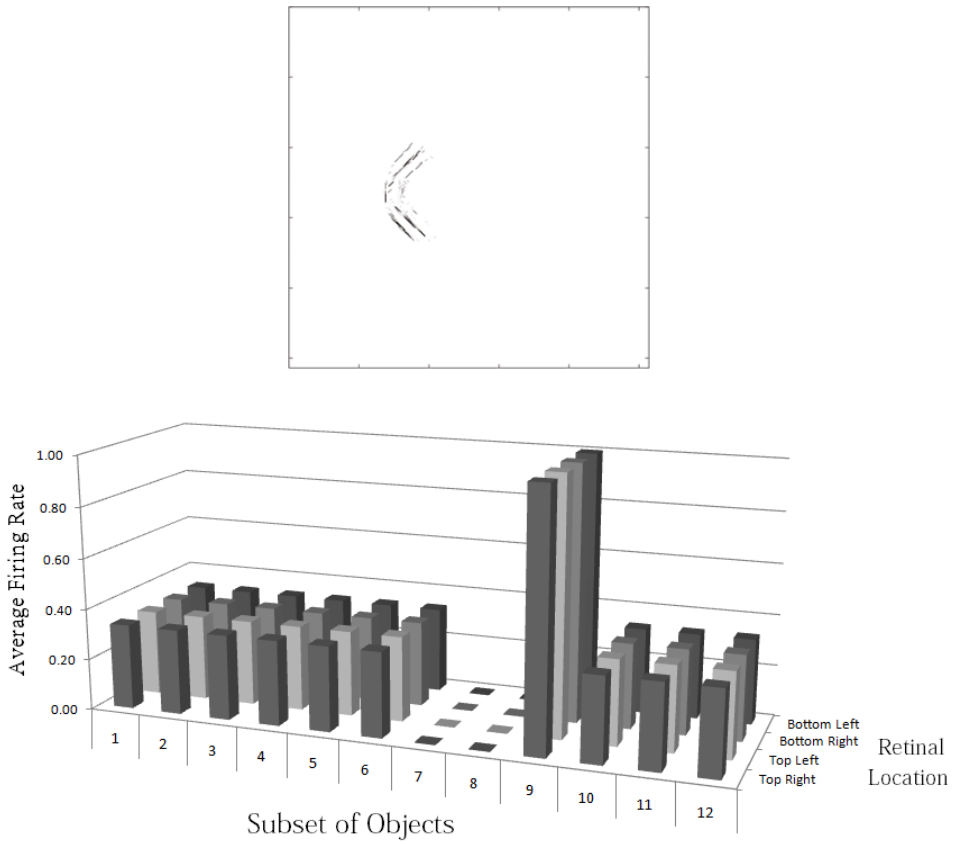

Figure 2.18: Simulation of network trained with the trace learning rule (1.8) and (1.9) as each of the visual objects is shifted across 4 different retinal locations: top right, top left, bottom right and bottom left. The objects had $n = 4$ sides, where each side has $p = 3$ possible boundary elements. The simulation was carried out with a competitive network architecture implemented within each layer of VisNet. The figure shows results after training for a typical output neuron in layer 4. Top: The input Gabor filters that had strong connectivity through the layers to the output neuron. It can be seen that the neuron has strong connections from a convex boundary element on the left of an object, which may be presented at four different retinal locations. Bottom: histogram showing the average firing rate response of the output neuron to the 12 subsets of objects that contain one of the boundary contour elements. That is, each of the data points (1-12) represents the average firing rate of the neuron across the 27 objects containing the following boundary elements: (1) right/concave, (2) right/straight, (3) right/convex, (4) bottom/concave, (5) bottom/straight, (6) bottom/convex, (7) left/concave, (8) left/straight, (9) left/convex, (10) top/concave, (11) top/straight, (12) top/convex. Each of these results is given for the objects placed in the 4 different retinal locations. It can be seen that the neuron responds maximally to the ninth subset of objects, which contains a convex boundary element on the left of an object as shown in the top subplot. In particular, the neuron responds maximally to this subset of objects over all 4 retinal locations.

responded selectively to a particular boundary contour with translation invariance. However, after training, there were significant numbers of neurons that learned to respond selectively to a particular boundary contour element over all 4 retinal locations. It can be seen that all of the different types of boundary contour are well represented by different subsets of output neurons. Interestingly, the numbers of such perfectly discriminating and translation invariant cells gradually increased through successive layers from 2 to 4. This observation reflects the increase in the size of receptive fields of neurons through successive layers of VisNet, which permits neurons in higher layers to respond to objects over a larger region of the retina.

2.4.1.7 Sensitivity of neuronal responses to object rotation

The neurophysiological study of Pasupathy and Connor (2001) showed that the responses of neurons were sensitive to the position of a local boundary contour element with respect to the centre of mass of a visual object. For example, a neuron might encode a convex contour at the top right (e.g. 45 degrees) of an object. This implies that the responses of such a neuron

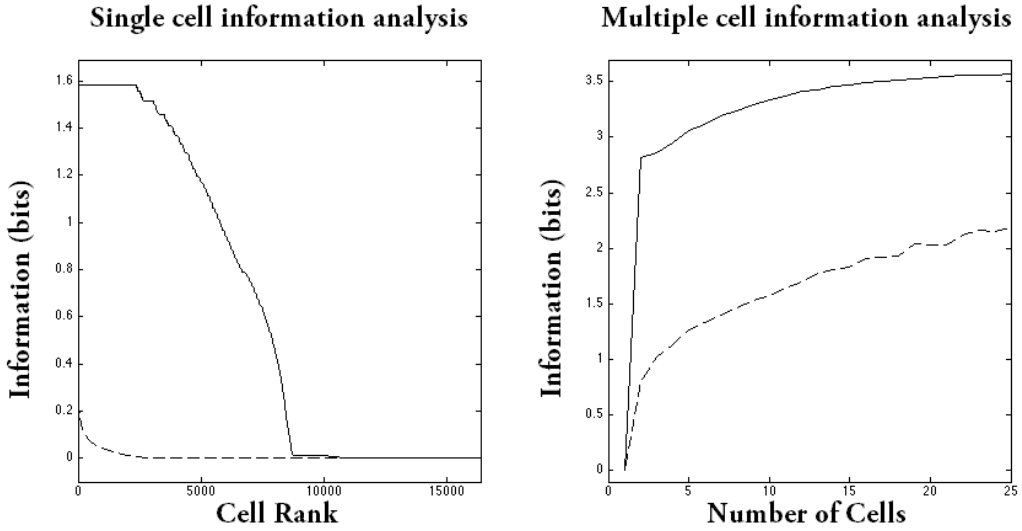

Figure 2.19: Simulation of network trained with the trace learning rule (1.8) and (1.9) as each of the visual objects is shifted across 4 different retinal locations. The objects had $n = 4$ sides, where each side has $p = 3$ possible boundary elements. The simulation was carried out with a competitive network architecture implemented within each layer of VisNet. The figure shows single cell information analysis results (top) and multiple cell information analysis results (bottom). The single cell information measures for all output layer neurons are plotted in rank order according to how much information they carry, and the multiple cell information measures are plotted according to the number of cells used in the analysis. Results are presented before training (broken line) and after training (solid line). Training the network on the set of p^n whole objects over the 4 retinal locations led to many top layer neurons attaining the maximal level of single cell information of $log_2(p)$ bits. Furthermore, after training, the multiple cell information measures asymptoted to the maximal level of $log_2(n \times p)$ bits.

will be sensitive to the rotation of an object that contains such a boundary contour element. I tested whether our model also demonstrated sensitivity to object rotation by running additional simulations in which the visual objects underwent in-plane rotation during training and testing.

The visual objects had $n = 3$ sides and each side had $p = 2$ possible boundary elements. Each object was rotated through 120 degrees in steps of 2 degrees during training and testing. This ensured that each boundary element shape (concave or convex) was sampled in every rotational position (albeit in increments of 2 degrees). The VisNet architecture used in this simulation consisted of three Self-Organising Map (SOM) layers. During training, the feedforward synaptic connections were modified using the Hebb learning rule (1.7). The Hebb learning rule was used in preference to the trace learning rule (1.8), (1.9) because it can be assume that in natural environment, objects are not likely to keep rotating so often.

The simulation results are shown in Figure 2.21, which illustrates the firing characteristics of a typical output neuron after training. The Gabor filter plot (b) showed that the neuron had learned to respond to a local convex contour at about 75 degrees. This was confirmed by the average firing rate response of the neuron to the 6 subsets of objects containing the 6 different boundary contour elements shown in plots (c)-(h). The response of the neuron was thus sensitive to the rotational position of the contour within an object. Specifically, the neuron fired maximally when there was a local convex contour at about 75 degrees. This result is consistent with the physiological findings reported by Pasupathy and Connor (2001).

2.4.2 Study 2: VisNet simulations with visual stimuli of Pasupathy and Connor

In Study 2, the visual stimuli presented to VisNet were similar to the artificial stimuli used in the neurophysiological experiments of Pasupathy and Connor (2001) and are shown in Figure 2.4. This allowed direct comparison between the learned response characteristics of the neurons

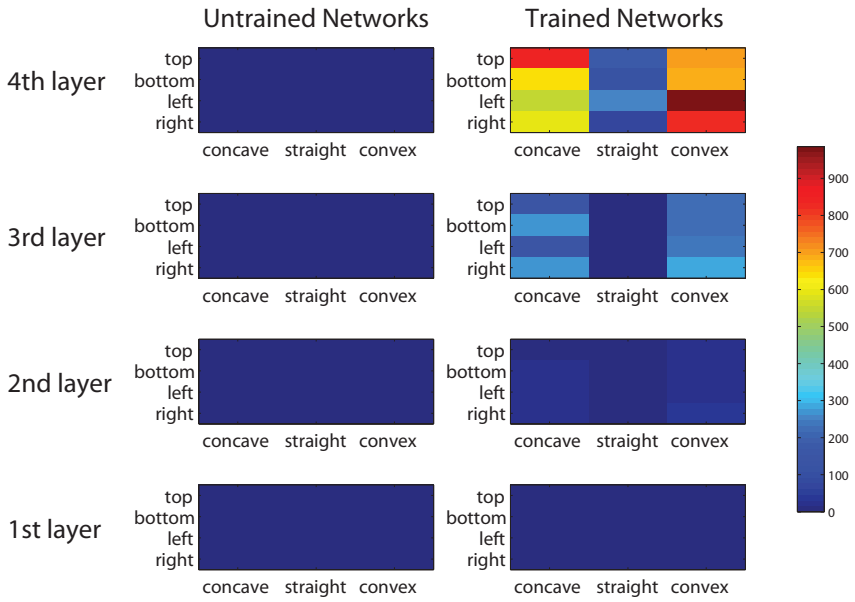

Figure 2.20: Simulation of network trained with the trace learning rule (1.8), (1.9) as each of the visual objects is shifted across 4 different retinal locations. The objects had $n = 4$ sides: right, bottom, left and top. Each side has $p = 3$ possible boundary elements: concave, straight and convex. The simulation was carried out with a competitive network architecture implemented within each layer of VisNet. Results are presented before training (left) and after training (right). Each heatmap shows the numbers of neurons in each of the layers 1 to 4 that respond selectively to all objects that contained one particular type of boundary element, and that respond with translation invariance as the objects are shifted over all 4 retinal locations. Before training, there are no neurons in any of the layers that respond selectively to a particular boundary contour with translation invariance. However, after training, there are significant numbers of neurons that have learned to respond selectively to a particular boundary contour element over all 4 retinal locations. It can be seen that all of the different types of boundary contour are well represented by different subsets of output neurons. The number of such perfectly discriminating and translation invariant cells gradually increases through successive layers from 2 to 4.

in the VisNet model and the experimentally observed cell responses encoding local boundary information reported by Pasupathy and Connor (2001).

The stimuli were constructed by systematically combining sharp convex, medium convex, broad convex, medium concave, and broad concave boundary elements to form closed shapes. Furthermore, in order to make the stimulus set more realistic, more heterogeneity by varying the angular separations of the vertices used to construct the stimuli is introduced. Specifically, the top subset of objects shown in Figure 2.4 had vertices separated by $135\,^\circ/135\,^\circ/90\,^\circ$, while the bottom set of objects had vertices separated by $180\,^\circ/90\,^\circ/90\,^\circ$. Furthermore, the visual stimuli are also rotated through 360 degrees during training to provide more natural visual training because in natural environments, objects are seen in many different orientations on the retina. This meant that there was not such a clean statistical decoupling between the boundary elements as for Study 1. A total of 24 objects were constructed in this way. Nevertheless, I expected that with the new objects used in Study 2 there would still be sufficient statistical decoupling between the boundary elements to ensure that the network developed neurons during visually guided learning that responded to a localised region of boundary curvature similar to the performance of the model in Study 1.

For all simulations in Study 2, the VisNet architecture consisted of three layers of Self-Organising Maps, where each layer is composed of 64×64 neurons. During training, the feed-forward synaptic weights are modified using the trace learning rule (1.8), (1.9), which is needed to develop translation invariant neuronal responses.

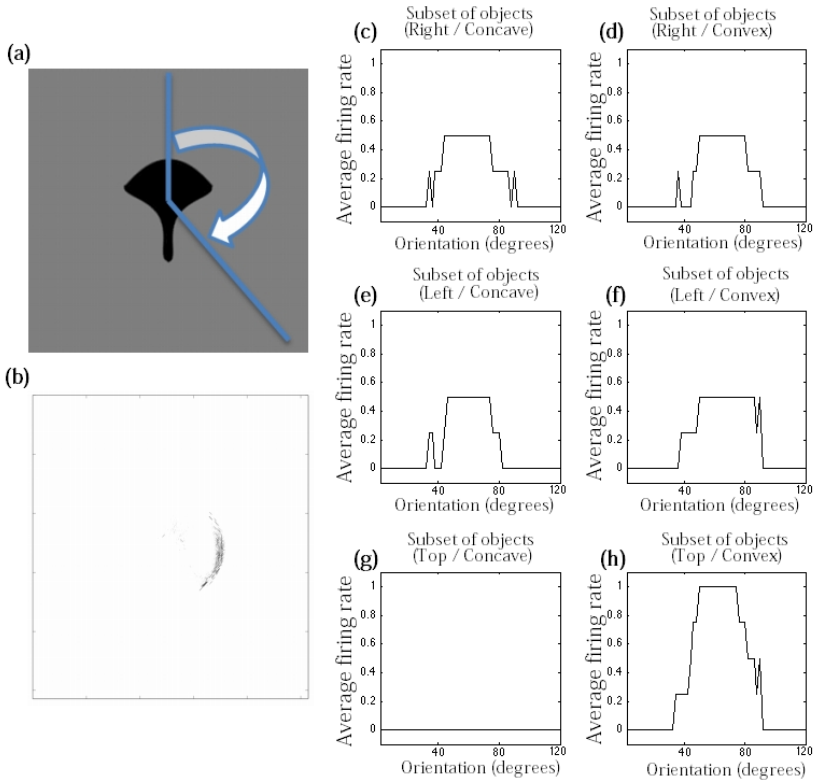

Figure 2.21: Simulation of network trained and tested with visual objects undergoing in-plane rotation. Results for an output neuron that learned to respond to a local convex contour at about 75 degrees are shown. (a) A typical visual object used in the simulation. The objects had $n = 3$ sides and each side had $p = 2$ possible boundary elements. Each object is rotated through 120 degrees in steps of 2 degrees during training and testing. This ensured that each boundary element shape (concave or convex) was seen in every rotational position (albeit in increments of 2 degrees). (b) The input Gabor filters that had strong connectivity through the layers to the output neuron. It can be seen that the neuron has strong connections from a convex boundary element at about 75 degrees of rotation. (c)-(h) These 6 plots show the average firing rate response of the neuron to the 6 subsets of objects that contain one of the different boundary contour elements as these objects are rotated through 120 degrees. These plots confirm that this neuron responds maximally when a concave contour appears at around 75 degrees orientation. The neuron does not respond in plot (g) because none of the visual objects used for this plot contains a convex contour on the top, which is what the neuron is tuned to.

2.4.2.1 Development of neurons encoding local boundary conformation in an object-centred frame of reference

In this simulation, VisNet was presented with all 24 visual stimuli shown in Figure 2.4. During training, each stimulus was rotated through 360 degrees in a single central location on the retina in steps of 10 degrees. The size of the images was 256×256 pixels, and the radius between the centre and each vertex of the objects was set to 50 pixels.

Figure 2.22 shows a comparison between the responses of a neuron recorded in area V4 of the primate ventral visual pathway by Pasupathy and Connor (2001) and a neuron recorded from our simulation, which exhibits a similar degree of selectivity. The neuron recorded by Pasupathy and Connor (2001) responds selectively to object shapes with an acute convex curvature at the top right of the object. The example neuron recorded in the VisNet simulation shows similar degree of selectivity. Moreover, many other neurons in the output layer of VisNet learned to respond selectively to particular combinations of local boundary curvature and position with respect to the centre of mass of the object. The network accomplished this even though the statistical independence of the boundary contour elements was not perfect.

To further analyse the detailed firing properties of each output neuron, its response to all objects as they were rotated through 360 degrees was recorded. Next the boundary contour

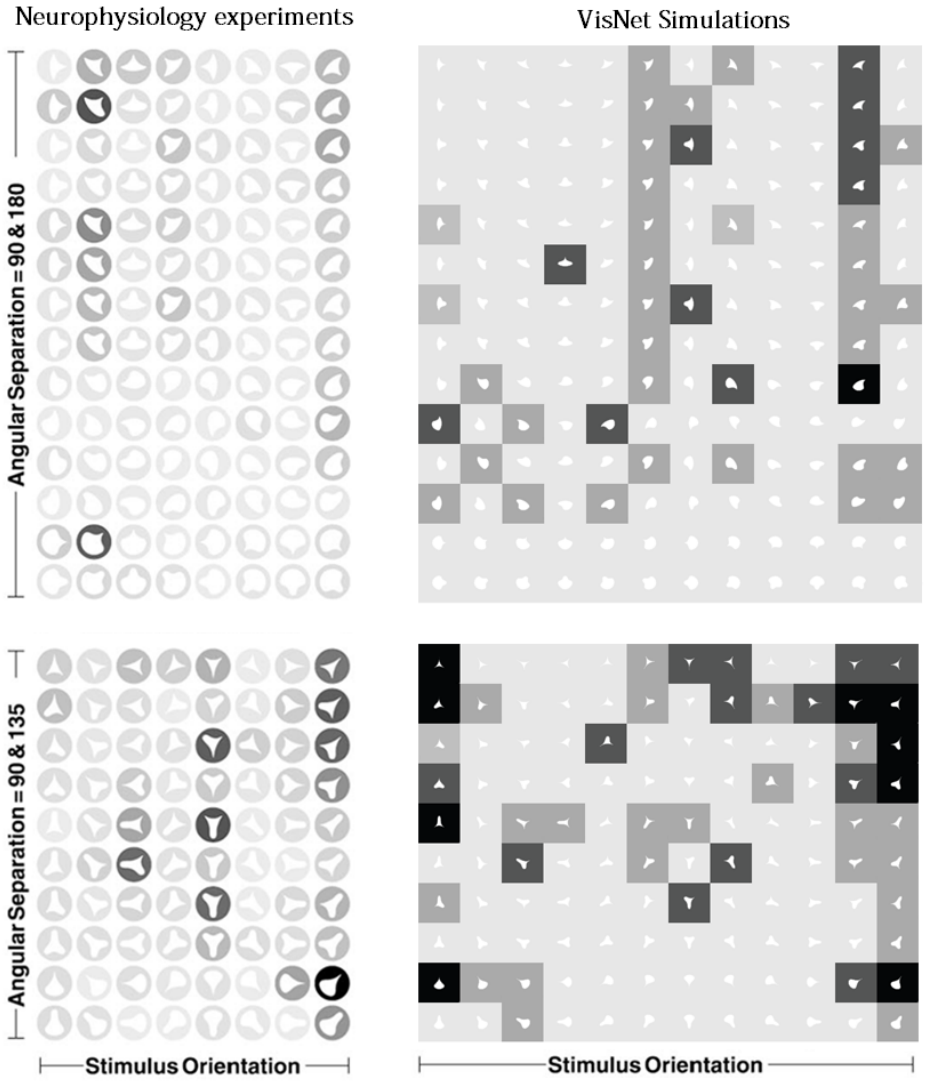

Figure 2.22: Comparison between the single neuron recording data of Pasupathy and Connor (2001) and corresponding results from VisNet simulations. On the left of the figure are shown the responses of a neuron recorded in area V4 of the primate ventral visual pathway and shown in Figure 5a of Pasupathy and Connor (2001). (Figure reproduced with permission.) Each object shape shown to the monkey is represented by a white icon, and the firing rate response of the neuron is represented by the surrounding shading with high firing denoted by black. Each row shows a different object shape, with each column corresponding to a different orientation of the object. It can be seen that the neuron responds selectively to object shapes with an acute convex curvature at the top right of the object. On the right of the figure are shown corresponding results for an output cell in layer 3 of VisNet, which has learned to respond with similar selectivity.

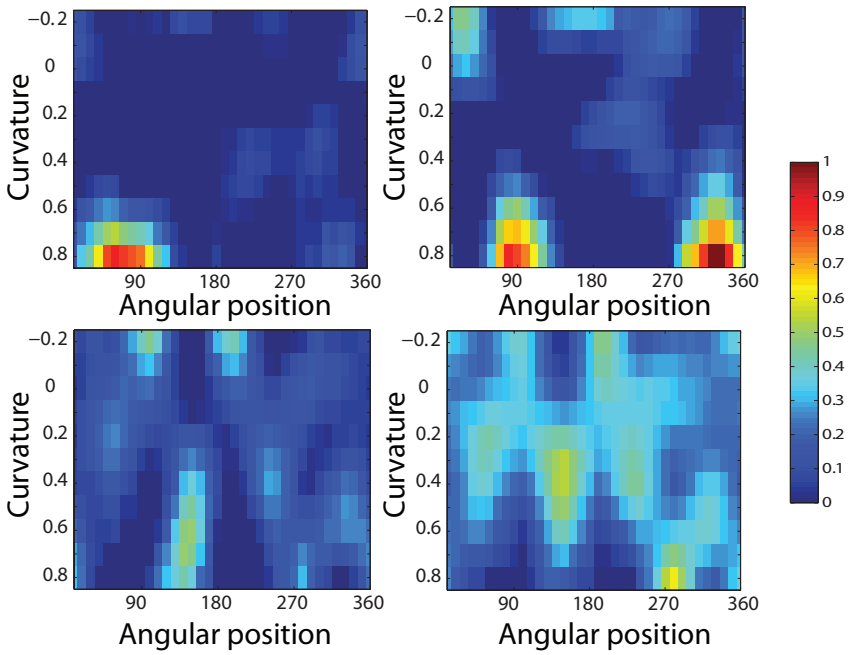

Figure 2.23: Heatmap showing the average responses of an output neuron to different combinations of local boundary curvature and angular position where the boundary curvature appears. The average is computed over all orientations (0-360 degrees) of all objects. The neuron plotted on the top-left responds maximally to object shapes with an acute convex curvature at the top-right (45 degrees). This is the same neuron that was shown on the right of Figure 2.22. The neuron plotted on the top-right responds maximally to object shape with an acute convex curvature at the top and the bottom-right. Two other cells that show different firing patterns are also plotted below to show the variability in the network.

of each object was segmented into multiple elements based on the positions where the rate of change of the curvature exceeded a fixed threshold. This then enabled us to calculate the average response of the neuron to each particular combination of local boundary curvature and angular position where that boundary curvature appears, where the average is computed over all orientations of all objects. Figure 2.23(top-left) shows a heatmap of the average responses of the output neuron shown on the right of Figure 2.22 to different combinations of boundary conformation and angular position. The plot confirms that this neuron responds maximally to object shapes with an acute convex curvature at the top right. The correlation coefficient between the result and a predicted result of a modelled V4 neuron based on Gaussian distribution, which is tuned to acute contours at 70 degree is strong (0.798) and confirms the selectivity.

In order to quantify the effect of training on the number of combinations of local boundary curvature and angular position that individual neurons were responsive to, for each neuron the number of local peaks in the heatmap of average firing rate against curvature and angular position was analysed, as shown in Figure 2.23. Specifically, for each neuron, the number of local peaks that were greater than 60% of the average firing rate across the heatmap was counted. Before training, 176 cells had one peak, 98 cells had two peaks, 63 cells had three peaks, and 44 cells had four peaks. After training, the distributions were 319 cells, 460 cells, 414 cells, and 374 cells. (These distributions were significantly different, chi-square = 17.58, df = 3, $P \ll 0.01$.) Thus, training led to a large increase in the number of neurons that were selectively tuned to either one or just a few boundary contour elements. The simulation results also predict the existence of individual neurons that are tuned to boundary elements in multiple locations. Consistent with this, Brincat and Connor (2004) have reported that some neurons in TEO do indeed respond to the co-occurrence of multiple adjacent contour elements.

2.4.2.2 Development of translation invariant neuronal responses

As noted in the earlier Section 2.4.1.6, Pasupathy and Connor (2001) and Brincat and Connor (2004) reported that neurons encoding the boundary conformation of objects also respond with translation invariance as an object is shifted across different retinal locations. Section 2.4.1.6 presented examples of translation invariance across retinal locations with simple shapes of objects. Here the result is extended with more abstract shapes shown in Figure 2.4, which are similar to those used in the neurophysiology experiments of Pasupathy and Connor (2001). To cope with the larger computational resource requirements, only the stimuli with an angular separation between vertices of $135\,^\circ/135\,^\circ/90\,^\circ$ were used, and the size of the image was reduced to 128×128 pixels. During training, the trace learning rule (1.8), (1.9) was used to modify the synaptic weights, which is needed for the neurons to develop translation invariant responses across the nine retinal locations.

In this simulation, during training each object was shifted across a 3×3 grid of nine different retinal locations, which are separated by horizontal and vertical intervals of 10 pixels. At each pixel location, the objects are presented in all orientations through 0-360 degrees in 10 degree steps. This means that during training the objects underwent two different kinds of transformation, both translation and rotation. It is therefore needed to consider how the two kinds of transformation should be temporally sequenced. To answer this, I need to consider how the trace learning rule operates. It is an important part of the underlying theory developed in this chapter that the trace learning rule (1.8), (1.9) is used to modify the synaptic weights in order to enable neurons to develop translation invariant responses across the different retinal locations. The trace learning rule operates by encouraging neurons to learn to respond to input patterns that tend to occur close together in time. This means that, in order for neurons to learn translation invariance, images corresponding to a particular object in different retinal positions must occur in temporal proximity. Contrariwise, the neurophysiology studies of Pasupathy and Connor (2001) and Brincat and Connor (2004) do not demonstrate a large degree of rotation invariance. Therefore, in order to replicate the experimentally observed response properties of translation invariance and rotational sensitivity, it is assume that typically the eyes shift about a visual scene more rapidly than the objects rotate on the retina. To simulate this effect, VisNet was trained as follows. During training, the orientation of each object was kept fixed at some initial angle while the object was shifted across all of the different retinal locations. Then the orientation of the object was adjusted by, for example, 10 degrees and the object was again shifted across all of the retinal locations. This procedure was repeated for all object orientations from 0-360 degrees in steps of 10 degrees. This training procedure ensured that images of each object in the same orientation but different retinal locations were closely clustered together in time. The trace learning rule was then able to develop translation invariant responses. On the other hand, it is assumed that images corresponding to different orientations of an object would not be clustered together in time, and so the neurons would not develop rotational invariance by trace learning. I admit that we do not have the data to verify this assumption, but this is an empirical issue that requires collecting data on natural statistics of eye-movements and object motions and is beyond the scope of the current study.

Figure 2.24 shows results for a typical output neuron after training. Each subplot shows the average responses of the neuron to different combinations of local boundary curvature and angular position. The top subplot shows the average neuronal responses over all nine retinal locations, while the remaining subplots show the average neuronal responses to each of the nine separate retinal locations. Although not perfect, it is evident that the neuron displays a pattern of selectivity for boundary curvature and angular position that is similar across the nine retinal locations. Thus, the responses of the neuron exhibit translational invariance, similar to the neurons reported in the neurophysiology experiments of Pasupathy and Connor (2001) and Brincat and Connor (2004). There are a number of factors that could potentially improve the translation invariance of neurons in the network. In particular, I expect that increasing

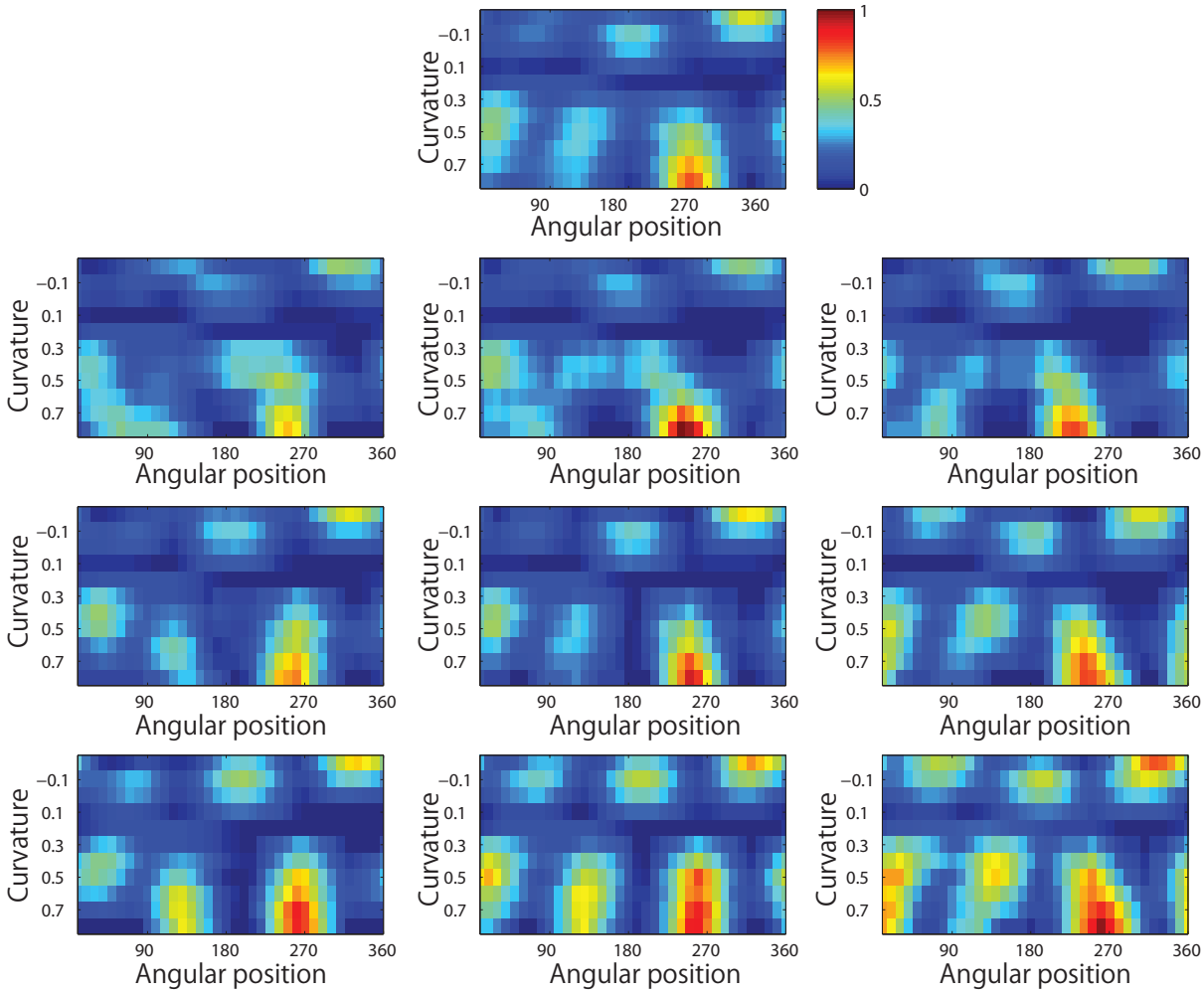

Figure 2.24: Demonstration of translation invariance after training the network on the visual stimuli shown in Figure 2.4, which were similar to those used in the neurophysiology studies of Pasupathy and Connor (2001). The figure shows the average responses of a typical output neuron over the nine retinal locations after training with the trace learning rule (1.8), (1.9). Each of the ten subplots shows the average responses of the neuron to different combinations of local boundary curvature and angular position where the boundary curvature appears (conventions as for Figure 2.23). The top subplot shows the average neuronal responses computed over all nine retinal locations. While the bottom subplots show the average neuronal responses computed separately for each of the nine retinal locations. It can be seen that the selectivity of the neuron is reasonably invariant across the nine retinal locations.

the resolution of the retinal input layer as well as increasing the numbers of neurons in the higher competitive layers should increase the capacity of the model to represent many different boundary contours in a translation invariant manner. However, the current network size is at the limit of what is practically feasible in terms of computational cost and simulation run times.

In order to quantify the distribution of such cells, the number of peaks of responses for each cell were calculated. Before training, 91 cells had one peak, 61 cells had two peaks, 34 cells had three peaks, and 24 cells had four peaks. After training, the distributions were 288 cells, 253 cells, 119 cells, and 158 cells. (These distributions were significantly different, chi-square = 1.99e+03, df = 3, P $\ll$ 0.01.)

2.4.3 Study 3: VisNet simulations with images of natural objects

In Study 3, VisNet was trained with images of natural objects in order to demonstrate that the learning mechanisms elucidated in this chapter and tested with artificially constructed visual stimuli in sections of Study 1 and 2 will indeed work effectively on real world visual objects. I

hypothesise that across many images of natural objects with different boundary shapes, there will be an effective statistical decoupling between localised boundary elements, which are defined by local curvature and angular position with respect to the centre of mass of the object. This should force the neurons in higher layers of the network to learn to respond to the individual boundary elements rather than the whole objects.

Some examples of the natural objects used in these simulations are shown in Figure 2.5 in the Section 2.2.3. The set of stimuli used in the simulations is composed of 177 realistic 3 dimensional objects. Various kinds of 3 dimensional objects are downloaded from Google 3D Warehouse, converted into grey-scaled images, and rescaled to fit on the centre of 256×256 retina. In order to enhance the realism of the visual images used to train VisNet, during training each of the natural objects is rotated in plane through 360 degrees in steps of 10 degrees.

After training, the neuronal responses in the network were examined with the test stimuli used for Study 2 (Figure 2.4), which are similar to those used in the original neurophysiology experiments of Pasupathy and Connor (2001). This allows direct comparison between the responses of neurons in the VisNet model and the experimentally observed firing characteristics of neurons that encode local boundary conformation as reported by Pasupathy and Connor (2001).

2.4.3.1 Development of neurons encoding local boundary conformation in an object-centred frame of reference

In this simulation, during training VisNet was presented with the 177 natural objects rotating through 360 degrees in one central location on the retina. Figure 2.25 shows the responses of a typical output neuron after training. This neuron learned to respond to an acute convex curvature at the bottom left of an object. Moreover, although not shown, many other neurons in the output layer of VisNet learned to respond selectively to particular combinations of local boundary curvature and angular position of the boundary element.

In order to quantify the distribution of such cells, the number of peaks of responses for each cell were calculated. Before training, 176 cells had one peak, 98 cells had two peaks, 63 cells had three peaks, and 44 cells had four peaks. After training, the distributions were 232 cells, 141 cells, 125 cells, and 103 cells. (These distributions were significantly different, chi-square = 176.82, df = 3, $P \ll 0.01$.)

This result showed that VisNet was able to develop these neuronal responses even though the network had been trained on many natural visual objects without artificially constructing the boundary shapes from artificially predefined elements.

2.4.3.2 Development of translation invariant neuronal responses

I then tested whether neurons in VisNet can also develop translation invariant responses when the network was trained on the natural objects shown in Figure 2.5. Each of the natural objects was shifted across a 3×3 grid of nine different retinal locations, which were separated by horizontal and vertical intervals of 10 pixels. At each pixel location, the objects were presented in different orientations through 0-360 degrees in 10 degree steps. The temporal sequencing of these two kinds of transforms was the same as described above. During training, the trace learning rule (1.8), (1.9) was used to modify the synaptic weights to enable the neurons to develop translation invariant responses across the nine retinal locations.

Figure 2.26 shows results for a typical output neuron after training. Each subplot shows the average responses of the neuron to different combinations of local boundary curvature and angular position. The top subplot shows the average neuronal responses over all nine retinal locations, while the remaining subplots show the average neuronal responses to each of the nine separate retinal locations. It can be seen that the neuron responds selectively to objects with a

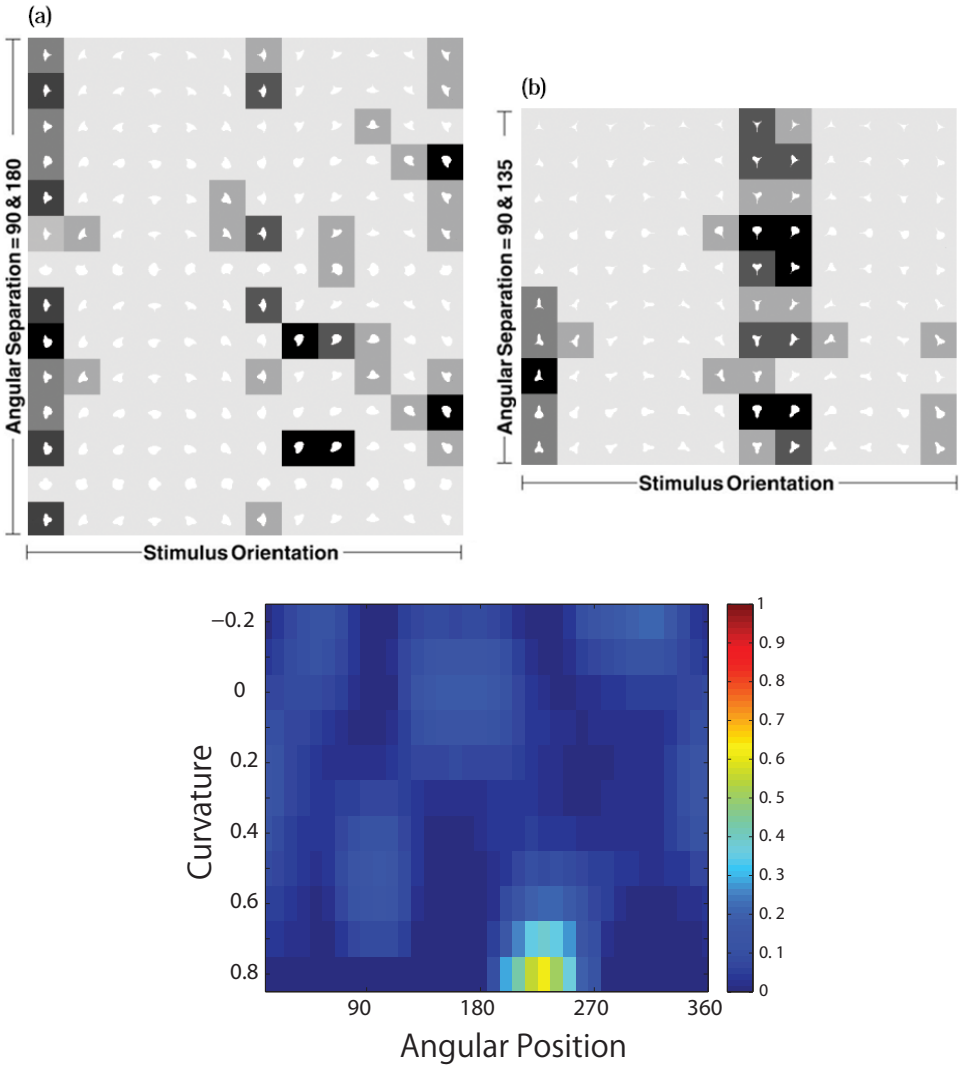

Figure 2.25: The development of neuronal responses in VisNet that encode combinations of local boundary curvature and rotational position after the network has been trained with images of natural objects as shown in Figure 2.5. The figure shows the responses of an output neuron which has learned to respond to an acute convex curvature at the bottom left (225 degree) of an object. Parts (a) and (b) show the responses of the neuron to objects with an angular separation between the vertices of 135°/135°/90° and 180°/90°/90°, respectively. Each object shape is represented by a white icon, and the firing rate response of the neuron is represented by the surrounding shading with high firing denoted by black. Each row shows a different object shape, with each column corresponding to a different orientation of the object. It can be seen that the neuron responds selectively to object shapes with an acute convex curvature at the bottom left of the object. Part (c) shows a heatmap of the average responses of the neuron to different combinations of local boundary curvature and angular position where the boundary curvature appears. The average is computed over all orientations (0-360 degrees) of all objects. The plot confirms that the neuron responds maximally to objects with an acute convex curvature at the bottom left.

high convex curvature at the top-left. Moreover, the responses of the neuron are similar across all nine retinal locations.

In order to quantify the distribution of such cells, the number of peaks of responses for each cell were calculated.

The distributions were that before training, 97 cells had one peak, 38 cells had two peaks, 25 cells had three peaks, and 31 cells had four peaks, whereas after training, the distributions were 349 cells, 148 cells, 90 cells, and 109 cells. (These distributions were significantly different, chi-square = 1.34e+03, df = 3, $P \ll 0.01$.)

Thus, the responses of the neuron are reasonably translation invariant, similar to the neurons reported in the neurophysiology experiments of Pasupathy and Connor (2001) and Brincat and

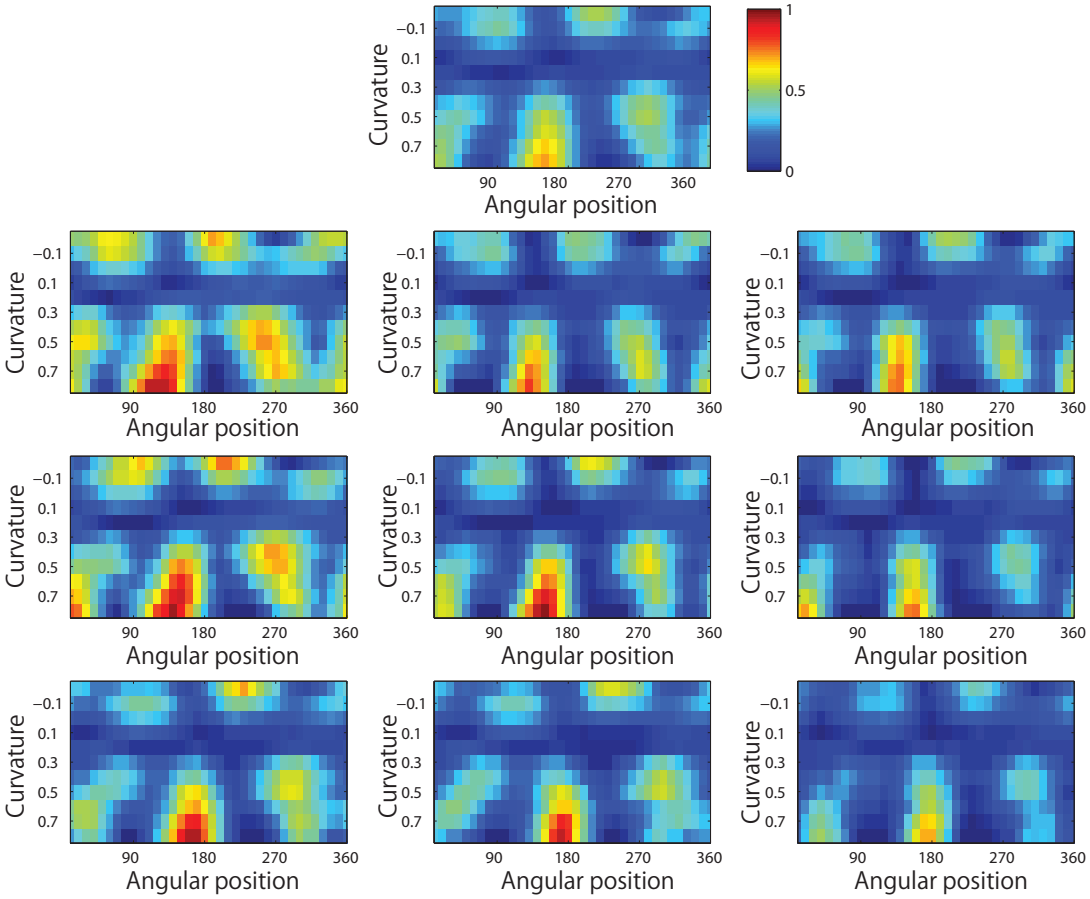

Figure 2.26: The development of translation invariant neuronal responses in the output layer of VisNet after the network has been trained with images of natural objects as shown in Figure 2.5. The network was trained with the trace learning rule (1.8), (1.9) in order to promote invariance learning across nine different retinal locations. The figure shows the average responses of a typical output neuron over the nine retinal locations after training. Each of the ten subplots shows the average responses of the neuron to different combinations of local boundary curvature and angular position where the boundary curvature appears (conventions as for Figure 2.23). The top subplot shows the average neuronal responses computed over all nine retinal locations. While the bottom subplots show the average neuronal responses computed separately for each of the nine retinal locations. The neuron has become tuned to an acute convex curvature at the top-left of an object. Moreover, it can be seen that the neuron responds similarly across the nine retinal locations.

Connor (2004).

In conclusion, the above results thus demonstrate that even when VisNet is trained on realistic natural visual objects, where the boundary shapes have not been carefully constructed from a pool of artificial elements, the network still develops neurons that respond selectively to the curvature and location of localised boundary contour elements in the frame of reference of the object. Moreover, with the help of the trace learning rule, these neuronal responses are also translation invariant as an object shifts across different retinal locations.

2.5 Discussion

In this chapter, I demonstrated that when a neural network model, VisNet, of the primate ventral visual pathway is trained on many objects with different boundary shapes, the neurons in the higher layers of the network learn to respond to localised boundary contour elements, which are defined by the curvature and location of the boundary element in the frame of reference of the object. Interestingly, neurons learn to respond to these boundary elements rather than learning to respond to the whole objects that were actually presented during training. Moreover, the neurons were able to learn to respond with translation invariance as visual objects are

shifted across different retinal locations. This was shown to be successful when VisNet was trained with either the artificially constructed visual stimuli used in Studies 1 and 2, or with images of natural visual objects in Study 3. A population of such neurons, representing many different boundary elements of different curvature and position within the object, could be used to provide a distributed coding of the entire boundary shape of an object. This has been demonstrated with real neurons in primate visual area V4 by Pasupathy and Connor (2002). As such, these neurons are likely to play an important role in how the primate visual system represents the shapes of objects.

The primary contribution of this chapter is to elucidate and test two key biologically plausible learning mechanisms that can combine to promote the development of these neuronal response characteristics. First, if the network is trained on many objects with different boundary shapes, where each boundary is comprised of a different constellation of contour elements, then this leads to a statistical decoupling between the boundary elements. This is sufficient to allow the competitive layers of VisNet to develop neurons that respond to individual boundary elements defined by curvature and position within the object, which are similar to the neurons reported in the physiological experiments conducted by Pasupathy and Connor (2001). Secondly, neurons may learn to respond with translation invariance across different retinal locations through the use of a trace learning rule. This kind of learning places constraints on the statistics of how the eyes move and visual objects change or transform on the retina. Specifically, it is assumed that the eyes usually shift about a visual scene faster than the objects are changing or rotating on the retina. These two mechanisms together provide a biologically plausible account of how neurons in the primate ventral visual pathway may learn to represent localised boundary contour elements of objects as revealed by Pasupathy and Connor (2001).

Furthermore, neurophysiological experiments carried out by Brincat and Connor (2004) have shown that neurons in the later stages of the ventral visual pathway, TEO, integrate information from multiple boundary contour elements. In our simulations, the number of cells that were tuned to combinations of multiple contours increased in the higher layers. Tracing back the feed-forward synaptic connectivity to these output neurons confirmed that their selectivities were built by combining inputs from neurons representing each local boundary contour in the preceding layer.

However, global analysis of object shape in the higher competitive layers of VisNet does not occur by merely averaging over neurons in intermediate layers that represent lower level boundary contour elements. In Elliffe et al. (2002), VisNet was trained on a set of visual objects that were constructed from the same basic alphabet of visual features, horizontal and vertical bars, which made up the boundaries of the objects. These objects thus had superset/subset relationships with each other. Nevertheless, individual neurons in the higher layers were able to learn to respond selectively to just one of the objects, with different neurons responding to each of the objects in the training set. Neurons that responded selectively to one of the objects did not respond to either simpler stimuli comprised of a subset of the features of the preferred object, or more complex stimuli comprised of all of the features plus additional features. This network behaviour results from the competitive interactions between neurons representing different object shapes in the higher layers. Thus, neurons in the higher competitive layers do not simply average over neurons in the intermediate layers, but are able to learn to respond selectively to specific combinations of low level visual features such as boundary contour elements.

The simulations reported in this present work are the first to show how neuronal responses encoding the local boundary conformation of objects may develop through a biologically plausible process of visually-guided learning. Both the Hebb learning rule (1.7) and trace learning rule (1.8), (1.9) used above are biologically plausible in that they are 'local' learning rules, which only use locally available biological quantities, such as the activity of the pre- and post-synaptic neurons, to modify the synaptic weights. This is in sharp contrast to other modelling studies that manually set up the synaptic weights in a non-local manner. In particular, the trace

learning rule drives the development of translation invariant neuronal responses. Convincing experimental evidence for the presence of trace learning in the primate visual system has been provided by Cox et al. (2005), and a plausible account of the synaptic basis of trace learning has been provided by simulations of biologically detailed integrate and fire neural networks carried out by Evans and Stringer (2012) and also later in this thesis in Chapter 7. Furthermore, the trace learning rule can be implemented in the afferent synaptic connections to all neuronal layers in the network, which avoids the biologically implausible need for separate layers for template learning and invariance learning as has been implemented in previous models. Another important factor that underpins the biological plausibility of the simulations carried out in this chapter is that the network model was always trained on whole objects rather than carefully pre-segmented and isolated parts of objects corresponding to local boundary elements. Indeed, in Study 3, VisNet was trained on a random assortment of whole natural visual objects. Nevertheless, the network was still able to develop neurons that were specifically tuned to localised boundary segments of objects. I also found the performance of the model to be extremely robust, which gives additional credence to the learning mechanisms explored in this chapter.

2.5.1 Future work

In future work, it will be important to investigate how VisNet learns to represent the spatial structure of objects when the network is trained with more realistic eye and head movements, more realistic network architectures that also incorporate top-down connections between layers and recurrent connections within layers, as well as a variety of more complex visual objects with internal structure.

The simulation studies reported in this chapter showed how neurons can develop translation invariant firing properties, but respond selectively to the orientation of an object. This was achieved by trace learning, which encourages neurons in higher layers to respond to subsets of input patterns that tend to occur close together in time. In this case, trace learning will encourage neurons to learn to respond to objects or object features with translation invariance across the retina. In order to simulate eye movements, the simulations reported in this chapter simply shifted each individual object in turn across a number of locations on the retina before moving on to presenting the next object. However, of course, these simulated eye movements were highly idealised. In future work I propose to record real eye and head movements as human subjects visually examine a test environment. This will, for example, permit objects to appear to rotate continuously in plane on the retina. If this is combined with trace learning, then this could encourage some neurons to develop rotation invariant responses. Therefore, it will be important to feed real recorded eye and head movements into the model to test whether the experimentally observed response properties continue to develop.

Also, the version of the VisNet architecture used in this chapter incorporated associative learning only in the bottom-up (feedforward) connections between successive layers of the network. Furthermore, no top-down connections was included in the model even though these are known to exist in the primate ventral visual pathway. The rationale for using this simplified architecture in the current study was that it is sufficient to replicate how neurons in V4, TEO, and TE are able to learn to encode the conformation of boundary contour elements at a particular position within an object. However, Zhou et al. (2000) have shown that the responses of neurons in earlier stages of visual processing such as V1 and V2, which have preferred responses to oriented edges, are also modulated by which side of a figure the edge occurs on. This is the case even when the figure/background cues lie well outside the classical receptive field of the neuron. This suggests that global image context specifying border ownership modulates the activity these neurons. This contextual information must be conveyed to these early stage visual neurons by some combination of top-down connections between layers and recurrent connections within layers. This problem is investigated in detail in Chapter 5.

Another issue that needs to be addressed in future modelling work is how the shape tun-

ing properties of neurons in V4 and other areas may evolve over tens of milliseconds during a stimulus presentation. For example, Yau et al. (2013) analysed the time course of curvature processing in V4. They found that the shape tuning of V4 neurons evolves over tens of milliseconds from individual contour fragments to multifragment shape representations. The modelling study presented in this chapter has not sought to explain how this dynamical tuning phenomenon may occur. However, given the above discussion about the potential combined bottom-up and top-down flow of visual information within the primate ventral visual pathway, I speculate that dynamical changes in tuning properties over time may arise as the hierarchical system gradually settles into a stable representation, with the firing of neurons in lower and intermediate layers eventually modulated by top-down signals. In future work, I intend to replicate this modelling study with a time-accurate, integrate & fire neural network, which is needed to faithfully simulate the temporal dynamics of real neurons. This will permit us to investigate the precise millisecond time course of the neuronal response properties. This problem is investigated in detail in Chapter 6 and Chapter 7.

Furthermore, another question is whether the approach proposed here can be extended to 3D shape. Yamane et al. (2008) have demonstrated the existence of neurons that encode the 3D configuration of localised surface fragments defined by their conformation, orientation and position with respect to the centre of mass of the object. A population of such neurons provides a distributed representation of an object's 3D shape. The response characteristics of these neurons are also invariant as the object is shifted through different locations on the retina. It will be important to evaluate if a model such as VisNet, trained using stereoscopic input, can begin to capture the partonomic structure of 3D objects. Furthermore, it will be critical to assess whether learning rules, such as trace learning, can still be used to generate translationally invariant recognition processes.

Further computer simulations should be aimed at understanding how neurons that encode 3D surface structure develop their firing properties through visually guided learning, and understanding how populations of such neurons could provide distributed representations of the full 3D structure of visual objects. To carry out these simulations, the architecture may be extended to incorporate stereoscopic visual inputs corresponding to the left and right eyes.

I expect that neurons may learn to represent 3D surface patches through learning principles similar to those proposed in this chapter. That is, when the network is trained on many different 3D object shapes constructed from combinations of surface patch elements, I expect that the surface patch elements will be statistically decoupled from each other. In this case, I hypothesise that the higher layers of VisNet will form representations of the statistically independent 3D surface patches. Furthermore, I hypothesise that such neurons may also develop translation invariant reposes through the kind of trace learning mechanisms demonstrated above.

However, theorists have long posited that the visual system in fact represents complex 3-dimensional shapes, such as a table or a chair, by decomposing it into volumetric parts with axial symmetry (Biederman, 1987). A recent fMRI study in humans has provided evidence for this at the level of the neuronal population, where it was found that the visual system explicitly represents the relationships between the medial axes of linked object parts (Lescroart and Biederman, 2013). Consequently, more recently, Hung et al. (2012) have investigated medial axis shape coding in the inferotemporal cortex. This work extended their studies of parts-based spatial representations to 'skeletal' representations involving a configuration of volumetric parts, where each part has an axis of radial symmetry or medial axis. The 3-dimensional structure of an object may then be represented by a combination of the relationships between the medial axes of the object parts as well as the conformations of the surfaces of the object parts. Hung et al. (2012) confirmed that individual neurons in IT do in fact encode a configuration of both medial axis and surface fragments. In future work, we shall investigate whether the computational learning mechanisms demonstrated in this chapter may also give rise to these kinds of skeletal representations.

Chapter 3

The Neural Basis of Face Representations

Experimental studies have shown that neurons at an intermediate stage of the primate ventral visual pathway, occipital face area (OFA), encode individual facial parts such as eyes and nose while neurons in the later stages, middle face patches, are selective to the full face by encoding the spatial relations between facial features. In this chapter, I perform a computer modelling study to investigate how these cell firing properties may develop through unsupervised visually-guided learning. A hierarchical neural network model of the primate's ventral visual pathway, VisNet, is trained by presenting many randomly generated faces to the network while a local learning rule modifies the strengths of the synaptic connections between neurons in successive layers. After training, the model is found to have developed the experimentally observed cell firing properties. In particular, I show how the visual system forms separate representations of facial features such as the eyes, nose, and mouth as well as monotonically tuned representations of the spatial relationships between these facial features. I also demonstrate how the primate brain learns to represent facial expression independently of facial identity. Furthermore, based on the simulation results, I propose that neurons encoding different global attributes simply represent different spatial relationships between local features with monotonic tuning curves or particular combinations of these spatial relations.

3.1 Introduction

The ability of the brain to analyse and recognize faces under natural viewing conditions is unmatched by today's computer vision systems. In order to achieve this singular ability, the primate brain develops and utilizes a rich tapestry of cells that encode different kinds of visual information about faces. For example, some neurons respond to the presence of facial features such as the eyes, nose, or mouth, while other neurons encode the many spatial relationships between these facial features (Freiwald et al., 2009). Some neurons also encode global properties such as facial identity or expression (Morin et al., 2014; Hasselmo et al., 1989a). Our ability to process and recognise faces utilises this rich tapestry of different kinds of visual information. Understanding how these diverse visual representations develop through sensory-guided learning may help to inform future research into computer vision for facial analysis and recognition.

3.1.1 Hierarchical representations of faces along ventral visual pathway

Functional magnetic resonance imaging (fMRI) studies in humans have revealed several cortical regions within the temporal lobe, which are exclusively dedicated to face processing (Perrett et al., 1992; Kanwisher et al., 1997; Pitcher et al., 2011; Zhang et al., 2012). In particular, there is evidence for hierarchical processing. For example, an early stage of processing, the occipital

face area (OFA) in the inferior occipital gyrus, has been found to contribute to face perception by responding to individual facial features such as the eyes, nose, and mouth (Pitcher et al., 2011). On the other hand, a later stage of processing, the fusiform face area (FFA) in the lateral fusiform gyrus, has been found to integrate such information by responding more strongly to intact rather than scrambled faces (Kanwisher et al., 1997; Zhang et al., 2012). Recently, it has also been reported that the face areas may also exhibit "faciotopy" where different cortical patches represent different face features, and the cortical distances between the feature patches reflect the physical distance between the features in a face (Henriksson et al., 2015).

In macaques, several face sensitive areas have been identified in the temporal lobe, which are known as *face patches* (Gross et al., 1972). It has been argued that the homologue of the FFA in macaques is the *middle face patch* (Tsao et al., 2006). In one single unit recording study by Freiwald et al. (2009), cartoon faces were presented to the monkey. The cartoons were systematically modified by varying the number of facial features present, as well as the spatial relationships between the features such as the distance between the eyes. It was found that the middle face patch also integrated information across facial features. That is, neurons were found to respond to different combinations of facial features (Figure 3.1) and the spatial relations between them (Figure 3.2). Furthermore, the tuning profiles of individual neurons that were selective to such spatial relations were typically ramp-shaped between two extremes, which even transgressed the limits of realistic face space. They reported that the responses of neurons encoding spatial relations between facial features were thus amplified for extreme values of these relations compared to intermediate values. Such amplifications may explain the neural mechanisms for the results of experiments showing that faces which are more deviant in appearance are recognized better than those that are more typical (Rhodes, 1997; Benson and Perrett, 1991; Bruce and Young, 2011).

In this chapter, I investigate through computer modelling how some neurons may learn to respond selectively to individual facial features, or subsets of facial features, even when the model is always trained on whole faces with all of the facial features present. I also investigate how some other neurons learn to encode the spatial relations between facial features, such as distance between the eyes, with monotonic tuning profiles.

3.1.2 Representations of global attributes of faces

In addition to individual facial features, the primate visual system is able to process various global attributes of faces such as identity, emotional expression, age, race, gender, etc. (Homola et al., 2012; Freeman et al., 2010; Morin et al., 2014; Hasselmo et al., 1989a). Past theoretical work has suggested that the different attributes of a face such as its identity and expression are processed by functionally and anatomically separated pathways. For example, a highly influential psychological model of face processing proposed by Bruce and Young (1986) hypothesized a series of distinct stages involved in face processing. Consistent with this model, experimental studies in macaques have reported that distinct sub-populations of neurons encode different global facial attributes across a number of areas of the primate visual system. For example, it has been shown that the inferior temporal gyrus contains cells that are primarily selective to facial identity, while the adjacent superior temporal sulcus (STS) contains cells that primarily respond to facial expression (Hasselmo et al., 1989a; Perrett et al., 1992; Engell and Haxby, 2007; Wegrzyn et al., 2015). Moreover, a recent study has reported that the number of neurons that encode global attributes such as identity and expression increase along the visual pathway (Morin et al., 2014). This is a quite extraordinary finding since primates are usually exposed to whole faces during visual development. The question is then how might the visual system separate the representations of these global facial attributes, which are always seen together, into different brain areas.

Based on such functional and anatomical specialization, Haxby et al. (2000) hypothesized the existence of a 'core system' that is dedicated for visual analysis in the temporal lobe. The core

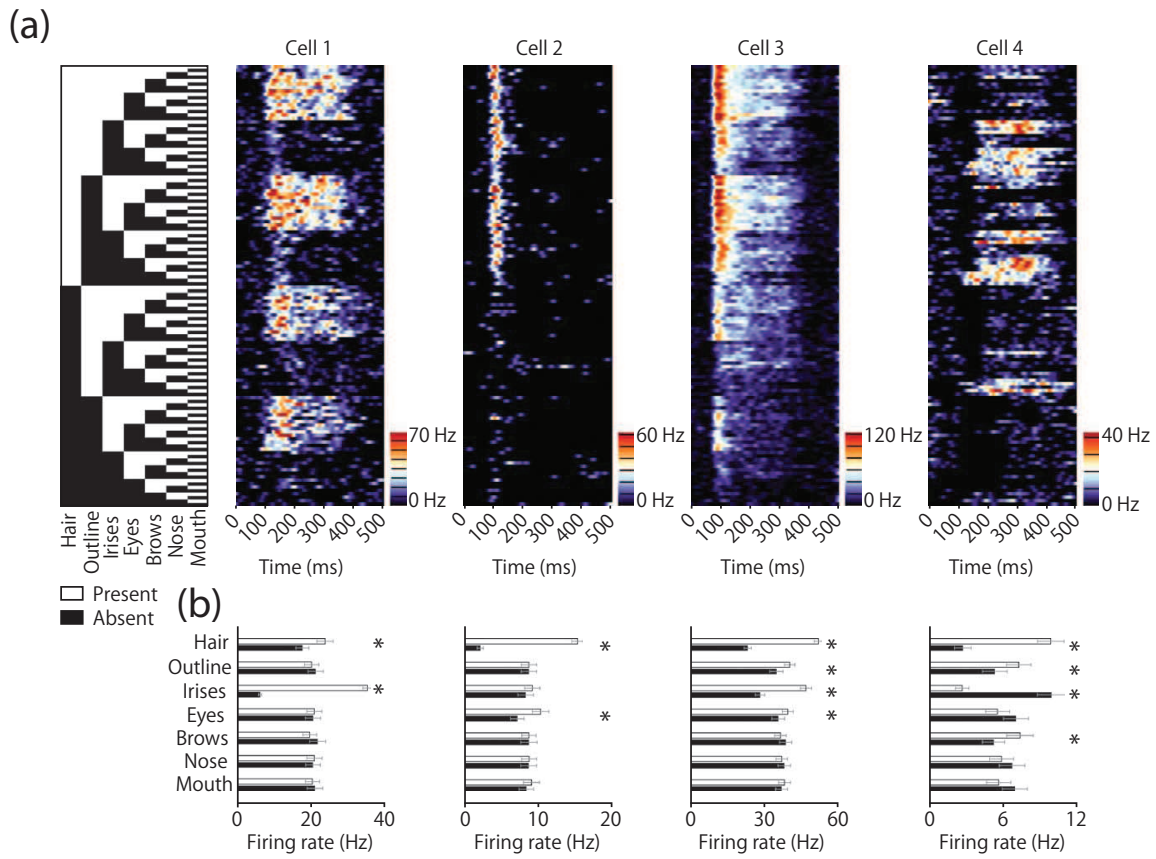

Figure 3.1: Physiological evidence from a single unit recording study carried out by Freiwald et al. (2009) showing neuronal selectivity for face parts in the primate ventral visual pathway. In this study, cartoon faces were shown to a macaque while the responses of neurons in the middle face patch were recorded. The face stimuli were varied across trials by varying which combination of facial features was present. The top panel (a) shows which facial features were present on each trial (left), and the corresponding responses of four example cells. All combinations of seven face parts (hair, outline, irises, eyes, eyebrows, nose, and mouth) were shown, including the whole cartoon face with all features (top row) and a gray background without any face features (bottom row). The responses of the four example cells to each of the face stimuli are shown as a function of time. The bottom panel (b) shows the average neuronal responses in the presence (white bars) or absence (black bars) of a given face part. * indicates significant modulation. Cell 1 fired significantly more strongly when irises were present and when hair was present. Cell 2 was influenced by two facial features, and cells 3 and 4 by four facial features. Cell 4 responded more strongly when irises were absent than when they were present. In cell 4, interactions between face parts were stronger than in the other cells, giving rise to less regular responses across stimulus conditions. Figures are reproduced with permission from Freiwald et al. (2009).

system includes the OFA that detects simple features of faces, the STS that processes changeable attributes of faces such as expressions, and the FFA that processes invariant attributes of faces such as identity.

However, these previous theoretical and experimental studies do not explain the precise learning mechanisms by which these neuronal representations of global attributes, such as identity and expression, may become mapped onto separate processing areas in the later stages of the visual system. Recently, Tromans et al. (2011) developed the first neural network model demonstrating how physically separated representations of facial identity and expression may develop through a biologically plausible process of unsupervised competitive learning. Nonetheless, this modelling study used highly idealised cartoon faces, in which these two global attributes were artificially encoded by different facial features. In the simulations described below, these learning mechanisms using much more realistic face stimuli produced using the FaceGen 3D face modelling software package, which generates stimuli based on real faces (FaceGen, 2013), are investigated. This permits a fine-grained study of how facial representations gradually develop through successive stages of processing within the network, until different forms of global

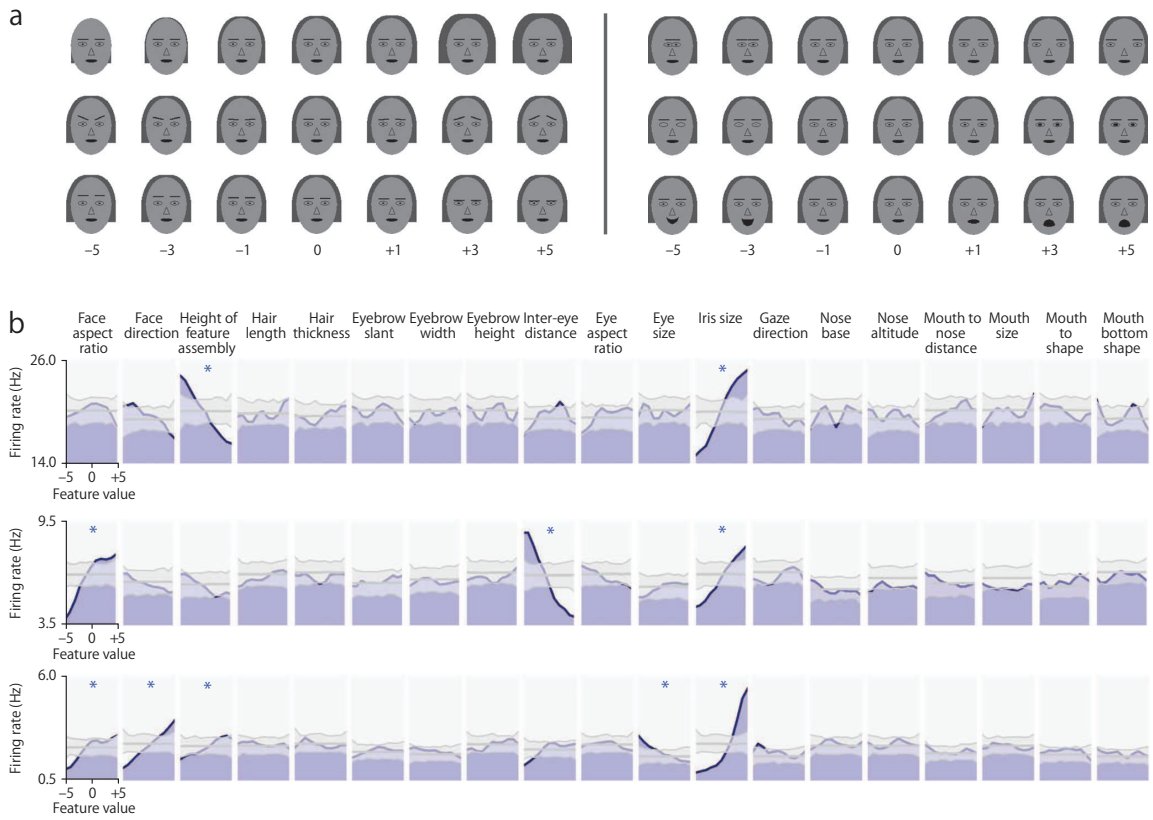

Figure 3.2: Physiological evidence from the single unit recording study carried out by Freiwald et al. (2009) showing neuronal selectivity for the spatial relationships between facial features with monotonic tuning curves. The top panel (a) shows example cartoon face stimuli for six different feature dimensions (hair width, eyebrow slant, eyebrow height, inter-eye distance, iris size and mouth shape) with seven feature values each spanning the entire range of values. The bottom panel (b) shows the response curves of three example cells to each of 19 feature dimensions. For each of the feature dimensions, the response curve (blue) is shown at a delay corresponding to maximal modulation. Maximal, minimal and mean values from the shift predictor are shown in gray. Asterisks mark significant modulation. Figures are excerpted with permission from Freiwald et al. (2009).

attribute eventually appear in distinct regions of the highest layers.

3.1.3 Computational modelling studies

While many modelling studies have investigated various kinds of processing in the primate visual system (Eliasmith et al., 2012; Serre et al., 2005), most of these investigations have not been concerned with uncovering the synaptic learning mechanisms by which the visual representations develop in the first instance. However, there is a large body of experimental evidence for learning of visual form recognition within the temporal lobes (Wallis, 2013). For example, Baker et al. (2002) showed that exposure to abstract shapes formed by combing multiple parts enhanced both parts-level and holistic shape tuning of neurons in the Inferior temporal cortex (IT). Studies with fMRI have also reported large-scale alteration of the organization and selectivity of temporal lobe cortex in humans after training with visual stimuli (Beeck et al., 2006; Gillebert et al., 2008). Additionally, although the discrimination of non-face objects is known to be more difficult than for faces, training on non-face objects improves the discrimination of these stimuli to nearly that of faces (Yue et al., 2006). Thus, how visual representations develop through a biologically plausible process of visually-guided learning is a key question that needs to be addressed by theoreticians, and is a fundamental aspect of the model simulations presented in this chapter.

Lades et al. (1993) presented the first self-organising neural model that developed representations of the spatial relationships between facial features. Their model employed a feature based approach to face recognition via active dynamic linking of features (von der Malsburg, 1981; von der Malsburg and Schneider, 1986). The model uses an input representation in which each face is convolved with a set of Gabor filters across the visual field. The output layer of the network constructs a graph representation of the face, with each node in the graph representing a particular facial feature, and each link representing the feature relation (Lades et al., 1993). The model was proposed to be biologically realistic because the development of the output face representations is unsupervised. However, it is not clear how a population of neurons in the brain may store facial representations in the form of graphs.

A more biologically plausible approach to modelling how transform (e.g. location or view) invariant representations of faces and non-face objects may develop through unsupervised, associative learning mechanisms in the higher stages of the ventral visual pathway was carried out by Wallis and Rolls (1997), which is VisNet (see Section 1.3). The inputs are represented as columns of V1-like spatial filter activation values similar to the model proposed by Lades et al. (1993). The architecture consists of four layers of competitive neural networks representing successive visual areas V2, V4, TEO and TE. During training with visual images of faces and other objects, the feedforward synaptic connections between successive layers were modified by a biologically plausible, local, associative learning rule.

The study showed that competitive learning allows neurons in the intermediate layers of the model to learn to respond to particular combinations of simple visual features present in faces and non-face objects. By building on these intermediate layer representations, the higher layers were then able to develop transform invariant representations of whole faces. More recently, Wallis (2013) has started exploring various aspects of recognition which are generally regarded as unique to faces such as holistic processing (Tanaka and Farah, 1993), configural processing (Leder and Bruce, 1998), sensitivity to inversion (Yin, 1969; Maurer et al., 2002), the other-race effects (Chance et al., 1982). Besides, the development of face representations within a more biologically accurate spiking neural network model with spike-timing dependent plasticity (STDP) has been presented by Masquelier and Thorpe (2007).

However, these previous studies have not yet fully explained how these representations correspond to those observed in single unit recording neurophysiology studies develop through successive layers of the model.

In the current work, I investigate how successive layers of VisNet develop representations of individual facial features and the spatial relations between these features as reported in the neurophysiological studies described in the introduction. In particular, I show how these cell firing properties may develop naturally through a biologically plausible process of visually-guided learning when the network is trained on realistic face images generated using the FaceGen 3D face modelling software (FaceGen, 2013).

3.2 Hypothesis

In this present study, the followings are considered: (1) how some neurons along the successive stages of processing learn to represent individual facial features such as the eyes, nose, and mouth given that the visual system is always exposed to whole faces, (2) how some neurons learn to represent particular spatial relationships between facial features, such as the distance between the eyes, with monotonic tuning curves, (3) how some neurons in later stages learn to respond to global attributes such as either a particular identity or expression, and (4) what is the relationship between spatial configurations of facial parts and global representations of face identity and expression.

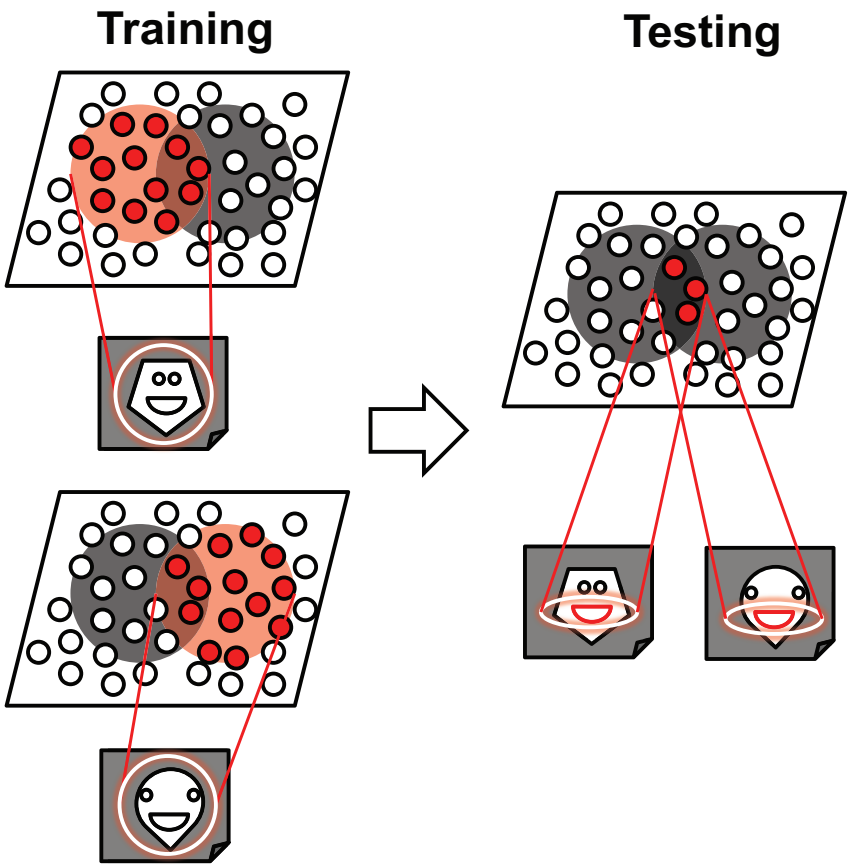

Figure 3.3: Illustration of how the network model develops neurons that have learned to respond to individual facial feature (mouth) from whole faces via statistical decoupling, which is similar to the mechanism of which V4 neurons may learn to represent the shapes of local boundary elements shown in Figure 2.3 (Eguchi et al., 2015). Left: during training, the network is presented with many different faces, where each shape is defined by a unique combination of facial features of different shapes. Two such faces are shown here. Each of these faces stimulates a different subset of neurons in the output layer of the network. The two faces shown have the mouth in common. As a result, this mouth becomes especially strongly connected, through associative learning in the feed-forward synaptic connections, with the intersection of the two subsets of output neurons shown. This intersecting subset of neurons will come to represent the mouth of the particular shape in the two faces. Right: during testing, whenever the mouth is part of a face, the same intersecting subset of output neurons will be activated. A similar learning process will drive the development of many other subsets of output neurons representing different individual facial features.

3.2.1 How some neurons learn to represent individual facial features such as the eyes, nose, and mouth

3.2.1.1 The representation of individual local facial features

In Chapter 2, I considered how neurons in V4 learn to respond selectively to the shape and location of localised boundary contour elements in the frame of reference of the object, and how neurons in the area TEO learns to respond to localised combinations of boundary contour elements (Eguchi et al., 2015). A biologically plausible solution for the development of such cells is provided by showing that the statistical decoupling (Stringer et al., 2007; Stringer and Rolls, 2008) which occurs between different forms of boundary contour element over a large population of different object shapes is a sufficient mechanism for the process. I hypothesise that a similar learning mechanism may operate to enable the network to learn to represent the individual face parts within a whole face as shown in Figure 3.3.

Let us assume that each face is comprised of n different kinds of local facial feature such as the eyes, mouth, and facial outline, and that each such facial feature may occur in p different possible shapes. In this context, presenting a whole face to VisNet can be seen as presenting an n-tuple of different facial features simultaneously to VisNet. With p possible shapes for each

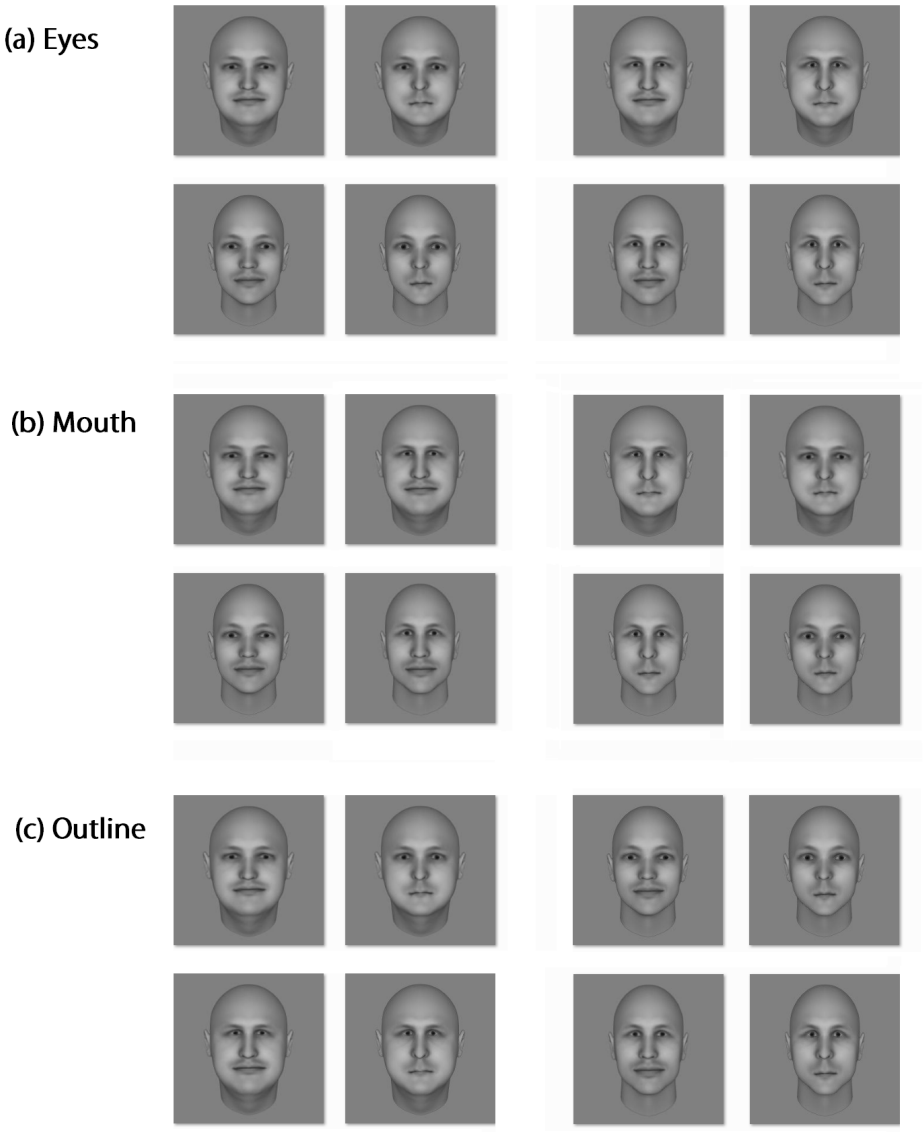

Figure 3.4: Example set of realistic faces used to train VisNet. This set of faces was generated by systematically varying the $n = 3$ facial features eyes, mouth, and facial outline. In this example, there are $p = 2$ shape variations of each facial feature. This gives a total number of $p^n = 8$ faces that may be constructed by combining the facial features in different combinations. The top two rows show how the shape of the eyes is varied. These two rows show all possible $p^n = 8$ faces, where faces with the first shape of the eyes are shown on the left and the faces with the second shape of the eyes are shown on the right. Similarly, the middle two rows show how the shape of the mouth is varied, where faces with the first mouth shape are shown on the left and faces with the second mouth shape are shown on the right. Lastly, the bottom two rows show how the facial outline is similarly varied between two different shapes shown on the left and right.

facial feature, the number of distinct whole faces that may be constructed is p^n. This means that if the number of identifiable facial features n is constant, the number of possible whole face input patterns grows polynomially with p. This polynomial increase in the representational burden makes it increasingly difficult for the network to develop non-overlapping output representations of all of the faces which are comprised of unique combinations of n facial features. Therefore, I hypothesise that at a certain point, it becomes less likely that neurons represent all the possible p^n whole faces, consisting of n-tuples of features, but rather the neurons may start to represent individual facial features.

For example, consider the case shown in Figure 3.4. This figure shows a set of faces which are systematically composed of the combination of two possible shapes of the eyes, mouth, and

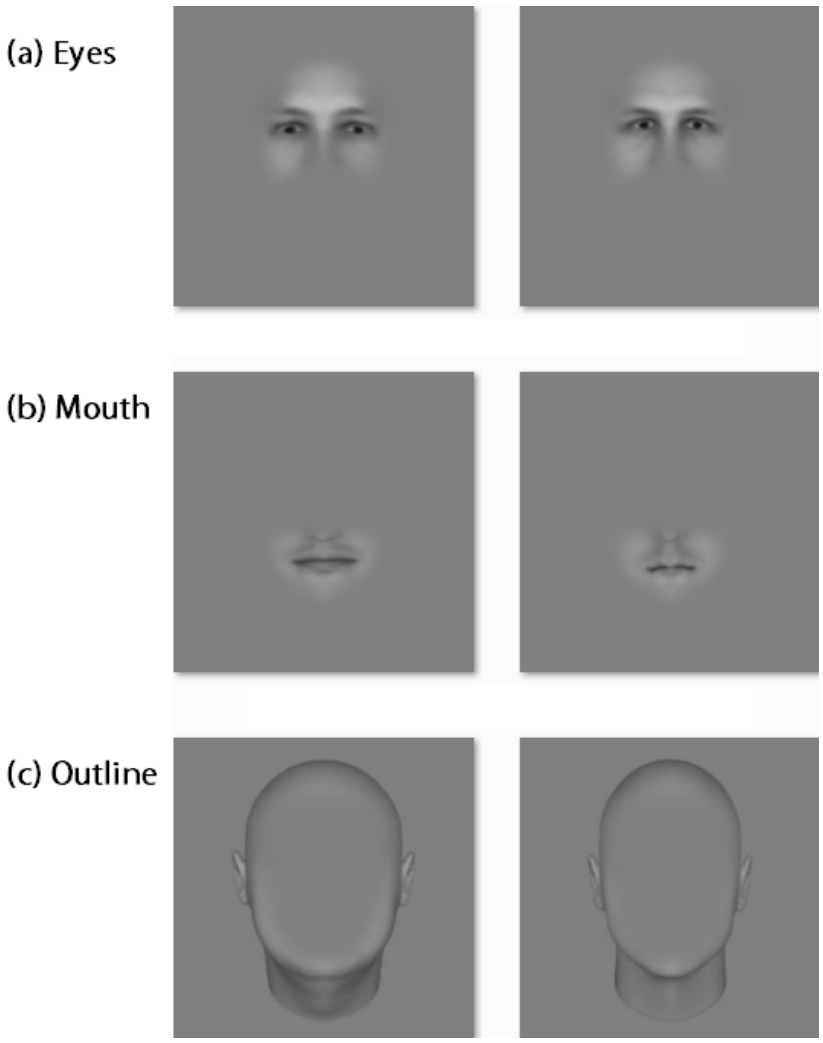

Figure 3.5: Example set of stimuli used to test VisNet after training. Each of the test stimuli is constructed by extracting one of the facial features used during training. In this example, each test stimulus contains one of the $n = 3$ facial features: (a) eyes, (b) mouth, and (c) facial outline. Each of the four facial features has $p = 2$ possible shape variations.

facial outline. Therefore, there are $n = 3$ facial features and $p = 2$ shapes of each facial feature, which may be used to construct a total of $2^3 = 8$ different faces. In the VisNet simulations reported in the first half of the study 1b in Section 3.4.1.2, the cell firing properties of the trained network between two conditions were compared: the network trained with only one face ($n = 3$ and $p = 1$) and the network trained with 8 faces ($n = 3$ and $p = 2$). In order to minimize the number of cells that happen to exclusively respond to a particular element due to the topologically distributed feed-forward connections of VisNet, each face was shifted across four different retinal locations during training. This would help to confirm the role of statistical decoupling between facial features of different shape, through exposure to many different faces comprised of different combinations of feature shapes, in the development of representations of individual facial features.

In order to test the hypothesis, the responses of neurons in VisNet to stimuli that contained just one of the facial features were recorded during training. An example set of such test stimuli is shown in Figure 3.5. These test stimuli allowed us to test whether neurons learned to respond to a specific shape of a particular facial feature as the number of possible shapes p is varied.

3.2.1.2 Shape invariant representations of local facial features

However, if the number p of possible shapes of each facial feature continues to increase, then the range of possible shapes for each feature will begin to form a continuum of gradually changing shapes. Each facial feature will then change its shape in a gradual and continuous manner across different faces. In this situation, the invariance learning mechanism continuous transformation (CT) learning (Stringer et al., 2006) may begin to operate as described in Section 1.4.1. If an individual facial feature is seen by the network in a large number p of gradually changing shapes, then the different feature shapes may be bound together onto the same shape invariant neurons in the higher layers of the network by CT learning. This will lead to the development of shape invariant representations of local facial features.

The concept of this learning process is somewhat analogous to that demonstrated in a previous simulation study (Tromans et al., 2012), which investigated the development of transform (view) invariant representations of individual rotating objects when multiple objects were presented rotating together during training. In this simulation study, the objects were presented rotating smoothly across many different views with only small (i.e. 1 degree) changes in orientation between successive transforms. It was found that if two objects were rotated independently of each other, leading to a statistical decoupling between any two particular views of the two objects, then the output neurons in VisNet learned to respond with transform (view) invariance to either one object or the other. The object specificity of the neuronal responses was driven by the statistical decoupling between any two particular views of the two objects during training, while the view invariance of the responses was driven by CT learning across the gradually changing views of each object. In this way, the network developed separate transform (view) invariant representations for each object.

In our setting, each of the facial features can be regarded as a different object, and the many possible shapes of each facial feature may be regarded as a near continuous space of gradually changing transforms. Because the changing shapes of any two different facial features are independent of each other across many faces, it is expected that the network will first develop neurons that respond to specific local facial features even when the network is always exposed to whole faces during training. However, as the number p of possible shapes of each facial feature continues to increase, then it can also be hypothesized that CT learning will begin to drive the development of shape invariant responses to specific facial features.

In the VisNet simulations reported in the later half of the study 1b in Section 3.4.1.2, only the shape of the eyes was varied, and how shape invariant eye selective cells may develop was tested. In particular, the effects of CT learning on the nature of the facial feature representations that developed in the network during training were tested by varying the number of shapes of the eyes p (5, 10, or 30 shapes) as shown in Figure 3.6. I hypothesized that as p increases, the facial feature representation turns from shape specific to shape invariant. In order to test the hypothesis, the responses of neurons in VisNet to stimuli that contained just eyes were recorded. An example set of such test stimuli is shown in Figure 3.7. These test stimuli allowed us to test whether individual neurons learned to respond to just one or a number of different eye shapes as the number of shapes p was increased.

As a result of the development of representations of individual facial features within the visual hierarchy, I conjectured that neurons in higher layers would start to process these representations and consequently develop various other related response properties. In particular, Young and Yamane (1992) have previously reported that the cells in the middle face patch carried information about the identity of individual faces distributed across the population. Furthermore, they showed that the code for macaques looking at faces of men could be summarized by two dimensions of configurations of facial parts: the amount of hair possessed by the face and an elongated to round face dimension. This results highlight the importance of both featural and configural information to the holistic process (Tanaka and Gordon, 2011).

I propose that representations of individual shapes of local facial features could contribute

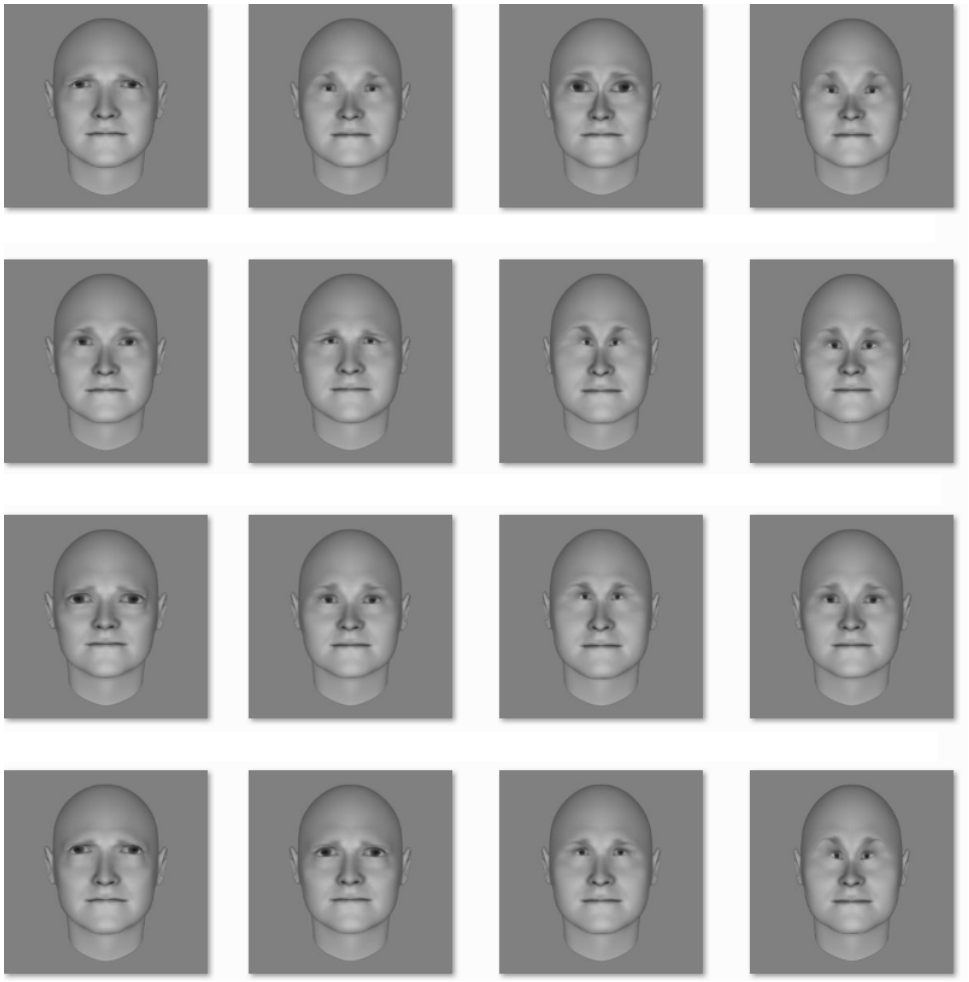

Figure 3.6: Example set of realistic faces used to train VisNet. This set of faces was generated by varying the shape of the eyes.

to the development of representations of the spatial relationships between these facial features (Freiwald et al., 2009) as well as the global properties of faces such as identity and expression (Hasselmo et al., 1989a; Morin et al., 2014). At the same time, a collection of shape invariant representations of individual facial features may contribute to the development of global representations of whole faces.

3.2.2 How some neurons learn to represent particular spatial relationships between facial features with monotonic tuning curves

In neurophysiology studies, neurons in the primate middle face patch have been found to encode spatial relationships between facial features, such as the distance between the eyes, with monotonic tuning curves (Freiwald et al., 2009). For example, some neurons that encode the distance between the eyes respond maximally when the eyes are furthest apart, and reduce their responses monotonically as the eyes get closer together. On the other hand, other neurons respond maximally when the eyes are closest together and reduce their responses monotonically as the eyes move further apart. I hypothesise that these monotonic tuning curves develop naturally as a result of competitive learning on the afferent connections into that cortical area when individual neurons receive connections from a physically localised region of the preceding area.

A basic competitive neural network architecture is shown in Figure 3.8. It consists of a layer of input neurons that send associatively modifiable synaptic connections to a layer of

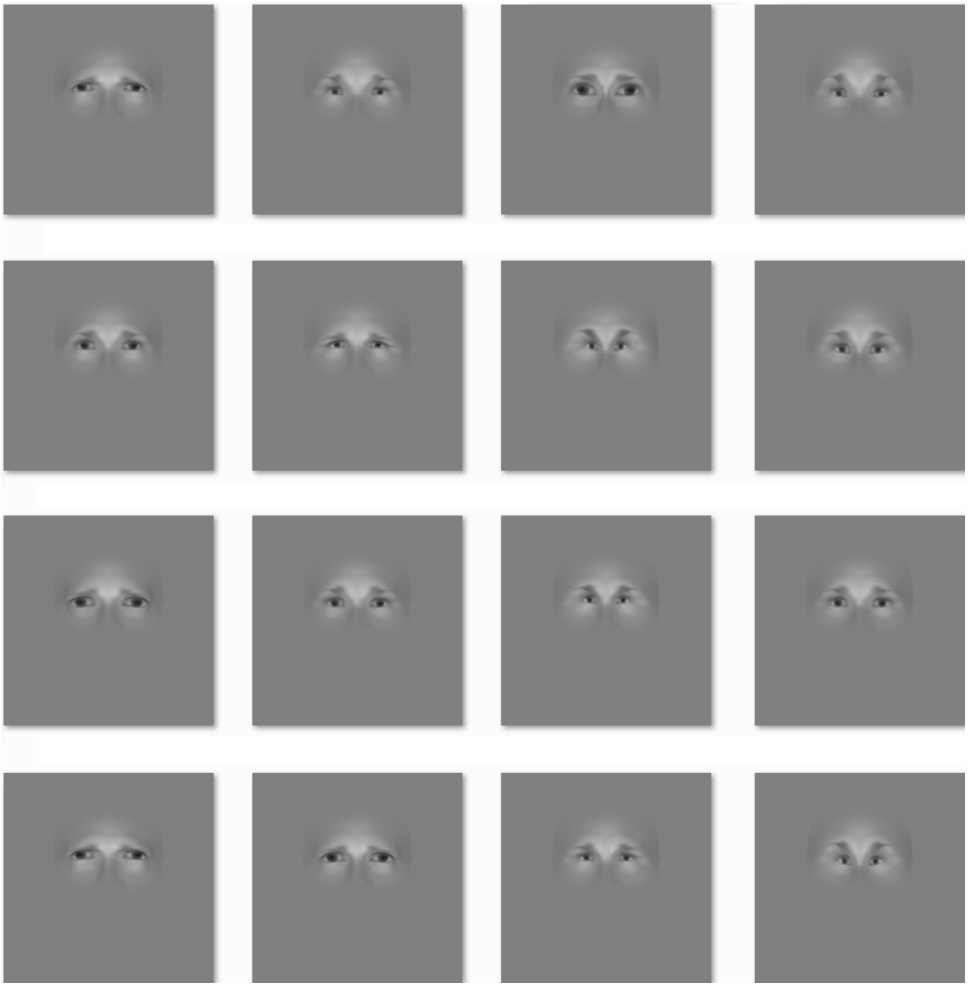

Figure 3.7: Example set of faces used to test VisNet. This set of faces only contained eyes, and was generated by varying the shape of the eyes between faces.

output neurons. The neurons in the output layer compete with each other through inhibitory interneurons to respond to incoming input patterns. Let us identify the neurons in the output layer of this model with a localised sub-population of neurons in the middle face patch, while the input layer contains a localised sub-population of neurons in a preceding cortical area. During learning, the afferent connections to the output neurons are modified by some form of associative learning. The synaptic weight vectors of individual output neurons are also bounded by some form of continual rescaling such as renormalisation, as has been reported in neurophysiology studies (Royer and Paré, 2003). The modification of the afferent connections to the output neurons drives the development of their response properties. These are standard elements of a competitive neural network architecture (Rolls and Treves, 1998). The question is how neurons in the output layer might learn to represent the spatial relationships between facial features with monotonic tuning curves.

Real neurons in the visual cortex of the brain receive afferent connections from a topologically localised region of the preceding cortical layer. Consequently, local sub-populations of neurons within an area such as the middle face patch may receive afferent connections from input neurons representing only a part of a particular feature space such as the distance between the eyes. In this case, the aforementioned middle face patch neurons may receive a full input representation when the eyes are an intermediate distance apart, but only a partial input representation when the eyes are either far apart or very close together. I hypothesised that this boundary effect

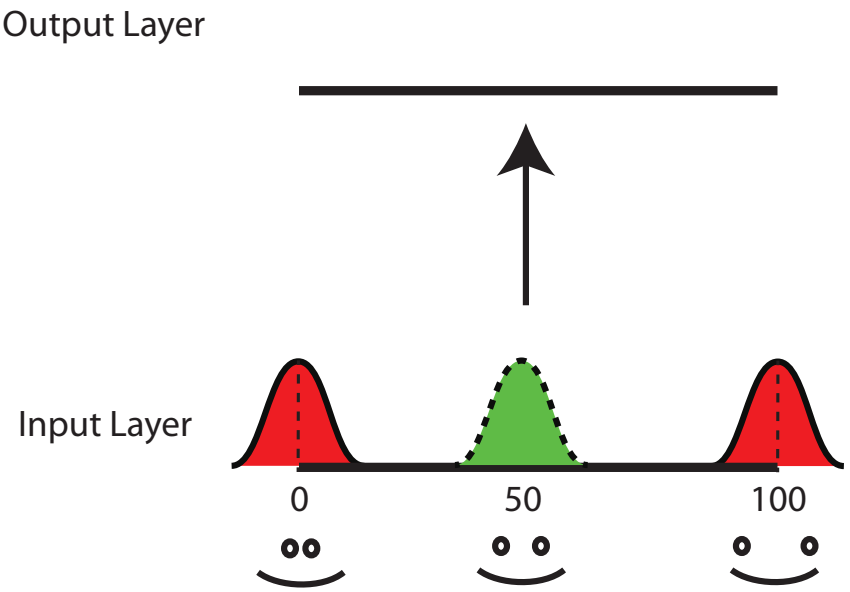

Figure 3.8: Figure showing the architecture of simple one layer network where Gaussian packet of neural activities in the input layer represent a finite 1-dimensional feature space such as the distance between the eyes. The curve filled with green shows the situation where a whole input packet is within the input layer whereas the curves filled with red show the situations where only part of the input packet is within the input layer.

at the extrema of the feature space may drive the development of monotonic tuning curves in the middle face patch neurons. Let us illustrate the argument in the context of the competitive network architecture shown in Figure 3.8 as follows.

Consider the situation where a localised sub-population of output neurons in the middle face patch receives afferent connections from a localised sub-population of input neurons within a preceding cortical area, as shown in Figure 3.8. Assume that the given input layer neurons represent part of a finite 1-dimensional feature space such as the distance between the eyes. In this idealised example, let the position in the space, i.e. the distance between the eyes, be represented by the position of a localised (e.g. Gaussian) packet of neural activity within the input layer, as shown by the green curve in Figure 3.8.

The key aspect of this architecture that drives the development of monotonic tuning curves in the output layer is what happens at the two boundaries of the space, when the eyes are either furthest apart or closest together. Specifically, if at each boundary only part (e.g. half) of the input packet is represented, as shown by the red curve in Figure 3.8, then the output neurons that learn to respond to these particular end locations will end up with relatively large synaptic weights. This is due to these output neurons becoming more tightly tuned to a smaller (end) region of the input layer by (Hebbian) associative learning based on co-activity of the input and output neurons, yet with the magnitudes of the synaptic weight vectors of these output neurons still renormalised over this smaller input region.

If the widths of the input activity packets are relatively broad, then when the input packet is shifted to a more central location within the feature space, the same output neurons, which learned to respond to the end locations, continue to win the competition and respond to the more central locations of the space as well. However, as the input packet shifts away from the ends of the feature space, the responses of these neurons will decline monotonically. Different sub-populations of neurons will learn to respond to the two ends of the feature space, with each sub-population reducing its responses monotonically as the input packet shifts away from its preferred end location.

However, I also hypothesize that the output neurons only develop monotonic tuning curves if the packet of activity in the input layer is wide enough; otherwise, the end effect breaks down

and output neurons develop peaked (e.g. Gaussian) tuning curves. Moreover, if the input space is circular with no end effects, then the output neurons should not develop monotonic tuning curves at all.

The VisNet model architecture shown in Figure 1.1 is designed to mimic these key aspects of cortical architecture that are needed for the development of monotonic tuning curves. The model is comprised of four competitive layers of neurons. Neurons within each layer receive afferent synaptic connections from a topologically corresponding, localised region of the preceding layer. The synaptic weights may be updated by associative (Hebbian) learning rules, with the weight vectors of individual neurons continually renormalised. It was therefore expected that when VisNet is trained on many realistic faces, neurons in the higher layers of model would learn to encode spatial relationships between facial features, such as the distance between the eyes, with monotonic tuning curves.

3.2.3 How some neurons in later stages of visual processing learn to respond to global attributes of faces such as a particular identity or expression

The primate visual system can process global attributes of faces such as identity, emotional expression, age, race, and gender (Homola et al., 2012; Freeman et al., 2010; Morin et al., 2014; Hasselmo et al., 1989a). For example, some neurons in the anterior IT (TE) respond selectively to facial identity, while other neurons in the superior temporal sulcus (STS) respond to facial expression (Hasselmo et al., 1989a; Perrett et al., 1992). The important question is how such selective cell response properties could develop given that the visual system is always exposed to both facial identity and expression simultaneously during early visually-guided learning and self-organisation in the visual system.

In earlier work carried out by Tromans et al. (2011), it was shown that when VisNet was trained on a large number of faces, the higher layers of the model developed neurons that either responded to the identity of a face regardless of its emotional expression, or responded to the facial expression irrespective of facial identity. The hypothesised learning mechanism was as follows. If VisNet is exposed to many possible combinations of facial identity and expression during training, then any particular identity is seen only rarely coupled with a particular expression. This creates a statistical decoupling between any particular facial identity and expression. This, in turn, forces individual neurons in the higher layers to learn to respond to either a particular identity regardless of expression, or particular expression regardless of identity. This was demonstrated successfully when VisNet was trained on a matrix of cartoon images of faces with varying identity and expression.

However, this earlier study used a highly idealised set of cartoon faces with two especially unrealistic properties. First, facial identity and expression were represented by different facial features, and thus had non-overlapping representations on the input layer. In particular, facial identity was represented by variation in the shape of the eyes and nose, while facial expression was represented by changes in the shape of the eyebrows and mouth. Thus, the representations of identity and expression were non-overlapping on the input retina, which is not realistic. With real faces, features such as the eyebrows, eyes, nose, and mouth will all contribute to the representations of both facial identity and expression in a more complex, distributed manner. Accordingly, in this chapter, I investigated whether VisNet will still form neurons that respond to either facial identity or expression even when the network is trained on more realistic faces created using FaceGen, where all of the facial features are involved in representing both facial identity and expression. I hypothesize that this can occur because it is a standard property of competitive networks that under the right conditions, they are able to develop separate (orthogonalised) output representations of distributed (overlapping) input patterns (Rolls and Treves, 1998). These learning mechanisms are demonstrated in simulations presented in the study 3a below.

Secondly, in the study carried out by Tromans et al. (2011), VisNet was trained on every

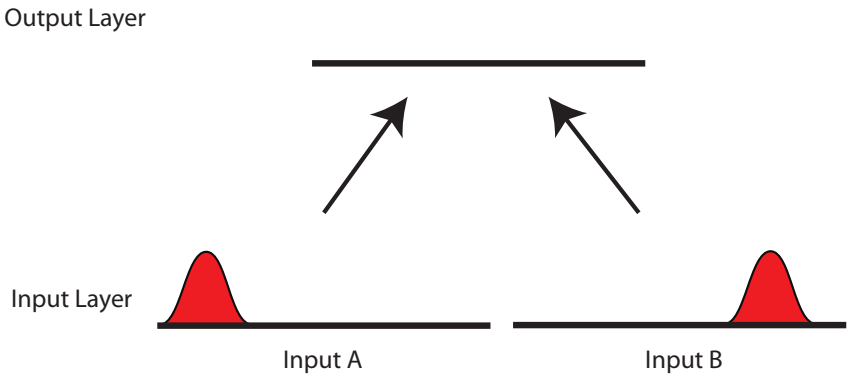

Figure 3.9: The architecture of a competitive neural network where the output neurons receive connections from two distinct populations of input neurons, A and B, which represent two different feature spaces such as facial identity and expression. The degree of overlap between the two input populations was varied across different simulations: 0% overlap, 50 % overlap, and 100 % overlap.

possible combination of facial identity and expression in order to ensure the strongest possible statistical decoupling between these two facial dimensions. This helped to force individual neurons in the higher competitive layers to learn to respond to either a particular identity or expression. However, with real life situations, we do not need to be trained on every possible combination of facial identity and expression in order to learn to recognise these two different facial attributes.

In fact, Tromans (2012) trained VisNet on realistic faces generated using a software FaceGen, the network failed to develop separate representations of facial identity and expression. I hypothesize that this failure was due to the the network being trained on a very dense set of different facial identities and expressions, with the both identity and expression varying almost continuously across their respective dimensions. This rather unnatural set of training faces may have increased the difficulty of neurons in the higher layers developing separate representations of facial identity or expression. In particular, an invariance learning mechanism known as Continuous Transformation (CT) learning (Stringer et al., 2006) may have caused individual neurons in higher layers to learn to respond simply to a large number of gradually changing faces. This is because CT learning is able to bind together smoothly varying input patterns, such as gradually changing faces, onto the same post-synaptic output neuron. In this way, CT learning may have dramatically reduced the selectivity of neurons for particular facial identities or expressions in the study of Tromans (2012).

I propose that this problem can be remedied by training VisNet on a more realistic, reduced set of face images, with only a limited number of different combinations of facial identity and expression chosen randomly during training. This ensures that the training stimuli do not cover a near continuum of every possible facial identity and expression, which should prevent CT learning from operating. This reduced set of training faces is more realistic than that used by Tromans et al. (2011) since real faces do not actually morph between each other very gradually. In the simulations reported in the study 3a below, training VisNet on the reduced set of realistic faces successfully led to the development of neurons that responded selectively to either facial identity or expression.

The mechanisms underpinning the above hypothesis, that competitive learning can map distributed (overlapping) input patterns to separate (orthogonal) output patterns, may be further elucidated by considering the operation of a simplified competitive network comprised of an input layer that sends associatively modifiable synaptic connections to an output layer. Let us assume that each output neuron receives input from two distinct populations of input neurons, A and B, as shown in Figure 3.9. Each input population represents a different finite 1-dimensional bounded feature space, such as identity or expression, where the location in the feature space

is encoded by the position of a Gaussian packet of activity. During training, Gaussian activity packets are shifted through both of the input populations simultaneously.

According to the basic principle of statistical decoupling, we should see the following effects during learning. If the activity patterns in the two input populations transform in lockstep, so that each location in A is paired with the same location in B, then the output neurons should fail to develop separate representations of the two input spaces. However, if the activity patterns in the two input populations vary independently of each other, so that each location in A is paired with many different random locations in B etc, then the output neurons should develop separate representations of the two input spaces.

In simulations described in the study 3b below, the effects of statistical decoupling were initially demonstrated for the case, where the representations of the two input spaces are perfectly orthogonal to each other. That is, the two input populations A and B have no cells in common. This situation is analogous to the past study with cartoon faces where identities and expressions were represented orthogonally by different facial features (Tromans et al., 2011). In this simple case, as expected, independent movement of the activity packets in the two input populations leads to the development of separate representations of the two feature spaces in the output layer.

Then I tested whether the same effect is seen when the two input populations are overlapping, i.e. share a number of neurons in common. I first tried 50% overlap between the two input populations, and then tried 100% overlap where each input neuron is a member of both populations A and B. In both cases, individual output neurons learned to respond to either the A population or B population as long as the activity packets in the two spaces moved independently during training. This result shows how competitive learning can form separate representations of two feature spaces, such as facial identity and expression, even when these two dimensions are represented by a common set of input neurons.

I hypothesise that a similar effect will be seen when the VisNet architecture, which consists of a hierarchy of competitive layers, is trained on realistic FaceGen faces, in which both facial identity and expression are represented in a distributed manner by all of the facial features such as the eyebrows, eyes, nose, and mouth.

3.2.4 What is the connection between neurons representing spatial relationships between facial features and neurons representing global attributes such as facial identity and expression?

Above, we have discussed how some neurophysiology studies have reported the existence of neurons that represent spatial relationships between facial features such as the inter-eye distance or height of the eyes (Freiwald et al., 2009), while other studies have reported neurons that encode global attributes of faces such as facial identity or expression (Hasselmo et al., 1989a; Perrett et al., 1992; Morin et al., 2014). It is reasonable to expect that the representations of global attributes are dependent upon different spatial configurations of facial parts, with different global attributes such as identity and expression influenced by different spatial configurations of facial parts. I now hypothesise that the cells that encode spatial relationships between facial features in fact largely overlap with the cells representing global facial attributes such as identity and emotion. That is, cells with responses that are correlated with particular global attributes of faces, such as a specific identity or expression, may actually be tuned to a particular spatial relationship between certain facial features that are indicative of that global attribute. For example, the responses of cells that are tuned to a specific shape of the mouth might be also correlated with a particular facial expression such as happy. In this situation, the cell might be regarded as contributing to representing both the shape of the mouth and facial expression.

It has long been known that the neural representation of faces in the primate visual system is distributed, with individual faces represented by many neurons and individual neurons participating in the representation of many faces (Rolls and Treves, 1998). The question is whether

Table 3.1: Parameters used for simulations

Parameter	**Value**			
(a) VisNet				
Gabor: Phase shift (ψ)	$0, \pi$			
Gabor: Wavelength(λ)	2			
Gabor: Orientation(θ)	$0, \pi/4, \pi/2, 3\pi/4$			
Gabor: Spatial bandwidth (b)	1.5 octaves			
Gabor: Aspect ratio (γ)	0.5			
No. of Layers	4			
Retina	$256 \times 256 \times 16$			
	1st layer	**2nd layer**	**3rd layer**	**4th layer**
Dimension	128×128	128×128	128×128	128×128
Num. of fan-in connections	201	100	100	100
Fan-in radius	8	8	12	16
Sparseness of activations	2 %	20 %	30 %	30 %
Sigmoid slope (β)	190	40	75	26
Learning rate (k)	1.0	1.0	1.0	1.0
Training Epochs	50	100	100	76
Excitatory Radius (σ_E)	1.4	1.1	0.8	1.2
Excitatory Contrast (δ_E)	5.35	33.15	117.57	120.12
Inhibitory Radius (σ_I)	2.76	5.4	8.0	12.0
Inhibitory Contrast (δ_I)	1.6	1.5	1.6	1.5
(b) Simplified Network				
No. of cells in each layer	100			
Sigmoid slope β	10			
Learning rate k	0.001			
Training epochs	3000			
Sparseness of activations	50 % (simulation 1) and 25 % (simulation 2)			
Inhibitory contrast δ_I	0.01			
Inhibitory radius σ_I	15			
Excitatory contrast δ_E	0.5			
Excitatory radius σ_E	5			

a global attribute, such as a particular identity or expression, is represented by a *random* subset of neurons, or whether individual neurons actually represent specific constituent features of the global attribute. In the latter case, a neuron that represents a particular curvature of the mouth might participate in the representation of a happy expression across a number of different facial identities. This is what I am proposing. In this case, the activity of neurons encoding a particular spatial relationship will also be correlated with a particular corresponding global attribute, and may be thought of as participating in the representation of that global attribute.

It is then possible that individual neurons in even higher layers learn to respond to specific combinations of neurons representing the spatial relationships between facial features that are correlated with a particular global attribute. These higher layer neurons would be tuned to a combination of all the spatial relationships comprising a particular global attribute, and so might be regarded as providing the most abstracted representation of the global attribute, that is, in a way that does not depend on the presence of any one particular spatial relationship between facial features.

3.3 Materials & Methods

3.3.1 Model Descriptions

3.3.1.1 VisNet Model

The main simulation studies presented in this chapter are conducted with an established biologically plausible neural network model, VisNet, of the primate ventral visual pathway, which was originally developed by Wallis and Rolls (1997). The detailed description is provided in Section 1.3. The values used in the current studies are given in Table 3.1(a).

Before the visual images are presented to the VisNet's input layer 1, they are preprocessed by a set of Gabor filters that accord with the general tuning profiles of simple cells in V1 (Jones and Palmer, 1987; Cumming and Parker, 1999; Lades et al., 1993). The filters provide

a unique pattern of filter outputs for each transform of each visual object, which is passed through to the first layer of VisNet. These filters are known to provide a good fit to the firing properties of V1 simple cells, which respond to local oriented bars and edges within the visual field (Jones and Palmer, 1987; Cumming and Parker, 1999). The input filters used are computed by Equations (1.1) and (1.2). In the experiments conducted in this chapter, an array of Gabor filters is generated at each of 256 × 256 retinal locations with the parameters given in Table 3.1(a). These parameters were selected based on those that previously optimised performance (Tromans et al., 2011; Rolls and Milward, 2000). The outputs of the Gabor filters are passed to the neurons in layer 1 of VisNet according to the synaptic connectivity given in Table 3.1(a). That is, each layer 1 neuron receives connections from 201 randomly chosen Gabor filters localised within a topologically corresponding region of the retina.

In this chapter, simulations utilised a self-organising map (SOM) (von der Malsburg, 1973; Kohonen, 1982) implemented within each layer. In the SOM architecture, short-range excitation and long-range inhibition are combined to form a Mexican-hat spatial profile and is constructed as a difference of two Gaussians as described in Section 1.3.3.2. The lateral inhibition and excitation parameters used in the SOM architecture are given in Table 3.1(a), which were selected based on those that previously optimized performance (Rolls, 2000; Tromans et al., 2011).

The parameters for the sigmoid activation function are shown in Table 3.1(a). They are similar to the standard VisNet sigmoid parameter values that were previously optimised to provide reliable performance (Stringer et al., 2006, 2007; Stringer and Rolls, 2008).

During training with visual objects, the strengths of the feed-forward synaptic connections between successive neuronal layers are modified by biologically plausible local learning rules, where the change in the strength of a synapse depends on the current or recent activities of the pre- and post-synaptic neurons. A variety of such learning rules may be implemented with different learning properties as described in Section 1.3.5.

3.3.1.2 Simplified network model

In order to analyse the learning mechanisms in greater detail, some complementary simulations were also carried out within a much simpler competitive neural network architecture with only one layer of fully connected, associatively modifiable synapses as shown in Figure 3.8. The network was trained and tested on 1-dimensional Gaussian input patterns, which provided an idealised representation of a 1-dimensional facial feature space such as the distance between the eyes. This abstracted neural network model allowed a more controlled investigation of the hypothesised mechanisms underpinning the development of the cell response characteristics of interest.

Firing rates of neurons in the input layer The population of input neurons represent the current position x within a 1-dimensional feature space, such as the distance between the eyes. Each input neuron j is set to respond maximally to a unique position x_j in the feature space. The firing rate r_j of each input neuron j is determined by a Gaussian distribution positioned at x_j with standard deviation σ as follows:

$$r_j = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-x_j)^2}{2\sigma^2}} \tag{3.1}$$

The representation across the population of input cells thus takes the form of a Gaussian packet of activity centred on the current position x within the feature space.

Activations of neurons and competition within the network Within the output layer of the network, the activation h_i of each output neuron i is set equal to a linear sum of the

inputs r_j from afferent neurons j in the preceding input layer weighted by the synaptic weights w_{ij} as shown in equation (1.3).

Next, lateral competition is applied between neurons in the output layer. In the simulations reported below, competition is implemented in one of two possible ways as follows.

In the first competition method, divisive inhibition, the activation h_i of each output neuron i is divided by the average activation $< h >$ across all output neurons.

The second type of competition method implements a combination of lateral inhibition and excitation between cells in the output layer in order to effect a self-organising map (SOM). Short-range excitation and long-range inhibition are combined to form a 'Mexican-hat' spatial filter, which is constructed as a difference of two Gaussians as follows:

$$I_a = -\delta_I \exp\left(-\frac{a^2}{\sigma_I^2}\right) + \delta_E \exp\left(-\frac{a^2}{\sigma_E^2}\right) \tag{3.2}$$

To implement the SOM, the activations h_i of neurons within the output layer are convolved with the spatial filter, I_a, where δ_I controls the inhibitory contrast and δ_E controls the excitatory contrast. The width of the inhibitory radius is controlled by σ_I while the width of the excitatory radius is controlled by σ_E. The parameter a indexes the distance away from the centre of the filter. The lateral inhibition and excitation parameters are given in Table 3.1(b).

Next, the contrast between the activities of neurons with the output layer is enhanced by passing the activation h_i of each output neuron, i, through a sigmoid transfer function as shown in equation (1.6). The sigmoid slope β is set to a fixed value throughout each simulation given in Table 3.1(b). However, the sigmoid threshold α is continually adjusted to control the sparseness of the firing-rates within the output layer.

Modification of synaptic weights during training At the beginning of each simulation, the synaptic weights from the input neurons to the output neurons are initialized with random values.

The simulation begins with a training phase. At each timestep τ during training, the firing rates of the input neurons are first updated to represent a new position x in the feature space, and then the firing rates of the output neurons are computed as described above. Then the synaptic weights are updated according to an associative (Hebbian) learning rule as described in equation (1.7)

To prevent the same few output neurons always winning the competition, the synaptic weight vector $\mathbf{w}_i$ of each output neuron i is renormalised after each learning update by equation (1.10) and (1.11).

3.3.2 Information Analysis

To quantify the performance in transformation invariance learning with VisNet, the techniques of Shannon's information theory have previously been used (Rolls and Treves, 1998). Information theory can be used to quantify how selective neurons are for particular stimuli, each of which may translate across different locations on the retina (see Section 1.5.2).

Two information measures were used to assess the ability of the network to develop neurons that are selective to the presence of stimuli but also invariant to their occurrence in different retinal locations (see Rolls et al. (1997); Rolls and Milward (2000)). These two measure use the responses from either individual neurons (single-cell information analysis as described in Section 1.5.2.1) or small ensembles of neurons (multiple-cell information analysis as described in Section 1.5.2.2).

3.4 Simulation Studies

In this chapter, three simulation studies using VisNet were carried out. As described earlier in Section 1.3, the standard network architecture is shown in Figure 1.1. The VisNet model was trained on realistic images of faces with different identities and expressions generated using the FaceGen face modelling software (FaceGen, 2013). FaceGen builds artificial 3D face images from templates taken from 273 high resolution 3D face scans. The images are averaged, and PCA is used to extract a set of variances from the mean representing facial features such as shape, colour and gender. This in turn gives a normal distribution from which a random coefficient can be chosen, creating a random, realistic face based on a range of alterable features. After training, I investigated whether the neurons in the higher layers of the network had developed response characteristics similar to those reported in neurophysiology studies.

In the series of studies conducted in this chapter, I tested whether VisNet developed neurons with response characteristics similar to what has been found in neurophysiology experiments. In the first study, I explored how neurons learn to respond to individual local facial features such as the facial outline, eyes, nose, and mouth, as well as specific global combinations of these features. The second study investigated how some neurons learn to represent the spatial relationships between particular facial features, such as the distance between the eyes, with monotonic tuning curves. Finally, in the third study, I explored how some neurons learn to represent the global attributes of either facial identity or facial expression.

However, unless otherwise stated, these VisNet studies were carried out by testing the same trained network. That is, the VisNet model was trained only once at the beginning, and then the same trained network was tested across all three studies for the various cell response properties. There is one exception to this in the first study, where the network is retrained to investigate how increasing the variation in facial features, such as the eyes, nose, and mouth, drives the development of neurons that respond to the individual features.

During the initial training of VisNet, the network was presented with 450 realistic human faces as shown in Figure 3.10 and 150 non-face objects as shown in Figure 3.11. The faces were randomly generated with different identities using the commercial software FaceGen, and the expressions of individual faces were also randomly set along a continuous dimension between happy and sad. Non-face objects were retrieved from Google 3D warehouse. All stimuli were grayscaled and projected onto an input retina that was 256×256 pixels in size.

In the second and third studies, some complementary simulations with the simplified network model with only one layer of fully connected, associatively modifiable synapses as shown in Figure 3.8 were carried out. The network was trained and tested on 1-dimensional Gaussian input patterns, which provided an idealised representation of a 1-dimensional facial feature space such as the distance between the eyes. This abstracted neural network model allowed a more controlled investigation of the hypothesised mechanisms underpinning the development of the cell response characteristics of interest. The parameters used in the simulations in this chapter are shown in Table 3.1(b). Full details of the network architecture are provided in Section 3.3.1.2. The purpose of these additional simulations was to investigate deeper into the underlying learning mechanisms using a more simplified and controlled setup.

3.4.1 Study 1: The neural representation of local facial features and combinations of features

3.4.1.1 a. Simulation results of VisNet

Freiwald et al. (2009) showed that cells in the middle face patch of the primate visual system responded selectively to individual facial features or particular combinations of features. Therefore, in this first study, I investigated the neural representation of individual local facial features, such as the facial outline, eyes, nose, and mouth, as well as global combinations of

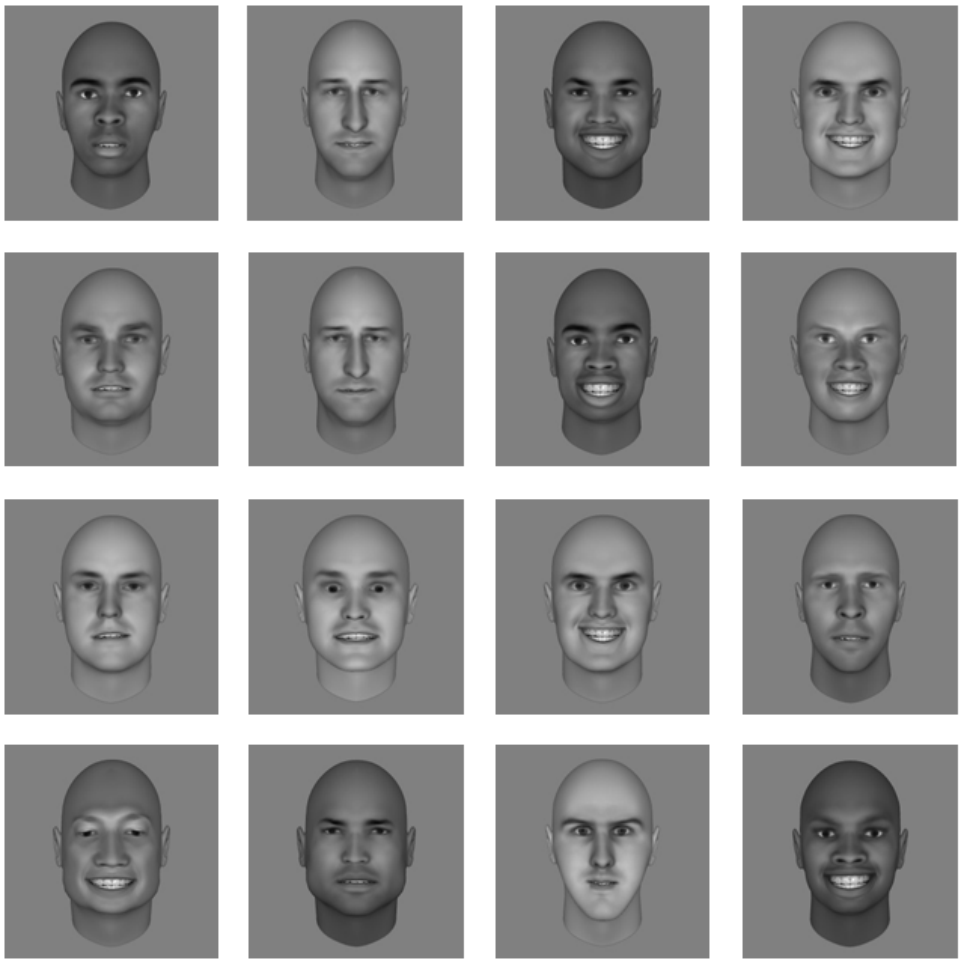

Figure 3.10: Examples of randomly generated faces with different identities used for training the VisNet network. These stimuli were generated using the commercial software FaceGen. The facial expression of each face was also randomly set along a continuous dimension between happy and sad.

these features, throughout the hierarchical architecture of VisNet. Specifically, it was explored whether such neuronal responses had developed in VisNet during the initial training on 450 realistic human faces as shown in Figure 3.10 and 150 non-face objects as shown in Figure 3.11.

The neurophysiology study carried out by Freiwald et al. (2009) showed that cells in the middle face patch were tuned to different combinations of facial features. For example, some neurons were tuned to the presence of only one particular facial feature, while other cells were tuned to a particular combination of either 2, 3, or 4 facial features. Therefore, to investigate whether similar cells had developed in VisNet, the network was tested on face stimuli that were comprised of all possible combinations of the four facial features: facial outline, eyes, nose, and mouth. For each possible combination of facial features, 15 different facial identities were created in order to test for generalisation across different facial identities. A subset of the face stimuli used for testing the network is shown in Figure 3.12.

Similar to the results reported by Freiwald et al. (2009), neurons were found to have learned to respond to different combinations of the facial features. Figure 3.13 shows the firing rate responses of five different 4th layer neurons to face stimuli constructed from different combinations of facial features for 15 distinct facial identities. For example, the first cell (113,1) shown in the figure is more likely to be activated when the facial outline is present. The second cell (82,67) responds strongly whenever the mouth is present. The third cell (80,61) responds when the facial outline and eyes are presented together. The fourth cell (102,62) responds most

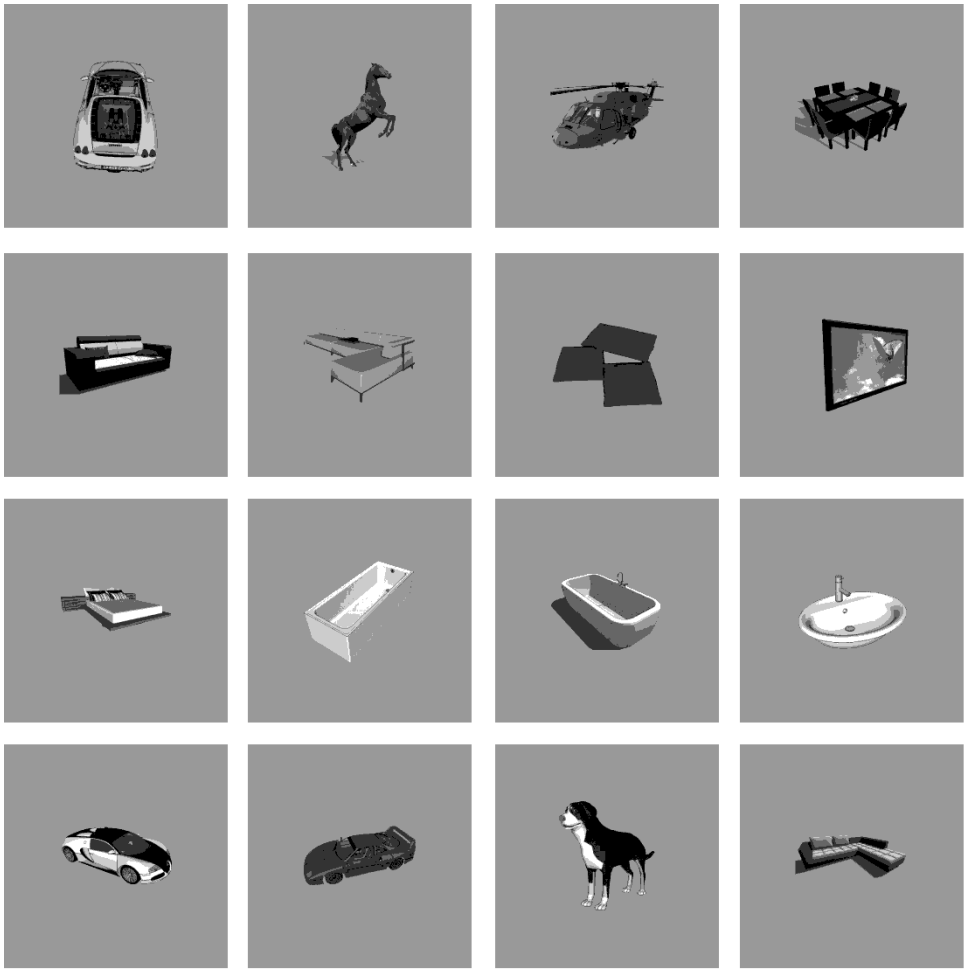

Figure 3.11: Examples of non-face objects used for training the VisNet network. Original images were retrieved from google 3D warehouse and then rescaled and grayscaled.

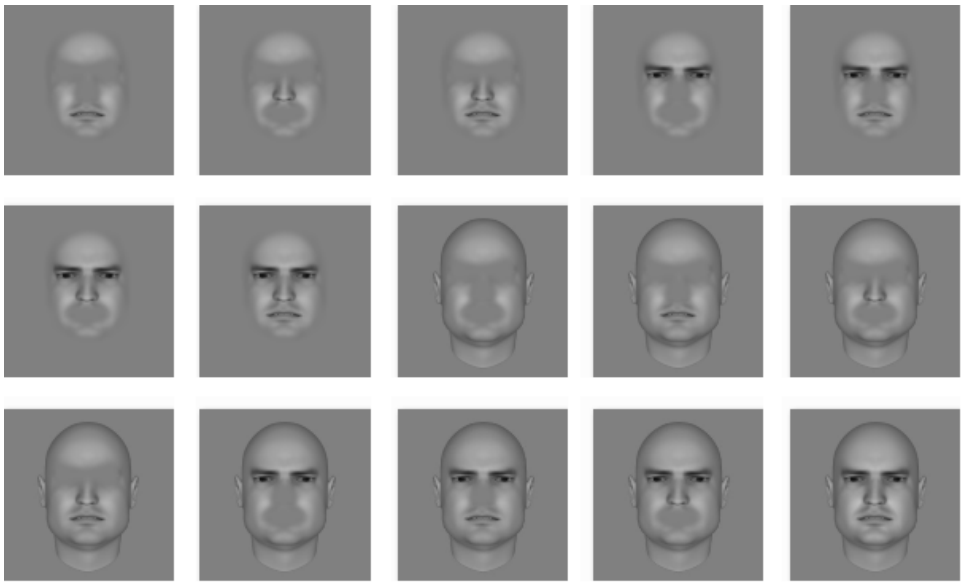

Figure 3.12: Examples of test faces which are composed of a different combination of the four facial features: facial outline, eyes, nose, and mouth. Different faces may have either 1, 2, 3, or 4 of these features present. The purpose of these face stimuli is to test for the existence of neurons that have learned to respond selectively to particular subsets of these facial features.

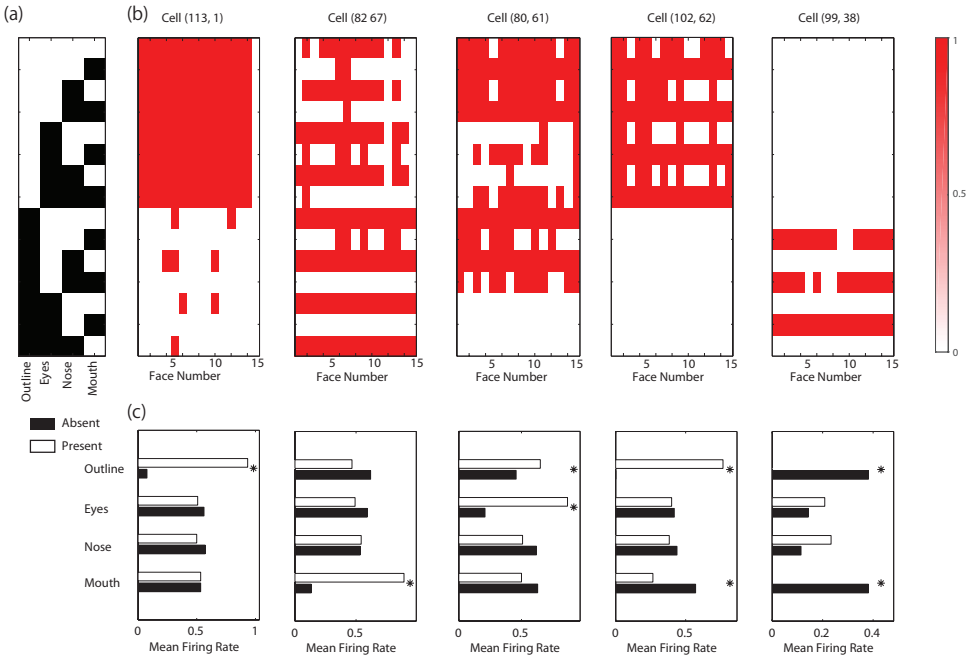

Figure 3.13: Simulation results showing the presence of 4th layer neurons that have learned to respond selectively to particular combinations of the four facial features: facial outline, eyes, nose, and mouth. The network was tested with faces constructed from all possible combinations of the four facial features. (a) The top left subplot shows the different combinations of facial features used to test the network, where each row corresponds to a different combination of facial features. (b) The five subplots on the top right show the responses of five different 4th layer neurons to faces constructed from different combinations of features. Each row corresponds to a different combination of facial features defined by the top left plot, and each column corresponds to a different facial identity. (c) The five subplots on the bottom show the average responses of the same 4th layer neurons to face stimuli with a given facial feature (white bars) and without the facial feature (black bars). Based on paired t-test, * indicates significant excitatory modulation of the neuronal responses by a particular facial feature ($P < 0.005$). For example, cell (113, 1) fired significantly more strongly when the facial outline was present ($P < 0.005$).

strongly when the facial outline is present but the mouth is absent. The fifth cell (99,38) responds most when the facial outline and mouth are absent. Similar cell selectivities were also found by Freiwald et al. (2009).

Figure 3.14 shows the number of 4th layer neurons that responded significantly more strongly ($P < 0.005$) to the presence or absence of a particular number (1, 2, 3, or 4) of facial features before and after training based on paired t-test over identities. The results confirm that, after training, different neurons responded maximally to different numbers of the facial features. Some cells were tuned to only a single facial feature, while other cells responded most to either 2, 3, or 4 facial features.

However, the figure still shows that many cells before training also seem to show the selectivity to a particular combination of facial features. This can be explained with the similar reason to the result reported in Section 2.4.1.5 about the unexpectedly large number of cells that exhibited a selectivity to a particular combination of multiple local contour elements. Basically due to simply the location specific stimulation in the input rather than the actual shape information, even a number of cells in an untrained network could happen to show the seemingly appropriate firing properties. This could have been investigated in more detail, the result reported in this simulation study at least shows that the training improved the performance significantly in terms of statistics, which serves the purpose for the aim of this particular study.

In order to further quantify the selectivity of neurons to individual face parts in the successive layers, single cell information analysis was conducted as described in Section 3.3.2. Figure 3.17 shows the single cell information plots for each layer of VisNet in which the testing stimuli are four different face parts (mouth, nose, eyes, and outline) for 15 distinct facial identities shown

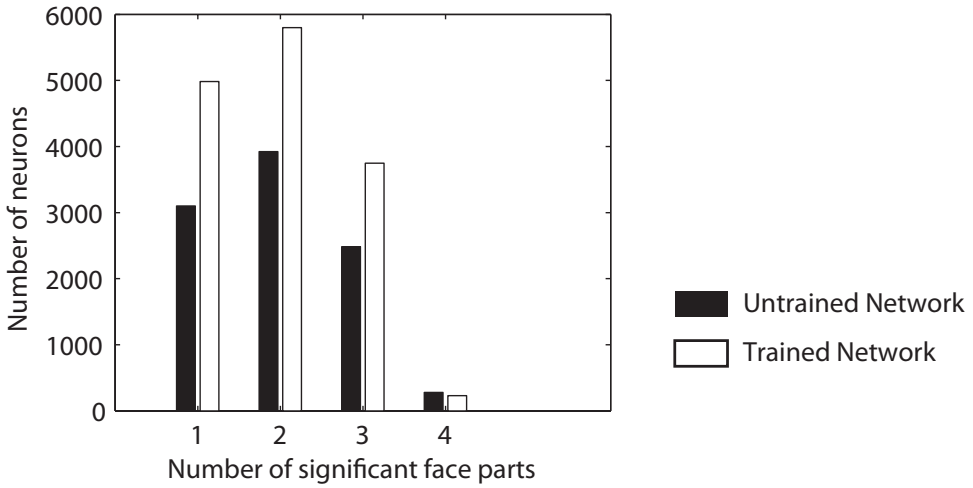

Figure 3.14: Frequency histogram showing the number of 4th layer neurons that responded significantly more strongly ($P < 0.005$) to the presence rather than absence of a particular number (1, 2, 3, or 4) of facial features. The network was tested as follows For each of the four facial features (outline, eyes, nose and mouth), the average response of each 4th layer neuron to (i) all possible faces that contain the feature and (ii) all possible faces that omit that feature were recorded. Then I computed for each neuron whether it responded significantly more to the presence rather than absence of that feature. This was done using a paired t-test. This procedure was repeated for all four facial features. Then, for each neuron, the number of facial features for which the neuron responded significantly more to the presence rather than absence of that feature was recorded. The histogram shows the number of neurons that responded more strongly to either 1, 2, 3 or 4 facial features. Results are shown before and after training.

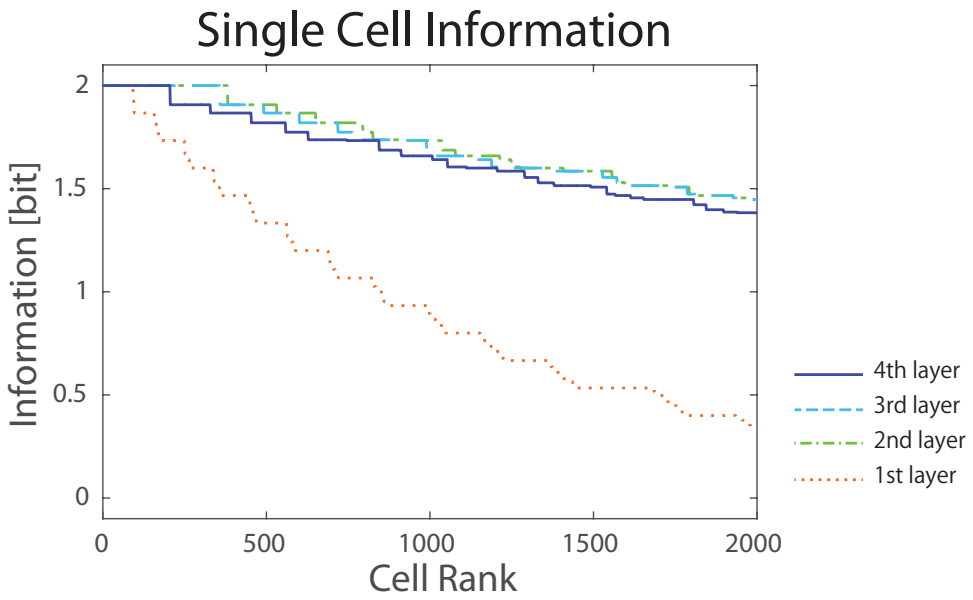

Figure 3.15: Information analysis of selectivity of neurons to specific face parts. VisNet was trained on 450 realistic human faces as shown in Figure 3.10 and 150 non-face objects as shown in Figure 3.11 and then tested on four different face parts (mouth, nose, eyes, and outline) for 15 distinct facial identities shown in Figure 3.12. In this analysis, I tested whether individual cells had learned to respond selectively to the presence of a particular face part. To do this, the amount of single cell information carried by cells about whether one of the four face parts was present in the test image was measured. The figure shows single cell information plots for different layers of VisNet: 1st layer (dotted line), 2nd layer (dash-dot line), 3rd layer (dashed line), and 4th layer (solid line). Since there are four different face parts, the maximum amount of information possible is $log_2(4)$, that is 2 bits. The results show that the number of cells that reached maximum single cell information is small in the first layer, is largest in the second and third layers, and then declines slightly in the fourth layer. This may imply the more holistic representations in higher layers.

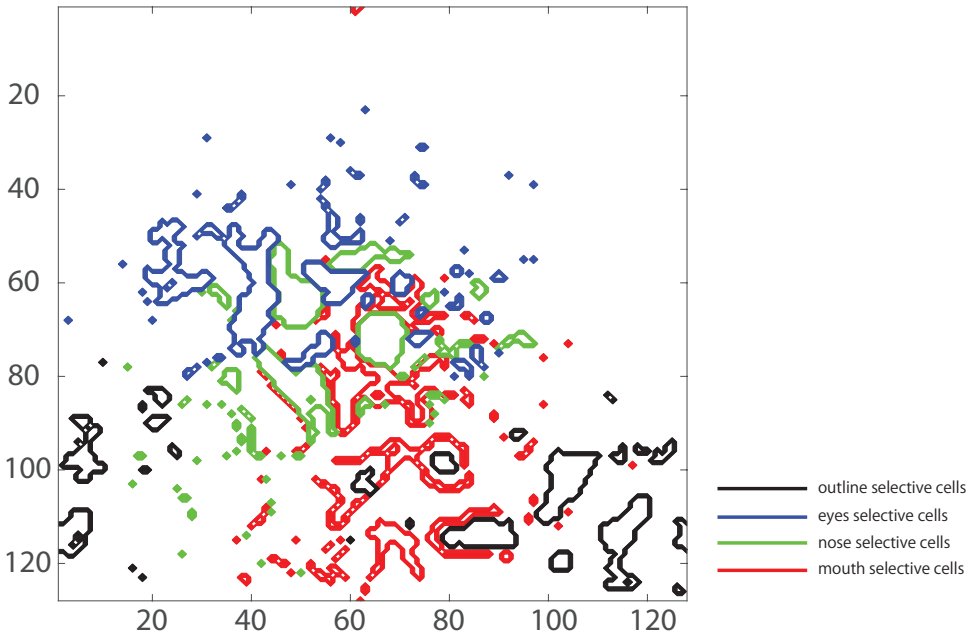

Figure 3.16: Map showing face feature selectivity of all 4th layer neurons to the mouth (red), nose (green), eyes (blue), and outline (black) features shown in Figure 3.12. The selectivity measure was computed based on single cell information analysis, and the 500 cells that carry the highest information for each facial feature are presented.

in Figure 3.12. The number of cells that reached maximum single cell information is small in the first layer, is the largest in the second and the third layers, but then declines slightly in the fourth layer. These simulation results reflect the hierarchical representation of faces in the primate visual system discussed in the introduction. Specifically, the simulation results mirror how the occipital face area (OFA) in an early stage of processing learns to respond to individual facial features, while the fusiform face area (FFA) in a later stage of processing subsequently integrates this information.

In addition, the 4th layer cells that carry highest single cell information were mapped for each facial feature to explore whether "facitopy" has been developed in our simulation as reported in the fMRI study of Henriksson et al. (2015). In their study, the cortical representations of facial features such as the eyes, nose and mouth were found to be arranged in a map that corresponded to their relative positions within the face. The contour plots shown in Figure 3.16 indicate each sub-region comprised of the top 500 cells that carry the highest single cell information for one of the four facial features: mouth (red), nose (green), eyes (blue), and outline (black). Consistent with the facitopy hypothesis, the distribution of the sub-regions are found to be roughly corresponding to the physical configurations of facial features within the face.

The above results show that along the hierarchy, the network develops separate representations of the local facial features such as the facial outline, eyes, nose, and mouth. However, the question is how the network learns to represent the individual facial features when the network is always exposed to complete faces comprised of all the facial features presented together during training. Since it was identified that the number of face feature selective cells is greater in the intermediate layers (i.e. 2 and 3) than in the output (4th) layer, in the following subsection, the focus of the analysis was moved on to underpin the learning mechanisms of feature selective neurons on the third layer of the network.

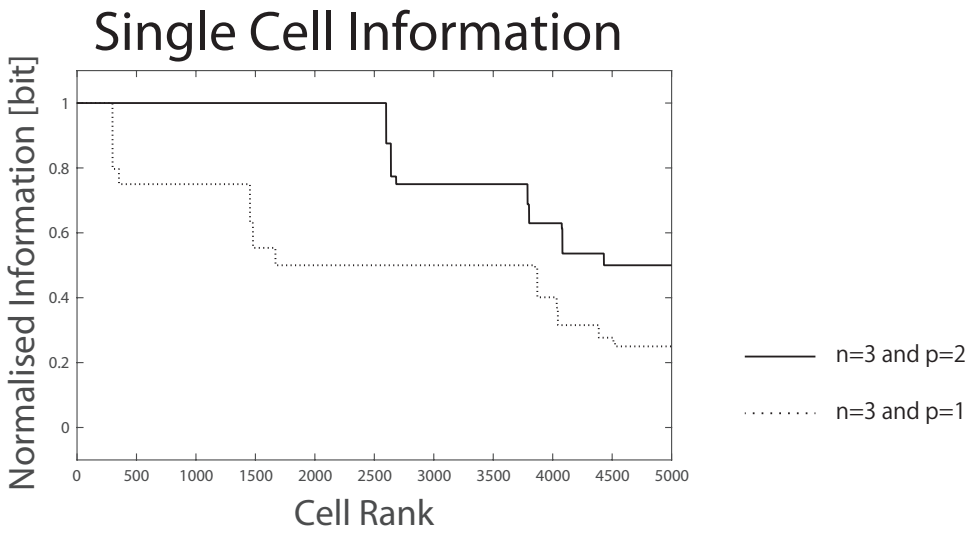

Figure 3.17: Results of simulations in which VisNet was trained on facial stimuli that were constructed by varying p, the number of possible shapes of each of the facial features. The face stimuli had $n = 3$ facial features, eyes, mouth, and outline, each of which was varied over p different shapes during training. (An example set of training stimuli where $p = 2$ is shown in Figure 3.4.) There were two separate simulations in which the training stimulus set had p equal to either 1 (dotted line) or 2 (solid line). In both simulations, the network was tested on the set of stimuli constructed by extracting only one facial feature from the facial stimuli used during training, as shown in Figure 3.5 for $p = 2$. The figure shows single cell information plots for the two separate simulations with p equal to 1 or 2. Each plot shows the normalized single cell information carried by each of the 3rd layer neurons (in rank order) about the presence of one of the $n \times p$ facial features. The maximum amount of information possible for the two simulations is $log_2(n \times p)$, that is 1.6 and 2.6 bits for simulations with p equal to 1 or 2 respectively. The results show that the number of cells that reached maximum single cell information increases as p increases from 1 to 2. This supports the hypothesis that the increased statistical decoupling between facial features across multiple faces as p increases from 1 to 2 forces neurons to learn to become more selective to particular facial features.

3.4.1.2 b. How the network learns to represent individual facial features through competitive learning driven by statistical decoupling between the features

Shape selective facial feature representations In Section 3.2, I hypothesised that some neurons would become tuned to particular facial features due to the statistical decoupling between any two of these features as the number of shape variations increases. Specifically, eyes of a particular shape and a particular shaped mouth would be seen together only rarely. This creates a statistical decoupling between these two particular features, which in turn makes it difficult for neurons to learn to respond to this particular combination. In order to carry out a controlled test of this hypothesis, two simulations in which VisNet was trained on faces with $n = 3$ variable facial features (eyes, mouth, and facial outline) were conducted. In the simulations, the number of shape variations of each of these facial features p was set to either 1 or 2. Accordingly, the number of distinct shapes of facial features to be learned is 3 and 6 ($n \times p$), and the number of whole faces presented during training is 1 and 8 ($p^{(n)}$), respectively. Additionally, in order to eliminate the cells that happen to exhibit facial feature selectivity due to the topologically distributed feedforward synaptic connections in the model, each face was presented in a 2×2 grid of 4 different retinal locations, which were separated by horizontal and vertical shifts of 10 pixels. For each simulation, after training using the temporal trace learning rule as described in equations (1.8) and (1.9) in Section 1.3.5.2, the network was tested with the set of face stimuli constructed by extracting just one of the three facial features as shown in Figure 3.5 for $p = 2$.

In order to quantify the performance, single cell information analysis was conducted as described in Section 1.5.2.1. Figure 3.17 shows normalized single cell information plots for two simulations in which the training stimuli were constructed with p set to either 1 or 2 shapes for each of $n = 3$ facial features, eyes, mouth, and outline. Each plot shows the information carried

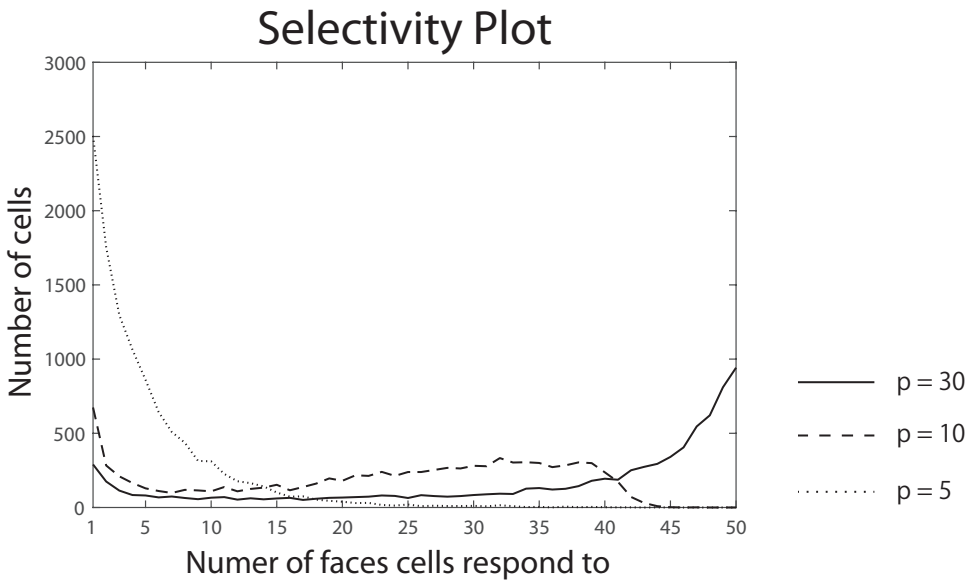

Figure 3.18: Results of simulations in which VisNet was trained on facial stimuli that were constructed by varying the number of possible shapes of the eyes as shown in Figure 3.6, and tested on facial feature stimuli that were constructed by extracting just the eyes of novel faces as shown in Figure 3.7. The plot shows how, as the network is exposed to more shapes of eyes during training, many cells start to exhibit shape invariant selectivity to the eyes.

by all of the 3rd layer neurons about a specific shape of one of the three facial features, where the neurons are plotted in rank order along the abscissa. The maximum amount of information possible for the simulations is $log_2(n \times p)$, that is 1.6 or 2.6 bits for $p = 1$ or 2 respectively. The result shows that the number of cells that learned to carry maximum single cell information increased as p was increased from 1 to 2. Thus, for the higher value of $p = 2$, neurons learned to be more selectively tuned to a specific facial feature shape, which was due to statistical decoupling between different shaped features across multiple faces.

Shape invariant facial feature representations At the same time, I hypothesized that as p increases, then CT learning will begin to bind together the different shapes of a particular facial feature leading to different subset of neurons that respond to all possible shapes of that feature. In particular, if the network is exposed to many different shapes of eyes covering a near continuum of gradually changing eyes, then the CT learning mechanism may bind together the different shapes of eyes onto the same subset of output cells, which would then respond to all eyes. Something similar would occur for the other facial features such as the facial outline and mouth. The end result of these learning mechanisms should be that as the number of shape variations p for each facial feature increases, more cells should learn to respond selectively to all possible shape variations of just one particular facial feature.

In order to confirm our hypothesis that CT learning was beginning to bind together the shape variations of each particular facial feature as p increased, another series of simulations was conducted. For each of three further simulations, the network was exposed to larger numbers of faces where the shape of eyes was varied over 5, 10, or 30 shapes during training as shown in Figure 3.6. Again, in order to eliminate the cells that happen to be exclusively responding to a particular facial feature due to the topologically distributed feed-forward synaptic connectivities, the faces were shifted across four different retinal locations during learning. After training using the temporal trace learning rule (1.8) and (1.9) described in Section 1.3.5, the network was tested with 50 eyes which were extracted from randomly generated faces as shown in Figure 3.7. Figure 3.18 shows the distribution of the number of cells that respond to different numbers of the shapes of eyes. The result when the network was trained with 5 faces is plotted with a

dotted line, the results with 10 faces is plotted with a dashed line, and the results with 30 faces is plotted with a solid line. It can be seen that over the three simulations, for larger values of p, the number of cells that have learned to respond to most of the shape variations of eyes increases. This confirms that as the number of shape variations of a particular facial feature increases, CT learning binds together these shape variations to produce neurons that respond selectively to one particular facial feature over all possible shapes.

In conclusion, the above simulations show how the network is able to develop neurons that respond to just one particular shape of a facial feature, or particular combinations of facial features, through the statistical decoupling that occurs between facial features when the network is presented with many different faces during training. Moreover, I showed how some neurons may learn to respond invariantly to all the different shape variations of a particular facial feature through continuous transformation (CT) learning when the number of feature shape variations across different faces is large.

Given these representations of individual facial features within the visual hierarchy, I conjectured that neurons in higher layers would start to process these representations and consequently develop various other related response properties. In particular, representations of individual shapes of local facial features could contribute to the development of representations of the spatial relationships between these facial features (Freiwald et al., 2009) as well as the global properties of faces such as identity and expression (Hasselmo et al., 1989a; Morin et al., 2014). At the same time, a collection of shape invariant representations of individual facial features may contribute to the development of global representations of whole faces.

3.4.2 Study 2: The representation of spatial relationships between facial features with monotonic tuning curves

3.4.2.1 a. Simulation results of VisNet

Freiwald et al. (2009) showed that some neurons in the middle face patch of the primate visual system encoded the spatial relationships between facial features, such as the distance between the eyes, with monotonic tuning profiles. We, therefore, tested whether such neurons had developed in VisNet during the initial training on 450 realistic human faces as shown in Figure 3.10 and 150 non-face objects as shown in Figure 3.11. For this purpose, a set of test face stimuli, in which the geometrical parameters of the facial features were systematically varied to be comparable with the physiological study conducted by Freiwald et al. (2009), was constructed. In particular, the dimensions of inter-eye distance, eye-brow angle, eye-height, and mouth shape were varied as shown in Figure 3.19. For each such dimension of spatial variation, faces with five different identities were used. And for each facial identity, ten face images were constructed by sampling ten different, evenly-spaced feature values of the relevant dimension. The ten selected feature values spanned the entire range of realistic values for that dimension. These face stimuli were presented to VisNet during testing, and the firing rate of each neuron in the network was recorded.

Figure 3.20 shows eight example neurons (a-h) found in the 4th layer of VisNet which represent different spatial relationships between facial features with monotonic tuning profiles. Neurons a and b encode inter-eye distance, neurons c and d encode eyebrow angle, neurons e and f encode eye height, and neurons g and h encode mouth shape. Visual inspection of the firing rate responses across the 4th layer confirmed that many neurons had developed monotonic tuning responses to variation in these four spatial relationships between facial features.

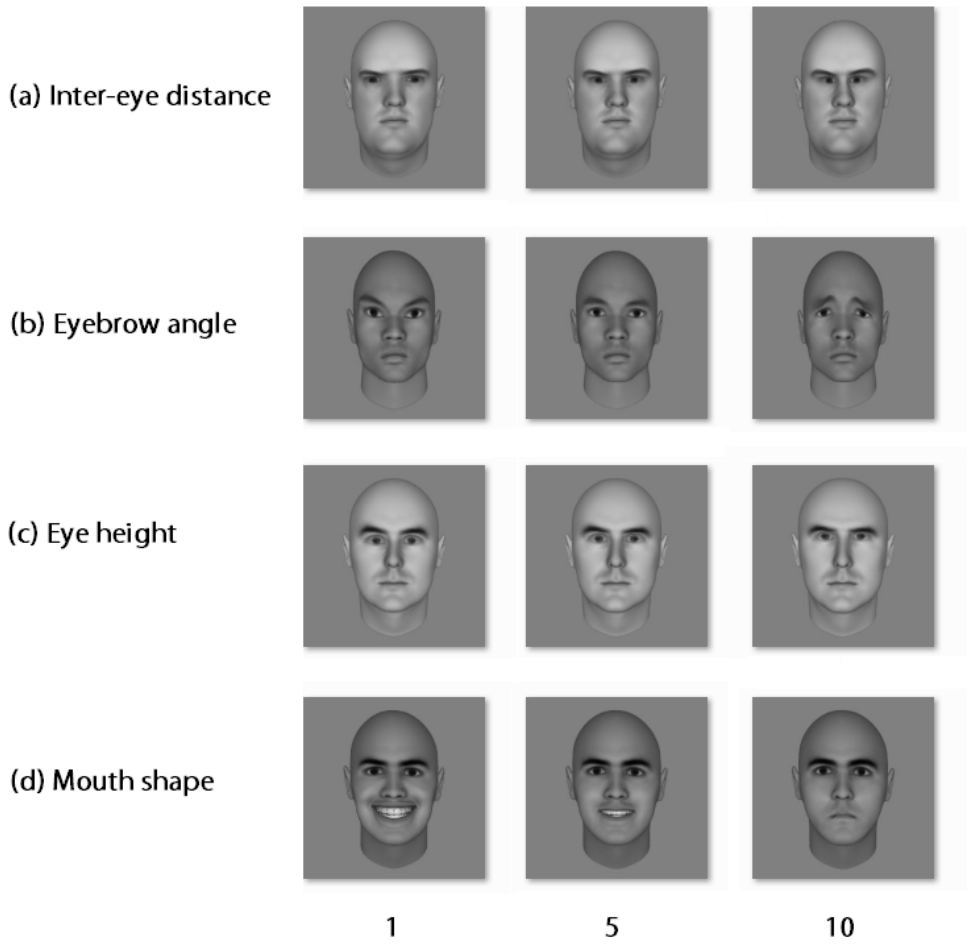

Figure 3.19: Examples of face stimuli used to test VisNet for the presence of neurons that had learned to represent the spatial relationships between facial features with monotonic tuning curves. Four different spatial relationships between facial features were varied during testing: inter-eye distance, eye-brow angle, eye-height and mouth shape.

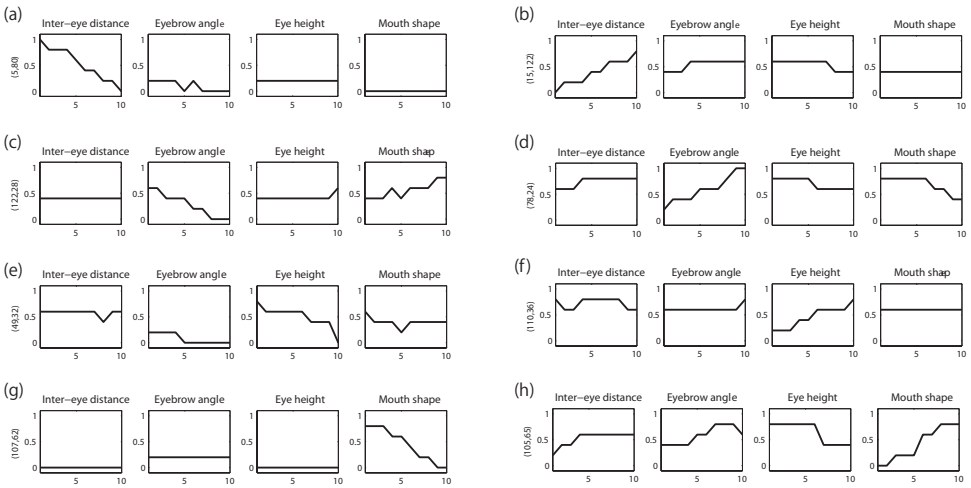

Figure 3.20: Firing rate responses of eight example neurons found in the 4th layer of VisNet which represent different spatial relationships between facial features with monotonic tuning profiles. The network was tested on the face stimuli shown in Figure 3.19. Each row shows the responses of a different neuron (a-h), while each column corresponds to a different kind of spatial relationship: inter-eye distance, eyebrow angle, eye height and mouth shape. The individual subplots show the firing rate responses of the neuron as the corresponding spatial relationship is varied across ten selected feature values. It is evident that the cells are tuned monotonically to different spatial relationships between the facial features: (a,b) inter-eye distance, (c,d) eyebrow angle, (e,f) eye height, and (g,h) mouth shape.

3.4.2.2 b. Simulation results of the simplified network model with one layer of synapses

In order to carry out a deeper investigation into the learning mechanisms by which neurons could develop monotonic tuning responses encoding the spatial relationships between facial features, further simulations in a simplified neural network architecture with one layer of synapses were carried out as described in Section 3.3.1.2 and shown in Figure 3.8. The network was trained and tested on 1-dimensional Gaussian input patterns, which provided an idealised representation of a 1-dimensional facial feature space such as the distance between the eyes. During training, a Gaussian packet of activity is imposed at a series of randomly selected locations on the input layer. At each location of the Gaussian input packet, activity is propagated to the output neurons, and then the synaptic weights are modified using a local associative (Hebbian) learning rule with synaptic weight vector normalisation as described in Section 3.3.1.2. The sigma value that controls the width of the Gaussian input packet, σ, was set to 10 unless otherwise stated. During the testing, the location of the Gaussian packet was moved from neurons 1 to 100 across the input layer, and the firing responses of the output neurons were recorded for each location. This abstracted neural network model allowed a more controlled investigation of the learning mechanisms responsible for the development of monotonically tuned responses among the output cells.

Figure 3.21 shows the development of monotonic tuning responses in the simplified network model with one layer of synapses. The figure shows results for three different simulations: (i) network trained with circularly arranged input neurons with wrap-around (top row), (ii) network trained with linearly arranged input neurons with no wrap-around (divisive inhibition) (middle row), and (iii) network trained with linearly arranged input neurons with no wrap-around (combined lateral inhibition and excitation) (bottom row). The columns show the following: (a) the width, σ, of the Gaussian activity packet imposed on the input layer during training and testing, (b) matrix of synaptic weights from input neurons to output neurons, (c) matrix showing activations of output neurons as a Gaussian activity packet is shifted through successive locations on the input layer, and (d) matrix showing firing rates of output neurons as a Gaussian activity packet is shifted through the input layer. The plots show that, regardless of the type of competition implemented, the trained networks with linearly arranged input neurons with no wrap-around (middle and bottom rows) have developed output neurons with monotonic tuning responses to the location of the Gaussian activity packet in the input layer. On the other hand, output neurons did not show monotonic responses in the network trained with circularly arranged input neurons with wrap-around. In such a circular network there are no such end effects on learning, which are needed to drive the development of output neurons with monotonically tuned responses.

Figure 3.22 shows further results for the three simulations shown in Figure 3.21. For each of these simulations, Figure 3.22 shows the behaviour of twelve typical output neurons in separate subplots. In particular, those cells are the cells indexed with 1, 10, 19, 28, 37, 46, 55, 64, 73, 82, 91, and 100 in the output layer. Each subplot shows how the activation and firing rate of the neuron vary as a Gaussian activity packet is shifted through the input layer. Figure 3.22 confirms that regardless of the type of competition, the trained network with linearly arranged input neurons with no wrap-around displays output neurons with responses that are monotonically tuned to the location of the Gaussian activity packet in the input layer. Moreover, some output neurons respond maximally when the Gaussian activity packet is presented at the left end of the input feature space, and their responses decline monotonically as the packet is shifted to the right. While other output neurons respond maximally when the Gaussian activity packet is presented at the right end of the input feature space, and their responses decline monotonically as the packet is shifted to the left (Figure 3.22(b,c)). This demonstrates that the network develops either monotonically increasing or decreasing responses along the feature space as was reported by Freiwald et al. (2009). However, output neurons failed to

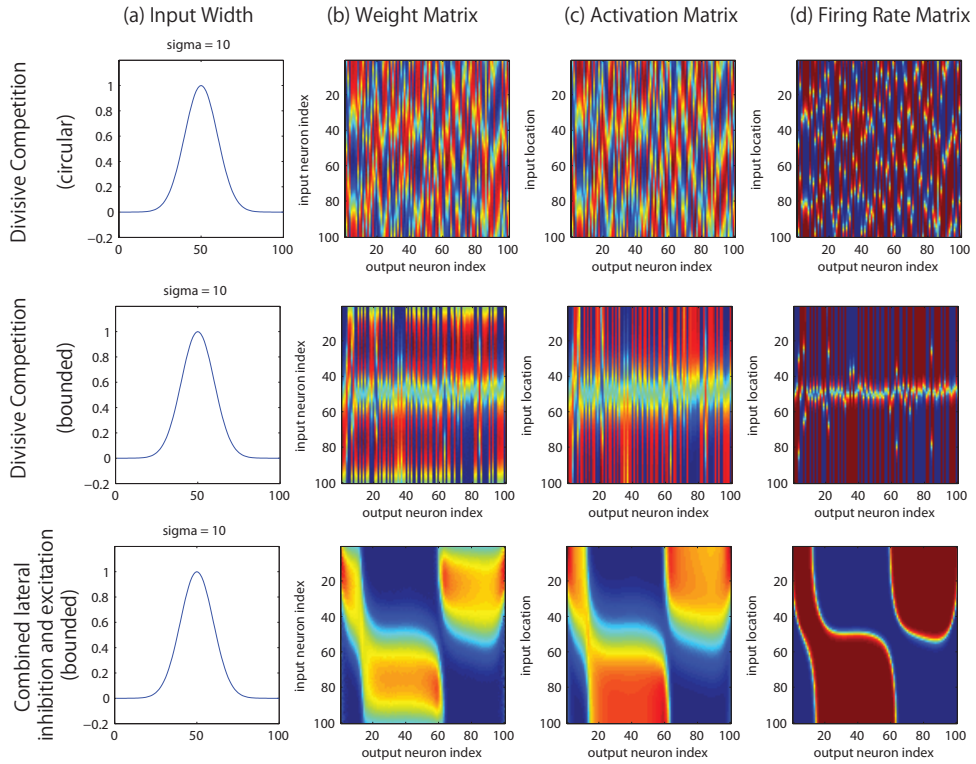

Figure 3.21: Simulation results investigating the development of monotonic tuning responses in the simplified network model with one layer of synapses as described in Section 3.3.1.2. The current location in the feature space is represented by the position of a Gaussian packet of activity imposed on the 1-dimensional layer of 100 input neurons. The results for three different simulations are shown in separate rows. Top row: network trained with circularly arranged input neurons with wrap-around. Middle row: network trained with linearly arranged input neurons with no wrap-around (divisive competition). Bottom row: network trained with linearly arranged input neurons with no wrap-around (combined lateral inhibition and excitation). The columns show the following: (a) the width, σ, of the Gaussian activity packet imposed on the input layer during training and testing, (b) matrix of synaptic weights from input neurons (ordinate) to output neurons (abscissa), (c) matrix showing activations of output neurons (abscissa) as a Gaussian activity packet is shifted through successive locations on the input layer (ordinate), and (d) matrix showing firing rates of output neurons (abscissa) as a Gaussian activity packet is shifted through the input layer (ordinate). Inspection of these plots shows that the trained networks with linearly arranged input neurons with no wrap-around (middle and bottom rows) have developed output neurons with monotonic tuning responses to the location of the Gaussian activity packet in the input layer regardless of the type of competition. However, output neurons did not show monotonic responses in the network trained with circularly arranged input neurons with wrap-around (top row).

show monotonic responses in the network trained with circularly arranged input neurons with wrap-around. Instead, the output neurons in the circular network developed peaked responses (Figure 3.22(a)).

These results are consistent with our original hypothesis described in Section 3.2. The key aspect of this network architecture that drives the development of monotonic tuning curves in the output layer is what happens at the two ends of the input space during learning. In the network trained with linearly arranged input neurons with no wrap-around, only part (e.g. half) of the input packet is represented at each end. In this case, the output neurons that learn to respond to the end locations develop relatively large synaptic weights. This is due to these output neurons becoming more tightly tuned to a smaller (end) region of the input layer by associative learning, but with the magnitudes of their synaptic weight vectors still renormalised over this smaller input region. If the widths of the input activity packets are relatively broad, then the same neurons continue to win the competition and respond when the input packet is shifted to a more central location within the input layer. However, as the input packet shifts away from the ends of the input layer, the responses of these neurons will decline monotonically. Different sub-populations of neurons will learn to respond to the two ends of the input layer, with each sub-population reducing its responses monotonically as the input packet shifts away

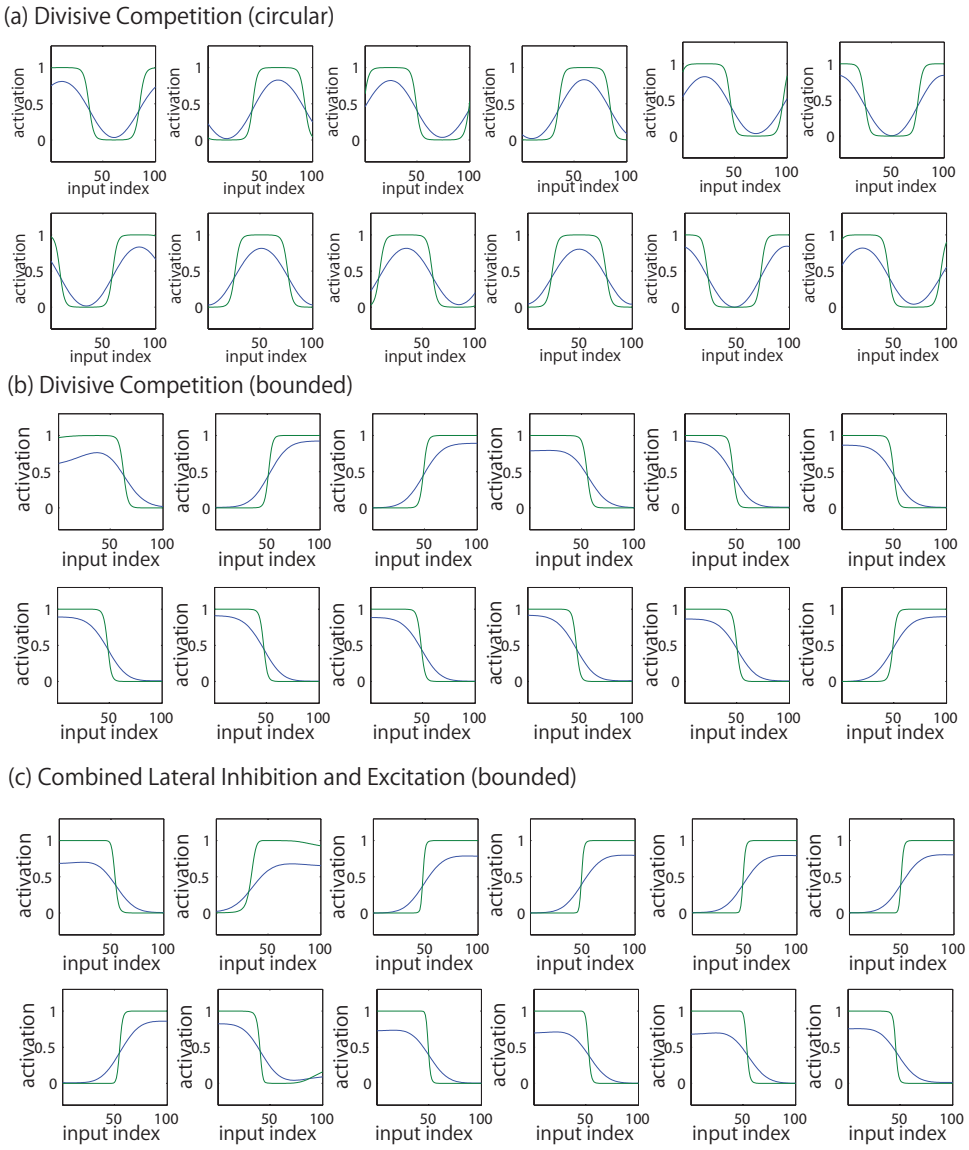

Figure 3.22: Simulation results investigating the development of monotonic tuning responses in the simplified network model with one layer of synapses as described in Section 3.3.1.2. The results for three different simulations are shown in separate blocks. Block a: network trained with circularly arranged input neurons with wrap-around. Block b: network trained with linearly arranged input neurons with no wrap-around (divisive competition). Block c: network trained with linearly arranged input neurons with no wrap-around (combined lateral inhibition and excitation). For each of the three simulations, the behaviour of twelve typical output neurons is shown in separate subplots. Each subplot shows how the activation (blue) and firing rate (green) of the neuron (ordinate) vary as a Gaussian activity packet is shifted through successive locations on the input layer (abscissa). It is evident that regardless of the type of competition implemented, the trained networks with linearly arranged input neurons with no wrap-around (block b and c) have developed output neurons with monotonic tuning responses to the location of the Gaussian activity packet in the input layer. However, this was not the case for output neurons in the network trained with circularly arranged input neurons with wrap-around (block a).

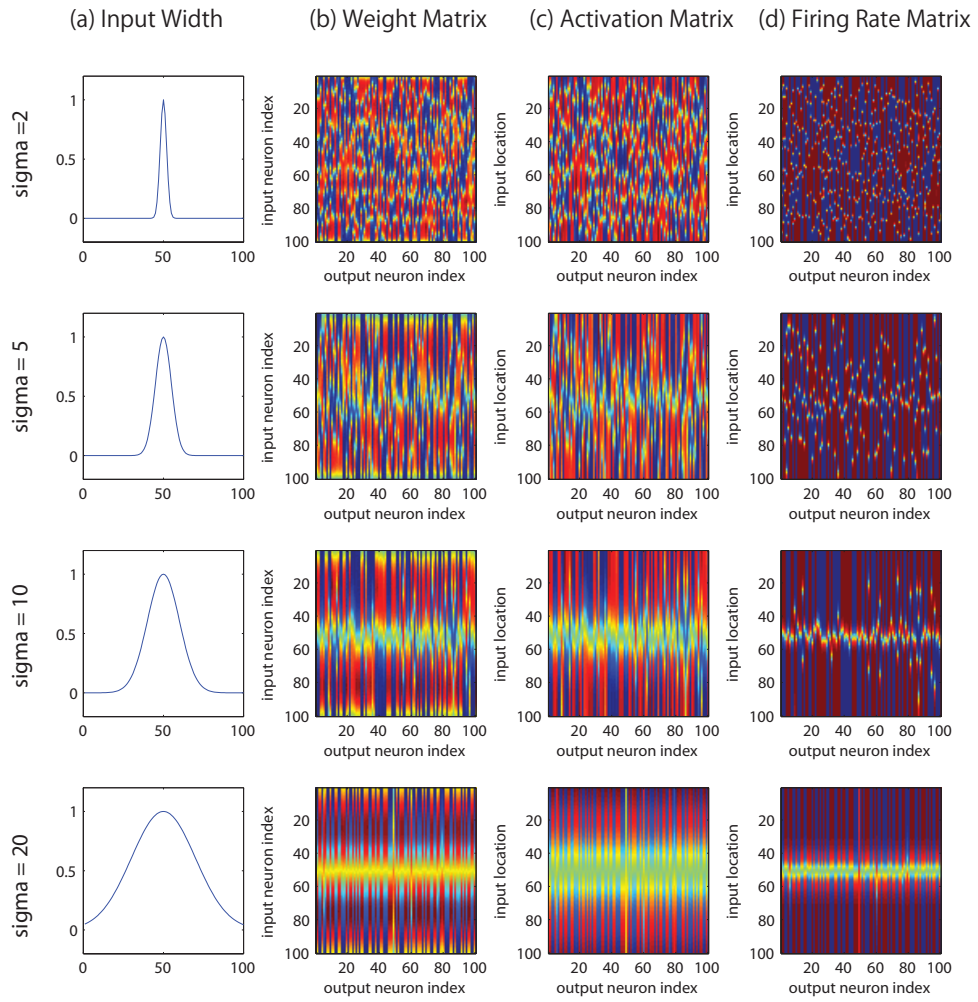

Figure 3.23: Simulation results investigating the development of monotonic tuning responses after training in the simplified network model with one layer of synapses. These simulations implemented divisive inhibition. Four simulations were run with different widths, σ, for the Gaussian packet of activity imposed on the 1-dimensional layer of 100 input neurons. The results for the four different simulations are shown in separate rows with σ set to 2, 5, 10 and 20 neurons. The columns follow the same conventions as in 3.21 and show the following: (a) the width of the Gaussian input packet, (b) synaptic weight matrix, (c) activation matrix, and (d) firing rate matrix. It can be seen that for a relatively small value of sigma equal to 2, the output neurons do not develop monotonic tuning responses. However, for a large value of $\sigma = 20$ neurons, the output neurons do display monotonic tuning profiles after training. Thus, the output neurons gradually switch to monotonic tuning curves with increases in the width, σ, of the Gaussian input packet.

from its preferred end location.

I also explored how varying the standard deviation σ that determine the width of the Gaussian activity packet imposed on the input layer affected the development of monotonic neuronal responses in the output layer. Figures 3.23 and 3.24 show the results of four simulations with divisive inhibition, where each simulation used a different value of σ set to 2, 5, 10, and 20, respectively. It can be seen that for a relatively small value of σ equal to 2, the output neurons do not develop monotonic tuning responses (Figure 3.23 (top row) and 3.24(a)). However, when σ is increased to 20, then the output neurons do develop monotonically tuned profiles after training (Figure 3.23 (bottom row) and 3.24(d)). Simulation results for σ equal to 5 or 10 show intermediate output behaviours. These results show that the output neurons gradually transition to developing monotonic responses as the standard deviation σ that determine the width of the Gaussian input packet increases. Thus, the width of the input packet needs to be reasonably large with respect to the size of the input space in order to drive the development of monotonically tuned output neurons.

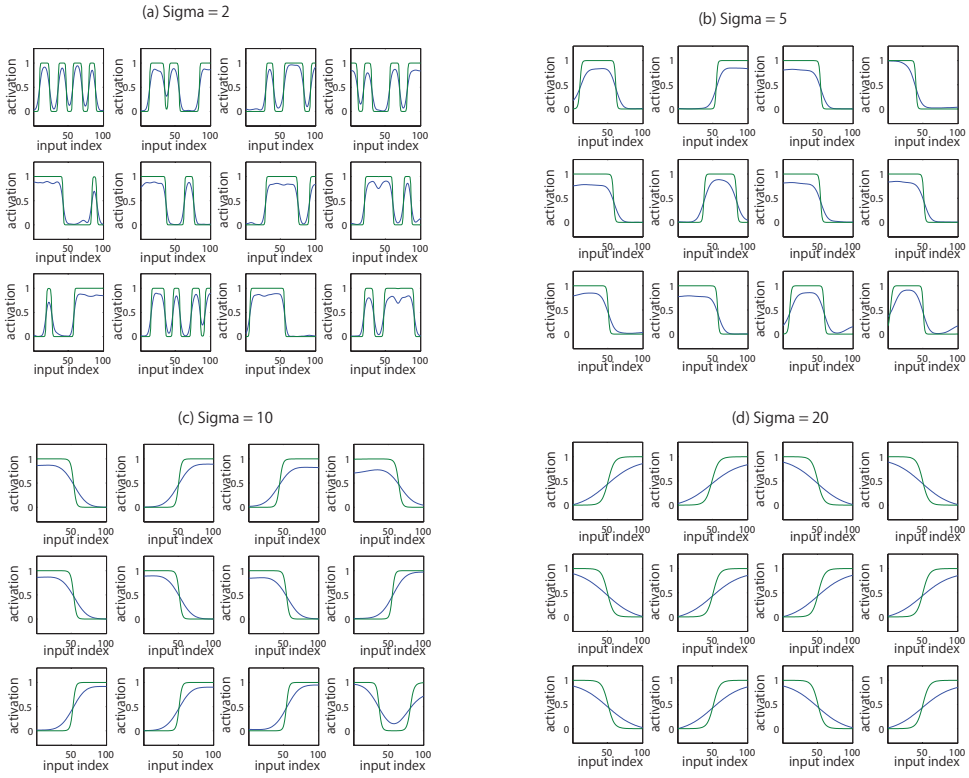

Figure 3.24: Simulation results investigating the development of monotonic tuning responses after training in the simplified network model with one layer of synapses. These simulations implemented divisive inhibition. Four simulations were run with different widths, σ, for the Gaussian packet of activity imposed on the 1-dimensional input layer. The results for the four different simulations are shown in separate blocks: (a) $\sigma = 2$ neurons, (b) $\sigma = 5$ neurons, (c) $\sigma = 10$ neurons, and (d) $\sigma = 20$ neurons. For each of the four simulations, the behaviour of twelve typical output neurons is shown in separate subplots. Each subplot shows how the activation (blue) and firing rate (green) of the neuron (ordinate) vary as a Gaussian activity packet is shifted through successive locations on the input layer (abscissa). It is evident that the output neurons gradually transition to developing monotonic responses as the width, σ, of the Gaussian input packet increases.

The above simulations showed how output neurons may develop monotonic tuning profiles when the network with divisive inhibition is trained with the input activity packet presented across all locations in the input feature space. However, Freiwald et al. (2009) showed that neurons in the middle face patch of the primate visual system maintained their monotonic tuning curves even when the monkey was presented with cartoon faces with unrealistically extreme spatial variations between the facial features, such as unrealistically large inter-eye distances, that could not have been encountered during prior visual experience. Accordingly, our next question was whether our model still develop output neurons that are monotonically tuned over the entire input feature space, including the extremal locations, if the model was trained with the input activity packet presented within only a limited central sub-region of the input feature space. Figures 3.25 and 3.26 show the results of three simulations in which the Gaussian activity packet was shifted over different sized central intervals, 25%, 50% and 75%, of the input layer during training. It can be seen that monotonic response curves still develop in the output layer when the input activity packet is presented within only 75 % or 50 % of the input feature space during training. Both of these two simulations still allow for some degree of truncation of the Gaussian activity packet at the two ends of the input layer, which is required for the development of monotonic tuning profiles in the output layer according to the hypothesis described in Section 3.2. These results thus confirm that the training set does not have to cover entire input space in order for the output neurons to develop monotonic tuning curves. This, in turn, offers an explanation for the experimental findings of Freiwald et al. (2009) that neurons in the monkey brain maintain their monotonic tuning curves even when

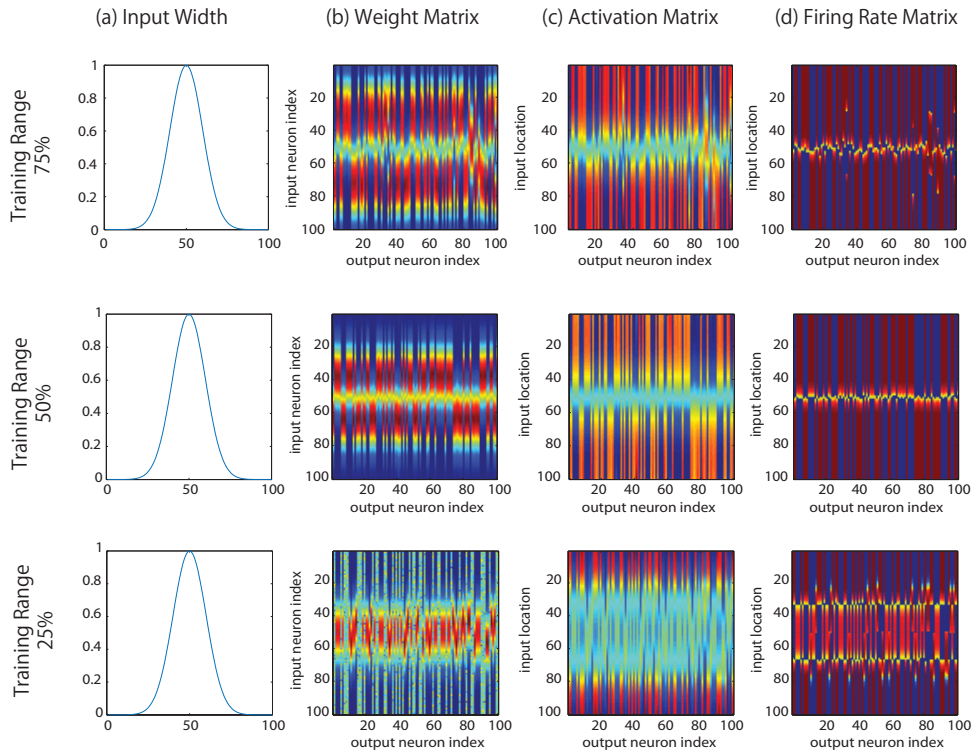

Figure 3.25: Simulation results investigating the development of monotonic tuning responses after training in the simplified network model with one layer of synapses. These simulations implemented divisive inhibition. Three simulations were run in which the Gaussian activity packet was shifted over different sized intervals of the input layer during training. The results for the three simulations are shown in separate rows. Top row: Gaussian activity packet is shifted during training over 75% of the input layer. Middle row: Gaussian activity packet is shifted over 50% of the input layer. Bottom row: Gaussian activity packet is shifted over 25% of the input layer. However, after training, the network is tested with the Gaussian activity packet presented at all locations on the input layer. The columns follow the same conventions as in Fig. 3.21 and show the following: (a) the width of the Gaussian input packet, (b) synaptic weight matrix, (c) activation matrix, and (d) firing rate matrix. It can be seen that the output neurons develop monotonic tuning when the Gaussian activity packet is shifted over 50% of the input layer during training (top and second row). However, as the Gaussian activity packet is shifted through less of the input layer during training, the output neurons gradually lose their monotonic responses.

presented with unrealistically extreme spatial variations between the facial features.

Lastly, I ran simulations to test whether the truncation of the Gaussian activity packet at the ends of the input layer during training played a key role in driving the development of monotonically tuned output neurons, as hypothesised in Section 3.2. In these simulations of the simplified network, the input layer was extended to include 100 extra neurons on either side of the 100 original input neurons, which gave a total of 300 input neurons. However, during training, the Gaussian activity packet was still presented only within the interval covering the original 100 input neurons. The inclusion of the extra 100 input neurons on either side of the original central region ensured that the Gaussian activity packet was not truncated at the original end locations. This should, according to our hypothesis described in Section 3.2, reduce the development of monotonic tuning profiles in the output layer. The result shown in Figure 3.27 support the hypothesis. In particular, some of the output neurons began to develop non-monotonic peaked (Gaussian) tuning responses as the end effects due to truncated Gaussian input packets broke down as the input layer was extended.

In conclusion, the above simulations show how the network is able to develop neurons that encode spatial relationships between facial features, such as the distance between the eyes, with monotonic tuning curves as reported in physiology (Freiwald et al., 2009). I proposed and provided evidence of the possible developmental mechanism of such cells, which is a result of competitive learning on the afferent connections into that cortical area when individual neurons

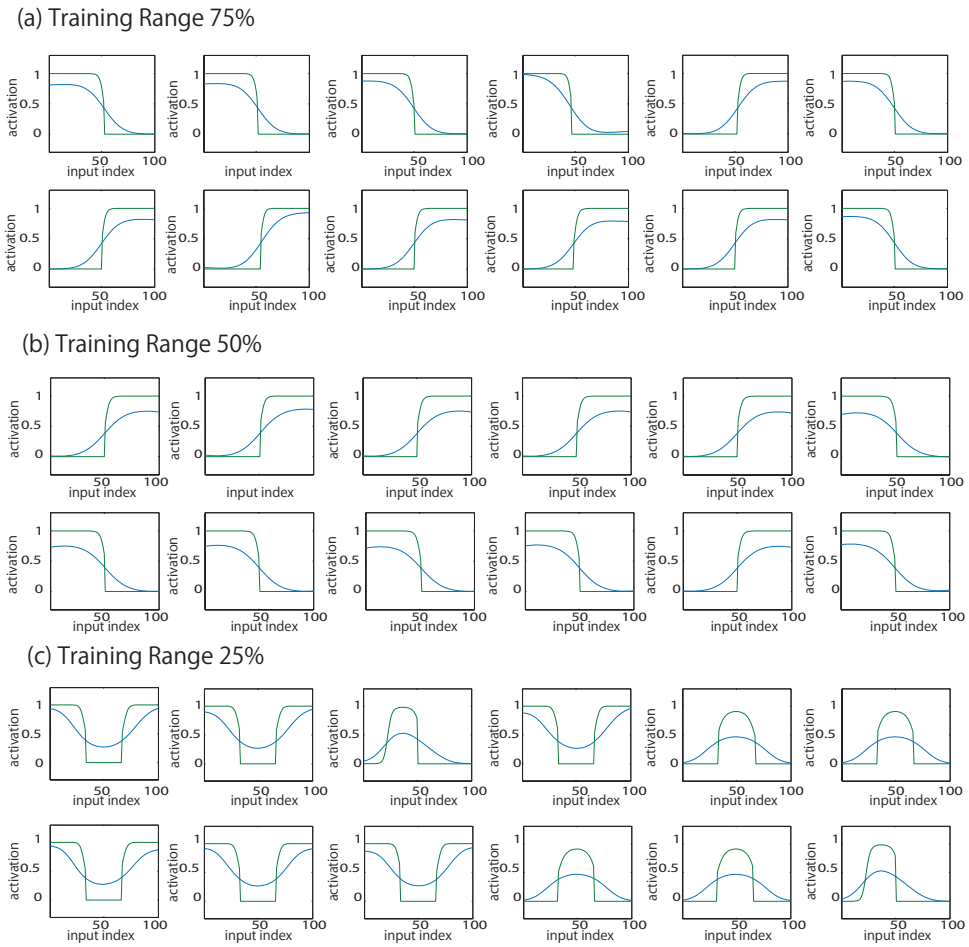

Figure 3.26: Simulation results investigating the development of monotonic tuning responses after training in the simplified network model with one layer of synapses. These simulations implemented divisive inhibition. Three simulations were run in which the Gaussian activity packet was shifted over different sized intervals of the input layer during training. Three simulations were run in which the Gaussian activity packet was shifted over different sized intervals of the input layer during training. The results for the three simulations are shown in separate blocks: (a) Gaussian activity packet is shifted during training over 75% of the input layer, (b) Gaussian activity packet is shifted over 50% of the input layer, (c) Gaussian activity packet is shifted over 25% of the input layer. After training, the network is tested with the Gaussian activity packet presented at all locations on the input layer. For each of the three simulations, the behaviour of twelve typical output neurons is shown in separate subplots. Each subplot shows how the activation (blue) and firing rate (green) of the neuron (ordinate) vary as a Gaussian activity packet is shifted through successive locations on the input layer (abscissa). It is evident that the output neurons display monotonic tuning when the Gaussian activity packet has been shifted over 50% of the input layer during training. However, the output neurons gradually lose their monotonic tuning as the Gaussian activity packet is shifted through a smaller interval of the input layer during training.

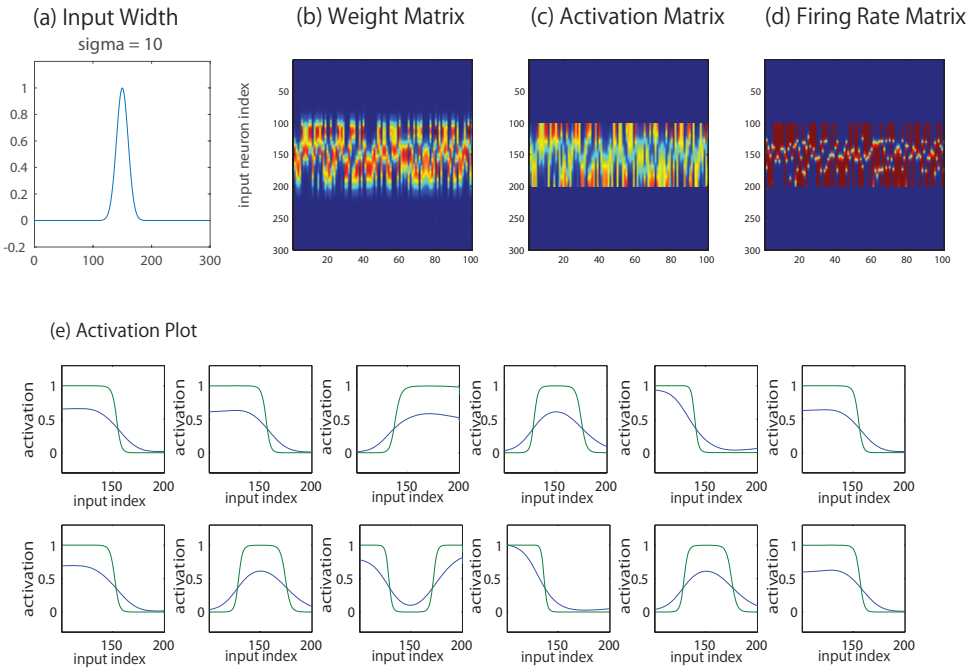

Figure 3.27: Simulation results investigating the development of monotonic tuning responses after training in the simplified network model with one layer of synapses. These simulations implemented divisive inhibition. In this simulation, the input layer was extended to include 100 extra neurons on either side of the 100 original input neurons to ensure that the Gaussian activity packet was not truncated at the original end locations. The columns follow the same conventions as in Fig. 3.21 and show the following: (a) the width of the Gaussian input packet, (b) synaptic weight matrix, (c) activation matrix, and (d) firing rate matrix. (e) In addition, the behaviour of twelve typical output neurons was shown in separate subplots. Each subplot shows how the activation (blue) and firing rate (green) of the neuron (ordinate) vary as a Gaussian activity packet is shifted through successive locations on the input layer (abscissa). It can be seen that some of the neurons show non-monotonic peaked (Gaussian) tuning profiles.

receive connections from a physically localised region of the preceding area leading to end effects.

3.4.3 Study 3: The representation of global facial attributes such as facial identity and expression

3.4.3.1 a. Simulation Results of VisNet

Neurophysiology studies have demonstrated the existence of separate clusters of neurons in the primate visual system that encode either facial identity or facial expression (Hasselmo et al., 1989a; Perrett et al., 1992; Morin et al., 2014). The question is how such cell response properties could develop. When Tromans et al. (2011) trained VisNet on cartoon faces of varying identity and expression, the network successfully developed separate clusters of neurons that encoded either facial identity or expression. However, when Tromans (2012) trained VisNet on a continuum of realistic faces generated using FaceGen, the network failed to develop separate representations of facial identity and expression. I hypothesised in Section 3.2 that this problem can be remedied by training VisNet on a more realistic, reduced set of face images, with only a limited number of different combinations of facial identity and expression. Another limitation of the study of Tromans et al. (2011) was that VisNet was trained on cartoon images of faces in which facial identity and expression were artificially represented by different facial features. I hypothesised in Section 3.2 that VisNet should still form neurons that respond to either facial identity or expression when the network is trained on more realistic faces where facial identity and expression may be represented by common facial features.

I tested whether neurons that responded selectively to either facial identity or expression had developed in VisNet during the initial training on 450 realistic human faces as shown in Figure

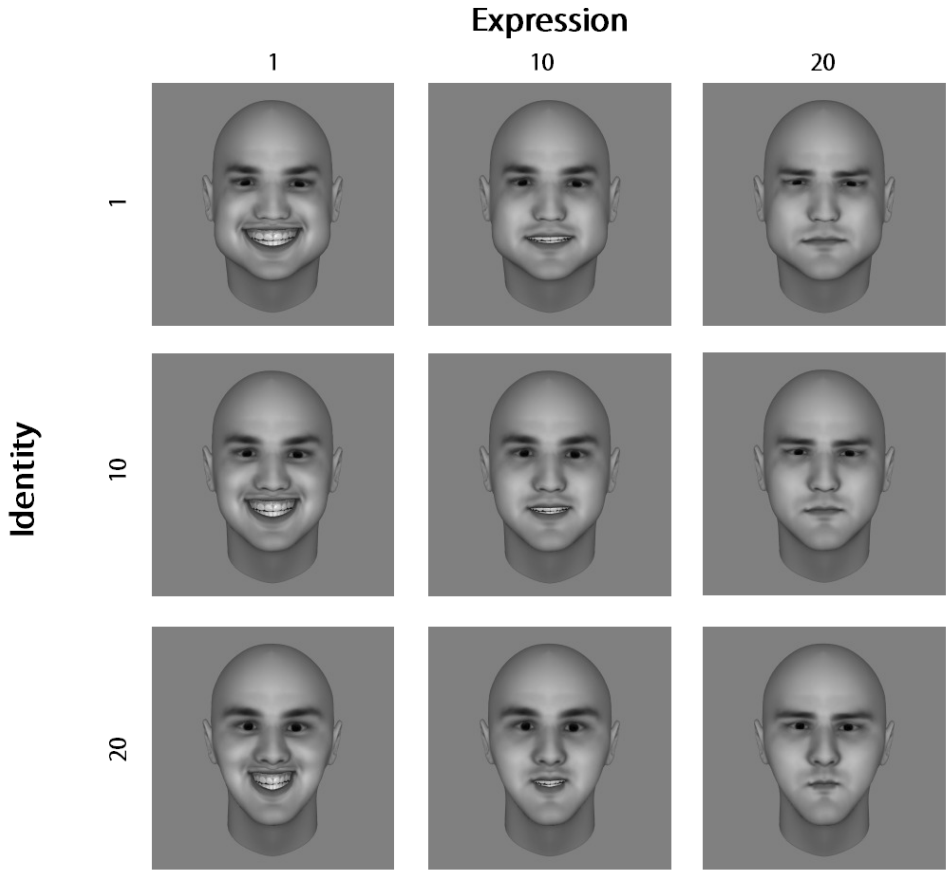

Figure 3.28: The face stimuli used to test VisNet for the existence of neurons that respond selectively to either facial identity or expression. A 1-dimensional space of 20 different facial identities, which varied gradually from one Identity A to another Identity B, was constructed. Each of these identities was then varied over a 1-dimensional space of 20 different expressions from Sad to Happy. This produced a matrix of 400 face stimuli constructed from 20 identities × 20 expressions.

3.10 and 150 non-face objects as shown in Figure 3.11. For this purpose, a new test set of realistic face stimuli using FaceGen was constructed as follows. A 1-dimensional space of 20 different facial identities, which varied gradually from an extreme Identity A to another extreme Identity B, was first created. Then each of these identities was varied over a 1-dimensional space of 20 different expressions from Sad to Happy. This resulted in a set of 400 face stimuli constructed from 20 identities × 20 expressions as shown in Figure 3.28. The trained network was tested on each face stimulus in the set, and the firing-rates of all neurons in the model were recorded.

Figure 3.29 shows the firing rate responses of typical neurons in the 4th layer of VisNet when tested on the facial stimuli representing combinations of identity and expression shown in Figure 3.28. Results are shown before and after training on the 450 realistic human faces shown in Figure 3.10 and 150 non-face objects shown in Figure 3.11. The individual plots in Figure 3.29 show how the firing rate of each neuron varies with facial identity and expression. Before training, the neuronal responses do not depend in a structured way on facial identity and expression (Figure 3.29(a)). However, after training, individual neurons have learned to respond selectively to localised regions of either the space of identities or space of expressions (Figure 3.29(b)). The first (top) row in Figure 3.29(b) shows neurons that have learned to respond to expressions near the right of the expression space bounded by Sad, while other neurons in the second row have learned to respond to expressions near the left of the expression space bounded by Happy. In contrast, the third row shows neurons that have learned to respond to identities on the top of the identity space bounded by Identity A, while other neurons in the

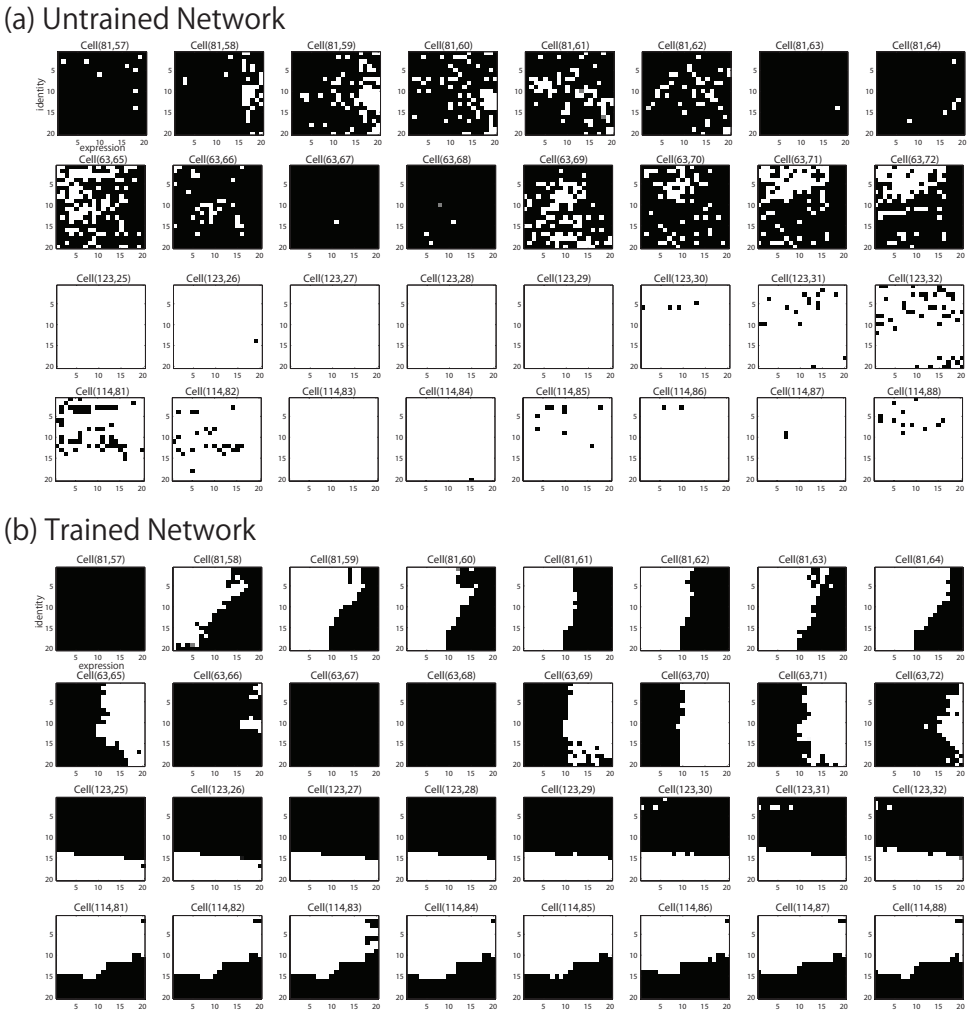

Figure 3.29: Firing rate responses of typical neurons in VisNet when tested on the facial stimuli representing combinations of identity and expression shown in Figure 3.28. Results are shown before training (a) and after training (b). Each row shows a different block of eight neurons in the 4th layer of the network. For each neuron, its firing rate as a function of facial expression (abscissa) and identity (ordinate) is plotted, with high firing denoted by black. Before training, the neuronal responses are quite unstructured with respect to facial identity and expression. However, after training, individual neurons respond selectively to localised regions of either the space of identities or space of expressions.

fourth (bottom) row have learned to respond to identities on the bottom of the identity space bounded by Identity B.

It can be seen that training VisNet has produced neurons that respond selectively to either particular identities or expressions. Furthermore, very interestingly, it can be seen that individual neurons have monotonic responses to the particular global feature dimension, i.e. identity or expression, which the neuron is tuned to. This is reminiscent of the neurons shown above in the study 2 in Section 3.4.2, which represent the spatial relationships between facial features with monotonic tuning curves. The question is could there be an underlying connection between these two kinds of neuron. This idea is further explored below. Figure 3.29 also shows that neurons that represent specific identities tend to be clustered close together, and the same is true for neurons that encode particular expressions. This is due to a combination of short range excitation and long range inhibition, effecting a self-organising map (SOM), within each layer of VisNet.

Figure 3.30 shows the results of analysing the amount of single and multiple cell information carried by 4th layer neurons in VisNet about facial identity and expression before and after training (Section 1.5.2). The left column of Figure 3.30 shows the amount of information about

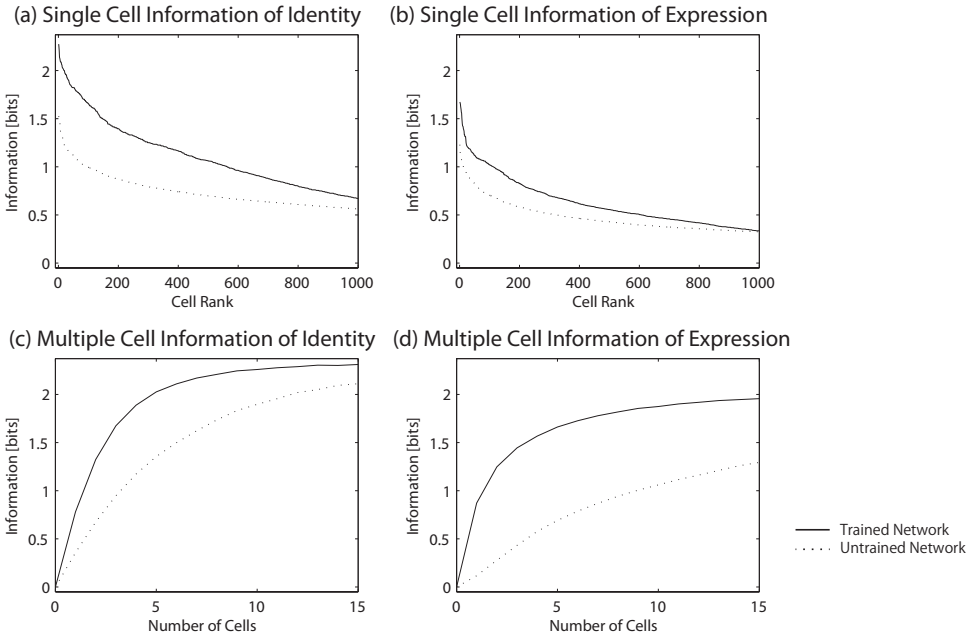

Figure 3.30: Analysis of the amount of single and multiple cell information carried by 4th layer neurons in VisNet about facial identity and expression before and after training. The left column shows the results of analysing the amount of single cell information (a) and multiple cell information (c) about identity conveyed by fourth layer cells before and after training. (a) shows the amount of single cell information carried by output cells plotted in rank order. The dotted line represents the untrained network while the solid line represents the trained network. This analysis involved quantising the identity space into five separate contiguous blocks. The maximal amount of information possible in this case is $log_2(5) = 2.32$ bits. The right column shows equivalent results of analysing the amount of single cell information (b) and multiple cell information (d) about expression conveyed by fourth layer cells before and after training. This analysis similarly involved quantising the expression space into five separate contiguous blocks. The maximal amount of information possible is again 2.32 bits. It is evident that training has significantly increased the amount of single and multiple cell information carried by 4th layer neurons about both facial identity and expression.

identity conveyed by fourth layer cells. This analysis involved quantising the identity space into five separate contiguous blocks. The maximal amount of information possible in this case is $log_2(5) = 2.32$ bits. The right column of Figure 3.30 shows equivalent results for the amount of information conveyed by fourth layer cells about expression. It can be seen that training has led to a substantial increase in the amount of single and multiple cell information about both facial identity and expression. Thus, information analysis confirms the enhanced selectivity of neurons for either identity or expression after the training. More than 100 neurons carry 1.5 bits or above of single cell information for facial identity, and around 100 cells carry 1 bit or above of single cell information for expression. This is consistent with the monotonic tuning curves shown in Figure 3.29. Such neurons respond to a localised region at one end of their preferred feature space, e.g. responding to Happy faces but not Sad faces. Such neuronal responses will carry at least 1 bit or more of information about the neuron's preferred feature space, i.e. identity or expression. However, different neurons have monotonic tuning curves with different slopes. This means that the distributed representation across a population of such neurons should still be sufficient to specify the exact identity or expression of a face stimulus. The reason why neurons were found to encode more information about identity than expression in our simulations might be related to the fact that with a set of realistic faces, such as the Ekman set (Friesen and Ekman, 1976), a pixel wise variation in identity tends to be greater than the variation in expression (Calder et al., 2001) so enabling easier discrimination for identity.

I next investigated what facial features, such as eyes, nose, and mouth, the different kinds of 4th layer neurons were responding to. This was done by tracing the connections that had been strengthened by learning from the 4th layer neurons back to the input Gabor input filters (see Section 1.5.3). Figure 3.31 shows the Gabor filters that have strong connectivity through

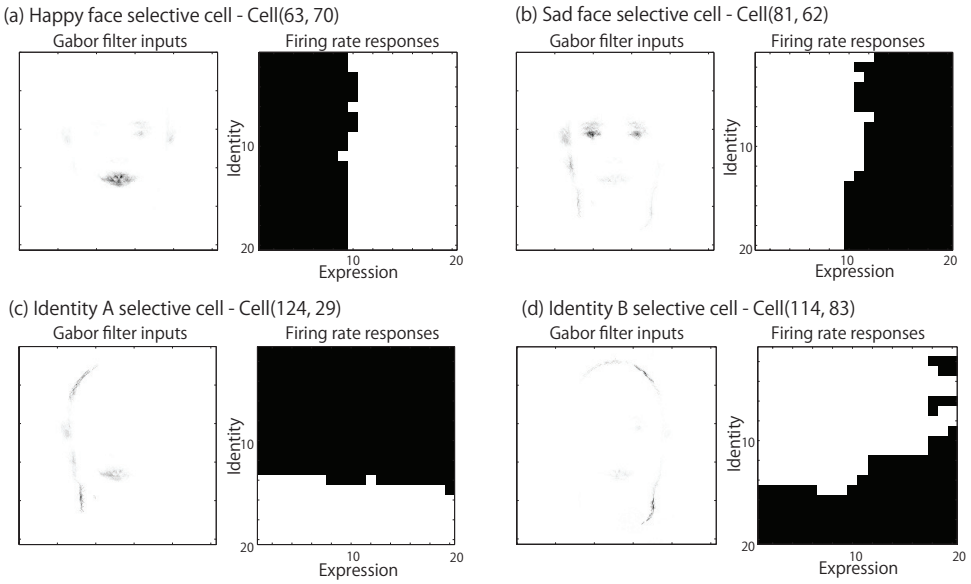

Figure 3.31: The input Gabor filters that have strong connectivity through the network to example 4th layer neurons in VisNet which are tuned to four global attributes: Happy (top left), Sad (top right), Identity A (bottom left), and Identity B (bottom right). Each of these four subplots shows the Gabor filters with strong connectivity to that neuron (left) and the neuron's firing rate responses to the facial stimuli representing combinations of identity and expression shown in Figure 3.28.

the network to example 4th layer neurons in VisNet which are individually tuned to one of four global attributes: Happy, Sad, Identity A, and Identity B. Each of the four corresponding subplots shows the Gabor filters with strong connectivity to that neuron as well as the neuron's firing rate responses to the facial stimuli representing combinations of identity and expression shown in Figure 3.28. It can be seen that the neuron tuned to Happy faces receives strong connectivity from Gabor filters representing the mouth. On the other hand, the neuron tuned to Sad faces receives strong connectivity from Gabor filters representing the eyes and eyebrows. Interestingly, the two neurons that differentiated between Identity A and Identity B received strong connectivity from Gabor filters representing the facial outline.

The above results indicate that there might be a relationship between neurons that represent particular global attributes, such as Happy, Sad, Identity A, and Identity B, and the kind of neurons discussed in the study 2 in Section 3.4.2 that encode the spatial relationships between local facial features such as the eyes, nose, and mouth with monotonic tuning curves. In fact, it could be that these apparently two different kinds of neuronal response characteristic is displayed by the same neurons. In other words, I wonder if the neuron that appears to respond to a global attribute such as Sad is simply responding to a particular spatial relationship between local facial features with a monotonic tuning curve.

To investigate this possibility, the four example cells shown in Figure 3.31, which represent particular global attributes such as Happy, Sad, Identity A, and Identity B, were taken to apply the same analysis that was used in Study 2 for Figure 3.20 with the test faces shown in Figure 3.19. These results are shown in Figure 3.32, which shows how the firing rate responses of the four neurons vary with the spatial relationships between facial features. Each row shows the responses of a different neuron, while each column corresponds to a different kind of spatial relationship: inter-eye distance, eyebrow angle, eye height and mouth shape. The individual subplots show the firing rate responses of the neuron as the corresponding spatial relationship is varied across ten selected feature values. It can be seen that some neurons responding to global attributes such as Sad and Identity A have monotonic tuning to particular spatial relationships between local facial features. For example, the cells that encode the facial expressions Happy and Sad essentially encode the shape of the mouth. While the cell tuned to Identity A encodes

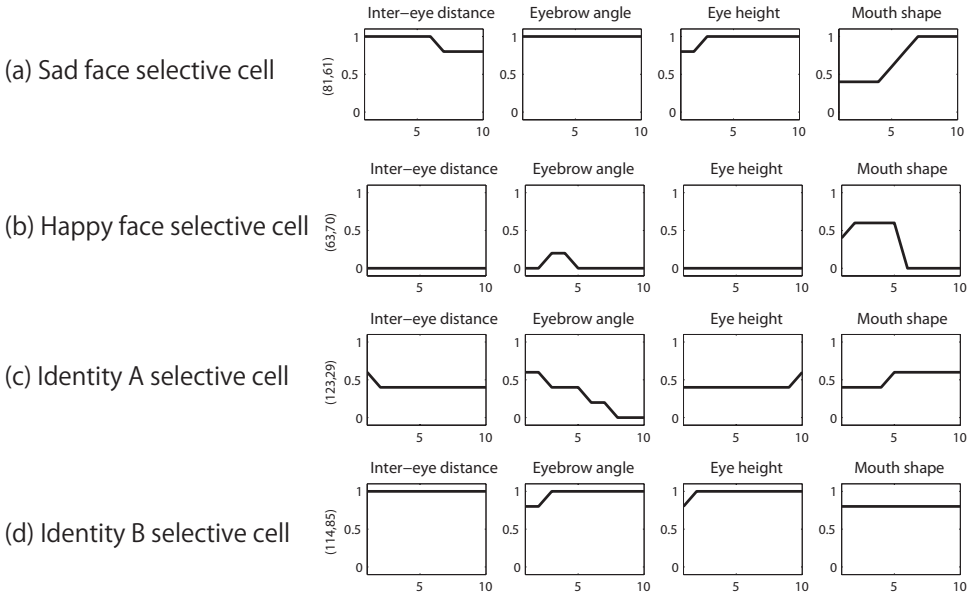

Figure 3.32: How the firing rate responses of four example neurons in the 4th layer of VisNet, which respond selectively to global attributes such as sad, happy, Identity A, and Identity B, depend on variation in the spatial relationships between facial features. Each row shows the responses of a different neuron, while each column corresponds to a different kind of spatial relationship: inter-eye distance, eyebrow angle, eye height and mouth shape. The individual subplots show the firing rate responses of the neuron as the corresponding spatial relationship is varied across ten selected feature values. It is evident that some neurons responding to global attributes such as sad and Identity A have monotonic tuning to particular spatial relationships between local facial features.

the eyebrow angle. The cell tuned to Identity B does not show a strong correlation to any of the four dimensions tested, but it is quite possible that this cell encodes a different spatial relationship between facial features that is not shown here. These results strongly support the notion that a neuron that appears to respond to a global attribute is simply responding to a particular spatial relationship between local facial features with a monotonic tuning curve. Thus, these two kinds of neuron may in fact be the same, with neurons encoding different global attributes simply representing different spatial relationships between local features with monotonic tuning curves or particular combinations of them.

Additional Study: the Representation of Six Basic Expressions So far in this chapter, VisNet has been trained on only two different facial expressions, happy and sad, including intermediate expressions. This leaves open the question of whether VisNet could learn to recognise a larger number of different expressions. This question is in part motivated by a recent study by Sormaz et al. (2016), which has shown that the perceptual similarity of five expressions (happy, sad, angry, disgust, and fear) could be predicted from the patterns of neural response in the STS.

To address this question, an additional VisNet study where the network was trained on a set of 100 randomly generated facial identities for each one of six basic expressions (Happy, Sad, Anger, Disgust, Fear and Surprise) was conducted. Then, the network was tested on a new set of randomly generated face stimuli for each of the 6 facial expressions. Specifically, for each expression, 10 different random facial identities were created in order to test whether the network representation of facial expression could generalise across the different facial identities. Figure 3.33 shows the results of the simulation. Each subplot in the top row shows the average responses of 10 cells, which are identified to carry the highest single cell information for a particular facial expression, to ten different randomly generated facial identities with that expression.

Although these results do not show perfect performance, it can be seen that the neurons shown in each subplot do respond more to faces with their preferred expression. The subplots

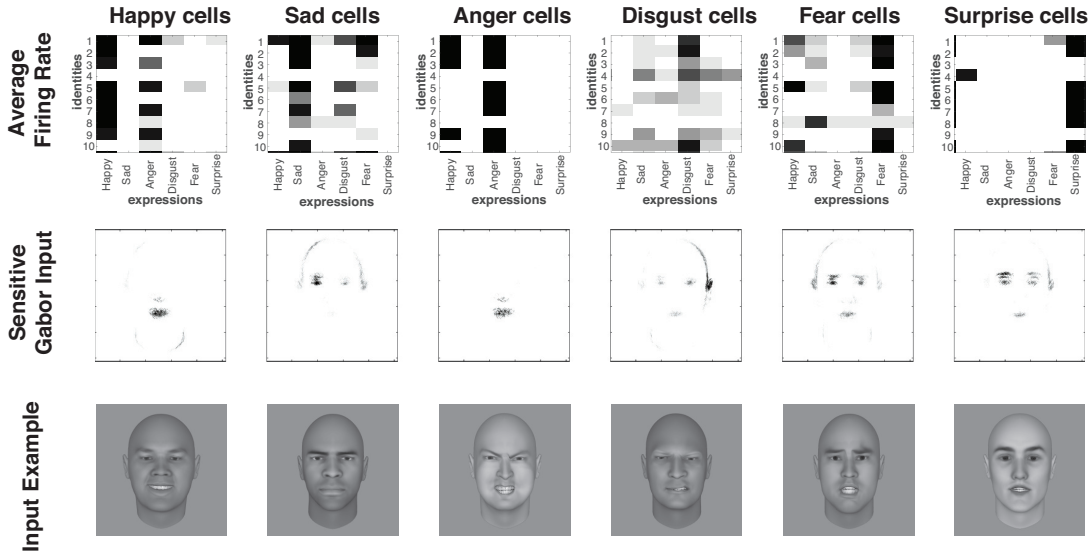

Figure 3.33: Results of the additional VisNet study where the network was trained on a set of 100 randomly generated facial identities for each one of six basic expressions: happy, sad, angry, disgust, fear, and surprise. The network was tested on face stimuli with the same six expressions. For each expression, 10 different randomly generated facial identities were created in order to test whether the network representation of facial expression could generalise across the different facial identities. Each subplot in the top row shows the average firing rate of the ten 4th layer neurons that carry the most information about one of the six expressions, with the neurons encoding each expression shown in a separate column. Each subplot shows the average responses of the ten neurons to ten different randomly generated facial identities with that particular expression. The subplots in the middle row show the average inputs from the gabor filters that are most strongly connected to the ten output cells that represent each expression shown in the top row. The images in the bottom show examples of the randomised face stimuli with the corrsponding facial expression used to test the network.

in the middle row show the gabor filters that are most strongly connected to the ten output cells in the top row that represent each expression. It can be seen that the Happy neurons (first column) are receiving strong connections from gabor filters representing the shape of the mouth, while the Anger neurons (third column) are receiving strong connections from a different part of the mouth. The Sad neurons (second column) are receiving strong inputs from gabor filters representing the shape of the eyes, while Fear neurons (fifth column) and Surprise neurons (sixth column) receive strong inputs from different parts of the eyebrows.

3.4.3.2 b. Learning mechanisms by which the network may form distinct representations of global facial attributes such as identity and expression : One layer network simulations

The question is how the network can develop separate output representations of different global facial attributes such as identity and expression if these attributes are always seen together at the same time. Moreover, I wonder how the same retinal input neurons are used to encode the two global attributes simultaneously. Somehow, through a hierarchical series of neuronal layers, the primate visual system must use competitive learning to separate these global attributes, which are initially encoded by overlapping sets of retinal input neurons, onto distinct populations of output cells.

In order to explore the mechanisms by which this transformation might take place, I ran simulations of an idealised one-layer competitive neural network as described in Section 3.3.1.2, but now with two 1-dimensional input spaces. One of the input spaces could be considered as encoding facial identity, while the other input space encoded facial expression. Each input space was represented by a 1-dimensional row of 100 neurons.

In some simulations the two input spaces were completely orthogonal to each other in that

they shared no input neurons. In this case, there was a total of 200 input neurons. On the other hand in other simulations, the two input spaces shared some neurons. In the case of completely overlapping input spaces the network contained a total of 100 input neurons, with these neurons ordered differently within the two spaces. The location of a face in each of these two spaces was encoded by the position of a Gaussian activity packet within that input space. The standard deviation σ governing the width of the Gaussian activity packets in both input layers was was set to 10 in all simulations.

At each timestep during training, an input stimulus was defined by Gaussian activity packets presented at random locations within each of the two input layers. In simulations with dependent motion, the two input spaces were fully statistically linked in that the Gaussian activity packets always occurred at corresponding locations in the two spaces. In simulations with independent motion, the two input spaces were statistically independent in that the locations of the Gaussian packets in the two spaces were entirely independent of each other.

The activities of the input neurons were fed through the feedforward synaptic connections to drive the responses of 100 neurons in the output layer. Combined lateral inhibition and excitation was implemented between neurons in the output layer in order to effect competition. During training, the feedforward synaptic weights were then modified using a local associative (hebbian) learning rule with synaptic weight vector normalisation as described in Section 3.3.1.2.

I explored how the output representations that developed in the network though learning were affected by (i) the degree of statistical independence between the two spaces during training, i.e. whether identity and expression varied independently of each other over the stimulus training set, and (ii) the degree of overlap between the input neurons encoding the two spaces, i.e. how many neurons the two input spaces had in common.

After training, the learned response behaviours of the output neurons were analysed using two methods. In the first method, the position of the Gaussian activity packet in one of the input spaces was systematically shifted through neurons 1 to 100, while the position of the Gaussian packet in the other input space remain fixed at the centre of that space. In the second method, Gaussian activity packets were presented at all 100×100 combinations of positions within the two input spaces, and the firing rate response table of each output cell was recorded for comparison with the results of the VisNet simulation reported in Figure 3.29.

Figure 3.34 shows how the *dependent* motion of Gaussian activity packets in two input layers during training affects the learned response properties of output neurons. In this simulation there was no overlap between the two input spaces. The top six subplots show the results of the first method of analysis. The three columns show the (a) weight matrix, (b) activation matrix, and (c) firing rate matrix of the population of output neurons. With dependent motion of the activity packets in the two input layers during training, the firing rate maps of the output neurons in response to the two input spaces largely overlap. Thus, the output neurons failed to develop separate representations of the two input spaces. The four subplots in the bottom row (d) show the second method of analysis. Each of the four subplots in the bottom row shows the firing rate responses for a different output neuron. Individual output neurons learned to respond to particular combinations of locations in the two input spaces that occurred together during training. Thus, these neurons had not learned to respond selectively to just one or other of the two input spaces.

Figure 3.35 shows how the *independent* motion of Gaussian activity packets in two input layers during training affects the learned response properties of output neurons. In this simulation there was again no overlap between the two input spaces. It can be seen in (a), (b) and (c) that the output neurons have developed separate representations of the two input spaces, with individual neurons responding to just one of the input spaces. In particular, the four output neurons shown in (d) each learned to respond selectively to a localised end region of one of the input spaces. For example, cell 82 shown in the first column of Figure 3.35(d) responds selectively to the right side of input space B regardless of where an activity pattern occurs in input

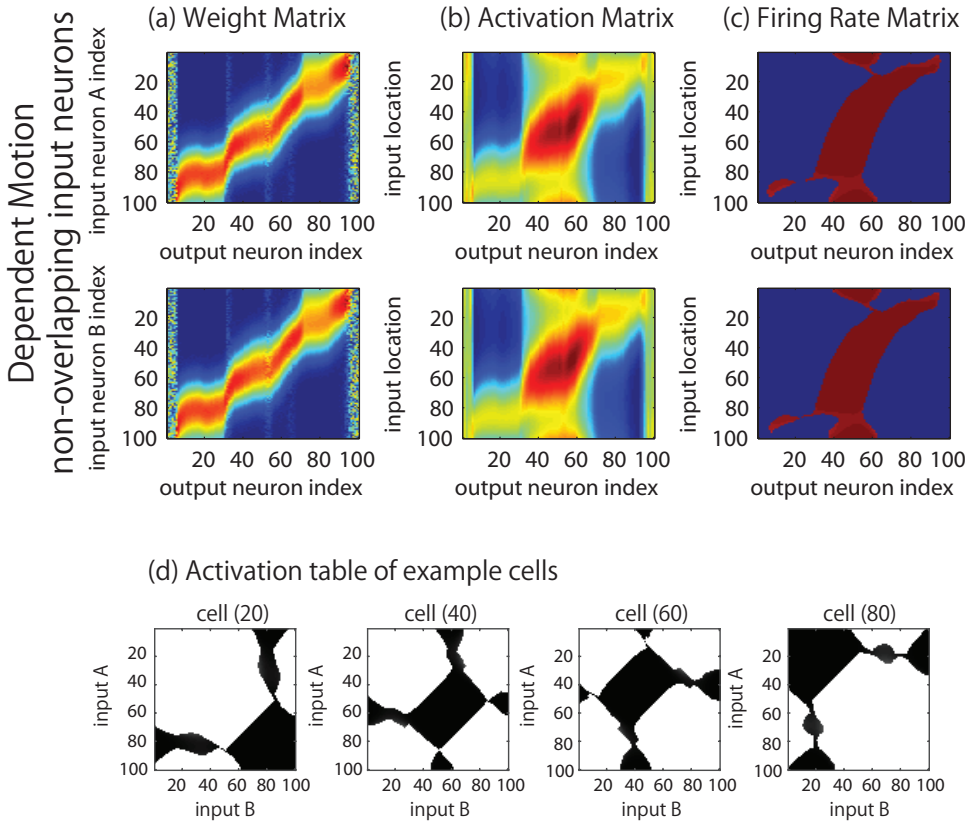

Figure 3.34: Simulation of a one-layer competitive network showing how the *dependent* motion of Gaussian activity packets in two input layers during training affects the learned response properties of output neurons. In this simulation there was no overlap between the two input spaces. The top six subplots show the results of the first method of analysis, in which the position of the Gaussian activity packet in one of the input spaces was systematically shifted through neurons 1 to 100, while the position of the Gaussian packet in the other input space remain fixed at the centre of that space. The first (top) row corresponds to shifting the activity packet through the first input space, while the second row corresponds to shifting the activity packet through the second input space. The three columns show the (a) weight matrix, (b) activation matrix, and (c) firing rate matrix of the population of output neurons. It can be seen that, with dependent motion of the activity packets in the two input layers during training, the output neurons have failed to develop separate representations of the two input spaces. The four subplots in the bottom row (d) show the second method of analysis, in which Gaussian activity packets were presented at all 100×100 combinations of positions within the two input spaces, and the firing rate response tables of each output cell were recorded. Each of the four subplots in the bottom row shows the table of firing rate responses for a different output neuron. It is evident that individual output neurons have learned to respond to particular combinations of locations in the two input spaces that occurred together during training.

space A. Similarly, cell 75 presented in the third column of Figure 3.35(d) responds selectively to the top of input space A regardless of the location of an activity pattern in input space B. Thus, the output neurons successfully developed distinct representations of the two input spaces when the motion of the Gaussian patterns in the two input layers was independent.

Next, the network was tested by gradually increasing the overlap of the two input spaces. Figure 3.36 shows the results of a simulation with 50% overlap, while Figure 3.37 shows the results of increasing the overlap to 100%. In both of these simulations, the Gaussian activity patterns moved independently through the two input spaces during training. It was found that even if the retinal input neurons are entirely overlapped, the output neurons still developed separate representations of two input spaces. This effect relied on the motions of the activity patterns in the two input spaces being independent during training.

I propose that these simulations may explain how the primate visual system develops physically separate representations of global facial attributes such as identity and expression, with individual neurons responding selectively to a localised region of one of these spaces, even though both attributes are encoded by the same population of retinal input neurons.

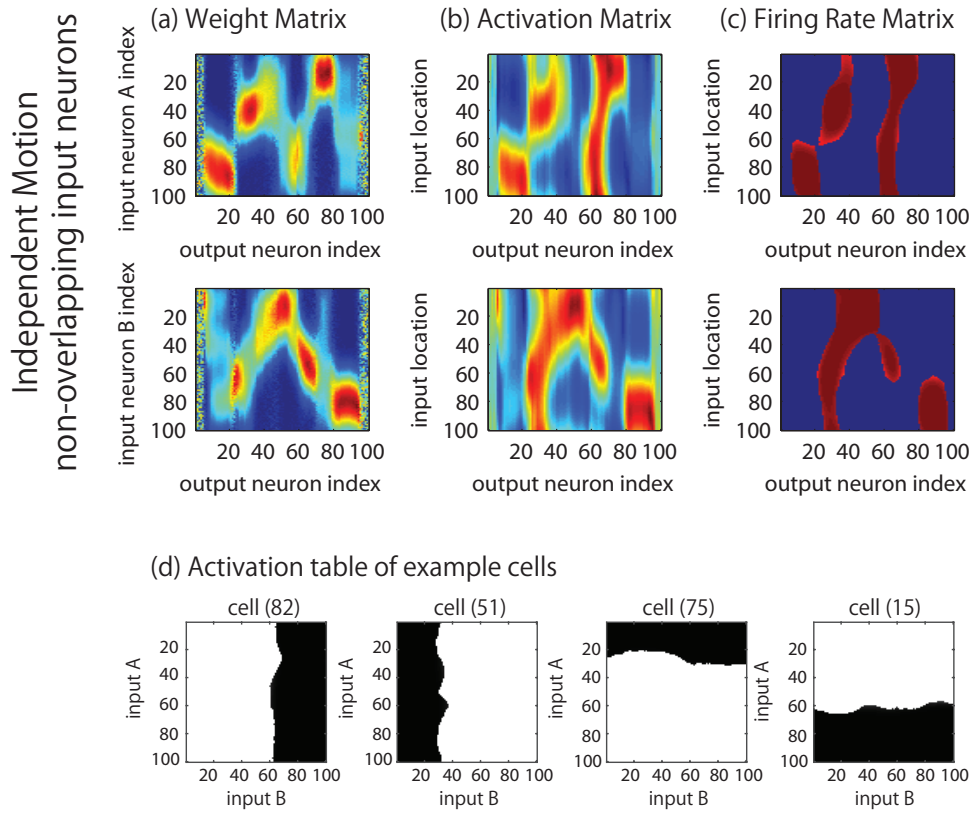

Figure 3.35: Simulation of a one-layer competitive network showing how the *independent* motion of Gaussian activity packets in two input layers during training affects the learned response properties of output neurons. In this simulation there was no overlap between the two input spaces. Conventions as in Figure 3.34. It can be seen in (a), (b) and (c) that the output neurons have developed separate representations of the two input spaces, with individual neurons responding to just one of the input spaces. In particular, the four output neurons shown in (d) have each learned to respond selectively to a localised end region of one of the input spaces.

3.5 Discussion

I presented biologically plausible neural network simulations of the visually-guided development of facial representations in the visual brain using completely unsupervised learning mechanisms with feed-forward visual processing. These simulations contrast with many current engineering approaches based on the feedback of error signals from higher- to lower-levels of representation to guide supervised learning of facial attributes such as identity and expression (Lawrence et al., 1997; Lisetti and Rumelhart, 1998; Taigman et al., 2014). Supervised learning by back-propagation of error (Rumelhart et al., 1986) is not a biologically plausible mechanism for learning facial representations in the brain. Although there exist back-projections in the visual system, it is not possible that these are carrying the kind of error signals needed by back-propagation of error learning (Stork, 1989). Hence, the simulations reported in this chapter represent an important theoretical advance in understanding how the visual system in the brain learns to represent the rich spatial structure of the faces.

In this chapter, a series of simulation studies investigating how visual representations of faces may develop in the primate visual system was conducted. In particular, VisNet was trained with realistic human face stimuli constructed using FaceGen. As a result, it was found that the network successfully developed various kinds of cells with response properties similar to those reported in neurophysiological studies. To further advance our understanding of the learning mechanisms involved, additional simulations were performed within simplified one-layer competitive network models.

Our initial simulations with the VisNet model showed the development of neurons that

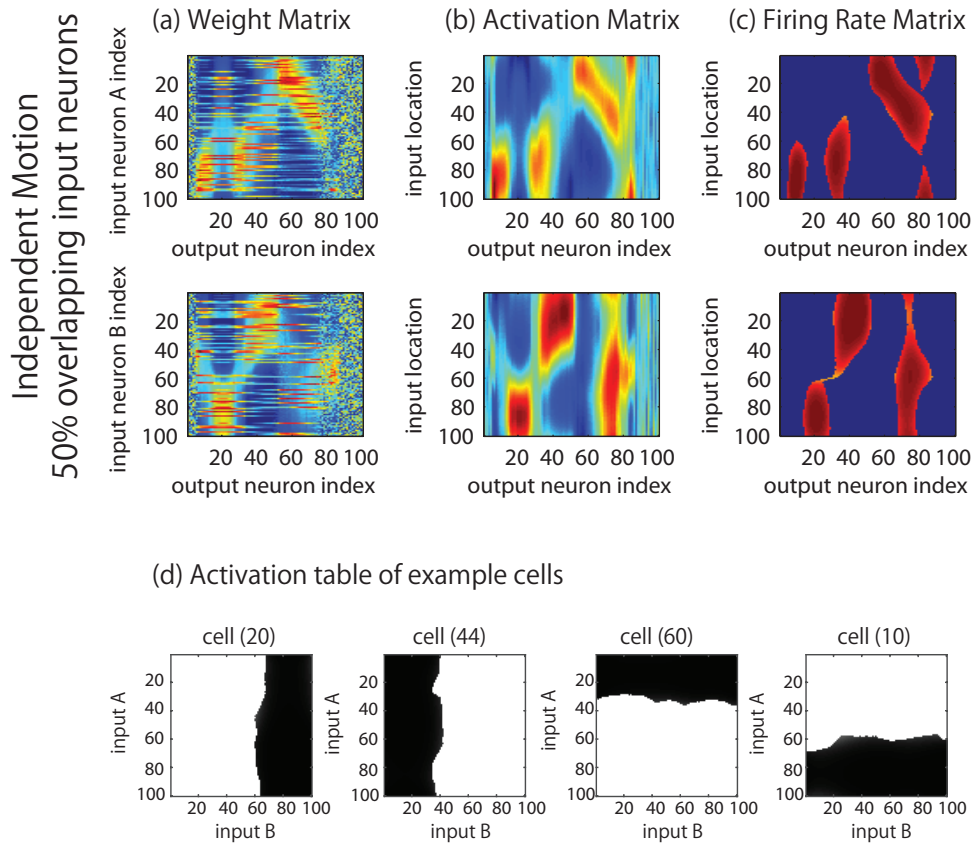

Figure 3.36: Simulation of a one-layer competitive network showing how the *independent* motion of Gaussian activity packets in two input layers during training affects the learned response properties of output neurons. In this simulation there was a 50% overlap between the two input spaces. Conventions as in Figure 3.34. Even though there is a 50% overlap between the two input spaces, subplots (a), (b) and (c) show that the output neurons still developed separate representations of the two input spaces. The four output neurons shown in (d) have each learned to respond selectively to a localised end region of one of the input spaces.

learned to respond to individual facial features such as the eyes and mouth, as well as combinations of these features, as has been reported in single cell recordings in the macaque brain (Freiwald et al., 2009). However, the question was how neurons might learn to respond to individual facial features if the facial features are always seen together within whole faces during training. Particular facial features such as the eyes occur in different shapes across different faces. Thus, across a population of faces the network will be exposed to different combinations of facial feature shapes on different occasions. This will lead to a statistical decoupling (Stringer et al., 2007; Stringer and Rolls, 2008) between the individual facial features, which I hypothesised may force the neurons in higher layers to learn to represent the individual features rather than whole faces. This hypothesis was confirmed in the VisNet simulations, where it was found that the output neurons switched to predominantly representing the individual facial features as the number of possible shapes of any facial feature p used to generate the set of training faces increased from 1 to 2.

I further hypothesised that as the number of shapes of any facial feature p increased further, an invariance learning mechanism known as continuous transformation (CT) learning would begin to drive the development of neurons that responded invariantly to many or all of the shape variations of a particular facial feature. Such neurons would represent a facial feature such as a mouth irrespective of the particular shape of that feature. This hypothesis was also confirmed in VisNet simulations as p was increased to 5, 10 and 30. At $p = 30$ there was a sharp rise in the number of neurons that responded to all 50 of the differently shaped eyes used to test the network.

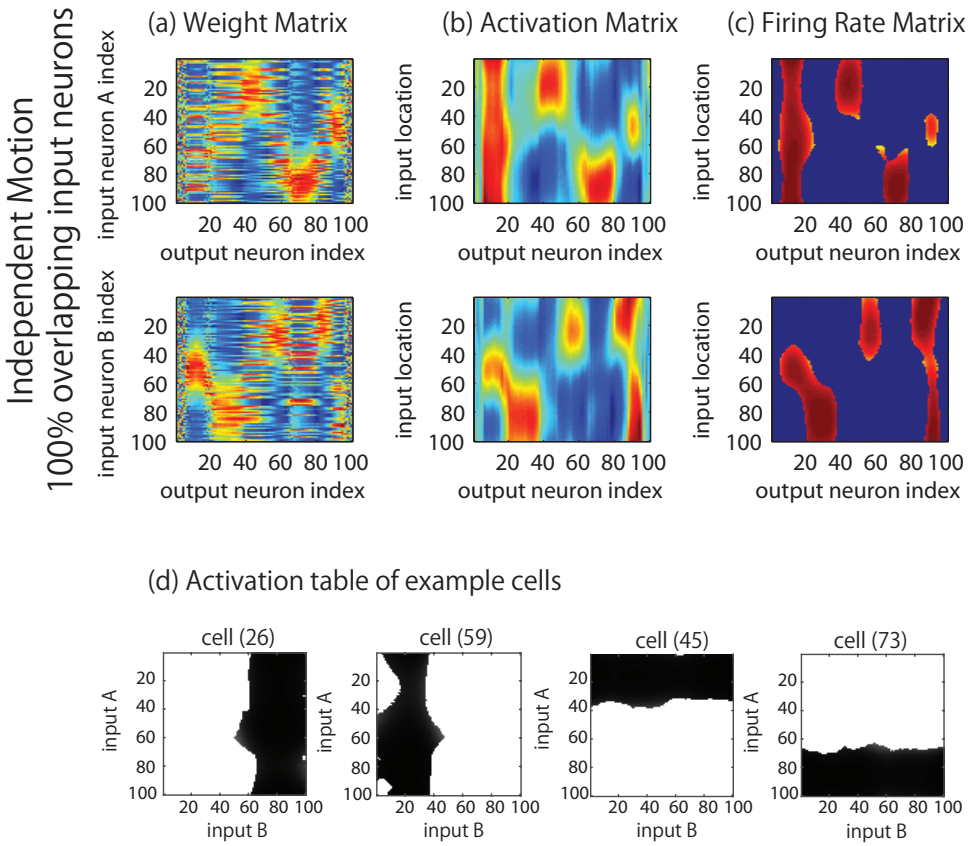

Figure 3.37: Simulation of a one-layer competitive network showing how the *independent* motion of Gaussian activity packets in two input layers during training affects the learned response properties of output neurons. In this simulation there was a 100% overlap between the two input spaces. That is, exactly the same set of input neurons was used to encode both of the two input spaces. Conventions as in Figure 3.34. Even though there is a 100% overlap between the two input spaces, subplots (a), (b) and (c) show that the output neurons still developed separate representations of the two input spaces. The four output neurons shown in (d) have each learned to respond selectively to a localised end region of one of the input spaces.

Furthermore, the VisNet simulations also developed some cells with monotonically increasing or decreasing tuning responses to gradually changing spatial relations between facial features such as inter-eye distance, as has been observed in neurophysiology studies (Freiwald et al., 2009). The question was how such monotonic response properties develop. In complementary simulations of a one-layer competitive network, it was found that the finite receptive field of a neuron due to a topologically restricted fan-in of afferent synaptic connections, as well as the nature of the competition within the output layer, both played important roles in the emergence of neurons with monotonic tuning.

3.5.1 Relationships between the global facial representations and local facial feature representations

It was also found that VisNet developed neurons encoding global facial attributes such as face identity and facial expression as reported in neurophysiology studies (Morin et al., 2014). The question was how different sub-populations of higher layer neurons can learn to respond selectively to either face identity or expression if the network is always exposed to both attributes simultaneously, and the same retinal input neurons represent both global attributes simultaneously in a complex distributed manner. In complementary simulations of a one-layer competitive network, it was shown that the network can develop separate representations of multiple perceptual input spaces such as facial identity and expression even if the input neurons encoding

these spaces are fully overlapping. In particular, this may occur when the input patterns vary independently between the different input spaces. This result provides a possible mechanism for the simultaneous development of multiple global facial representations such as facial identity and expression.

In the main simulation study reported in Study 3a in 3.4.3.1, it was shown that the cell that learned to be selective to happy faces had a higher sensitivity to the shape of the mouth. Interestingly, Gosselin and Schyns (2001) explored the specific visual information humans use to recognize global attributes of faces based on a technique called "bubbles," and they also found that the humans use information around the mouth for expression extraction. Their study has indicated that rather than the facial features which simply have the highest local variance between the considered categories, humans tend to use "partially efficient, not a formal, optimally efficient, feature extraction algorithm" (Gosselin and Schyns, 2001). In the additional simulation study conducted at the end of Study 3a, it was also presented that the Anger neurons (third column) are also receiving strong connections from the mouth. The Sad neurons (second column) are receiving strong inputs from gabor filters representing the shape of the eyes, while Fear neurons (fifth column) and Surprise neurons (sixth column) receive strong inputs from different parts of the eyebrows. These results would provide a predictions about the facial features that might be used for the processing of facial expressions in the brain.

Furthermore, one of the most important arguments raised in this chapter is that the neurons that encode global attributes of faces (such as facial identity and expression) and the neurons that encode a spatial relationship between facial features (such as inter-eye distance) are essentially the same. More specifically, I propose that neurons encoding different global attributes such as expression simply represent different spatial relationships between local features with monotonic tuning curves or particular combinations of these spatial relations. In this way, the population response of a set of facial features would be amplified for extreme compared with intermediate feature values along the visual pathway, and thereby explain why faces with more deviant appearances are recognized better than those which are more typical (Rhodes, 1997; Benson and Perrett, 1991; Bruce and Young, 2011). In particular, this proposal contrasts sharply with the idea of neurons being assigned in an entirely random distributed manner to represent particular facial identities. Instead, neurons encoding facial identity are in fact representing specific structural information about the faces they encode. Our simulation results provide convincing evidence for this argument.

3.5.2 The Representation of Faces and Non-face Objects

Recently, a modelling study carried out by Khaligh-Razavi and Kriegeskorte (2014) demonstrated that a number of unsupervised neural network models developed neuronal representations of faces that were highly correlated compared to the representations of non-face objects. These modelling results mirrored a similar effect found in actual data collected from monkey IT (Kriegeskorte et al., 2008b) and the human temporal lobe (Kiani et al., 2007). On the other hand, Khaligh-Razavi and Kriegeskorte (2014) showed that none of those unsupervised neural network models successfully captures the high correlations in the neuronal responses to non-face objects, which is also present in the brain.

In order to compare our results with those published by Khaligh-Razavi and Kriegeskorte (2014), the activity within the network was analysed by computing *representational dissimilarity matrices* (RDM) (Kriegeskorte et al., 2008a) for each layer of VisNet. Figure 3.38 shows the RDMs computed in response to 50 faces and 50 non-faces for each layer of VisNet before training (left column) and after training (right column). These results show that, after training, the output (4th) layer of the network demonstrates neuronal activity patterns that are highly correlated in response to pairs of stimuli from within one of the stimulus categories, i.e. faces or non-face objects, but are decorrelated in response to stimuli from different categories. It can also be seen that this effect gradually increases through successive neuronal layers of the

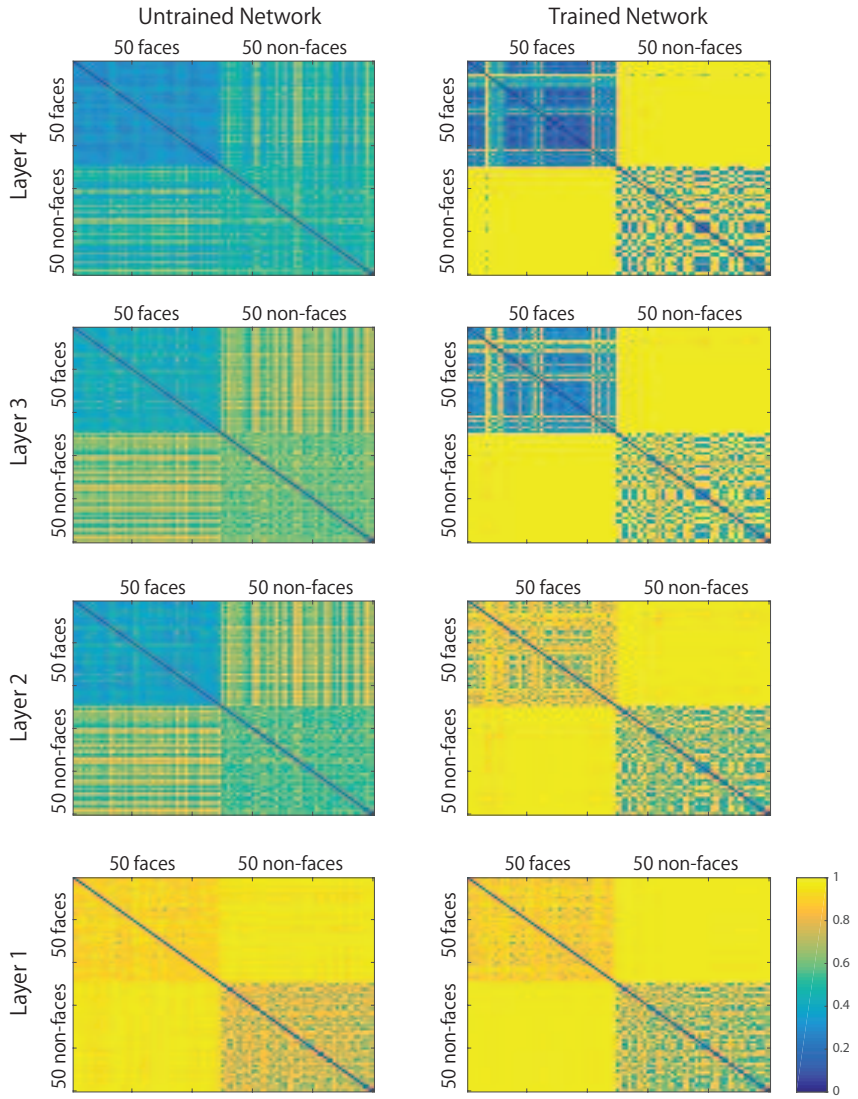

Figure 3.38: Representational dissimilarity matrices (RDM) (Kriegeskorte et al., 2008a) showing the correlations in network activity in response to 50 faces (Figure 3.10) and 50 non-faces (Figure 3.11) for each layer of VisNet before training (left column) and after training (right column). For each layer, the responses of all 128×128 neurons in response to each of the 100 test images were recorded. The Pearson correlations between the vectors of neuronal responses across the layer to each pair of test images were then computed. A representational dissimilarity matrix was then constructed for each layer where each element corresponding to a particular pair of test images was computed as 1 - the Pearson correlation. These results show that, after training, the output (4th) layer of the network demonstrates neuronal activity patterns that are highly correlated in response to pairs of stimuli from within one of the stimulus categories, i.e. faces or non-face objects, but are decorrelated in response to stimuli from different categories. It can be seen that this effect gradually increases through successive neuronal layers of the network.

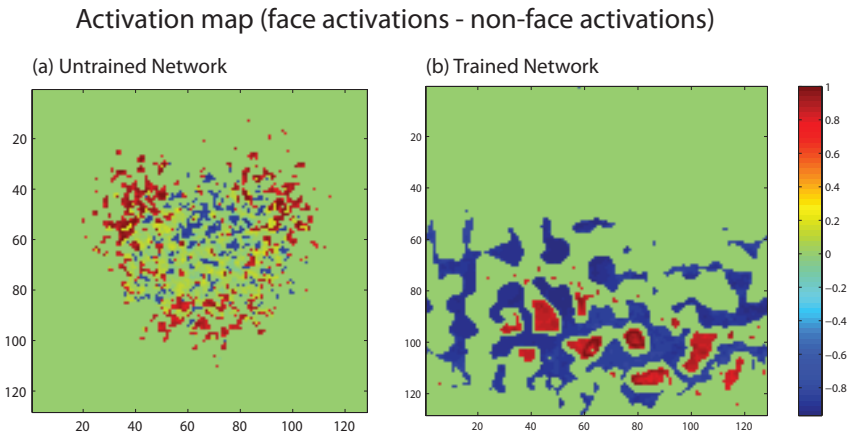

Figure 3.39: Map showing stimulus selectivity of all 4th layer neurons to the faces and non-face objects before training (a) and after training (b). The selectivity measure was computed for all cells that had an average firing rate response greater than or equal to 0.8 for at least one of the stimulus categories. The selectivity measure was calculated by subtracting the average firing rate response of each cell to the non-face objects from the average firing rate response to the faces. The selectivity measure is near +1 (red) for a cell that is selective to faces and near -1 (blue) for a cell that is selective to non-face objects. The selectivity measure was set to zero for those cells with an average firing rate response below 0.8 for both stimulus categories.

network.

This result contradicts VisNet's poor performance reported in Khaligh-Razavi and Kriegeskorte (2014). However, this inconsistency can be explained by the way the network was trained and the size of the network simulated in their study. In Khaligh-Razavi and Kriegeskorte (2014), the model was trained with a trace learning rule over two stimulus categories: 442 'animated images' (faces/bodies of humans/non-humans) and 442 'inanimated images' (natural/artificial objects). In theory, any visual stimulus in the same category should be associated together with the trace learning rule they implemented. This makes the network difficult to develop a representation that is exclusively dedicated to faces. Additionally, the size of the network they simulated was 16 times smaller than the network simulated in the current study, which may also have resulted in limiting the potential of the model. Accordingly, in contrast to the previously reported result in Khaligh-Razavi and Kriegeskorte (2014), our own simulations showed high correlations in the responses of the output layer of VisNet after training with objects from the same stimulus category whether faces or non-face objects. This implies a potential for models to develops similar kind of self-organisation to our brains through feedforward, unsupervised, visually-guided training.

Another important aspect of the visual processing can be found in the cortical structure of the brains. Even though the task of both face and object recognition is achieved along the ventral visual pathway, there is physiological evidence that faces and non-face objects are processed in distinct cortical areas. For example, it has been found that faces are preferentially processed in the occipital face area (OFA) (Pitcher et al., 2011) and several later cortical areas known as 'face patches' (Tsao et al., 2006). These effects were in fact seen in our simulations. When the network was trained on a mixture of faces and non-face objects, it was found that neurons that learned to represent faces tended to be clustered together within localised patches in the output layer, while neurons representing non-face objects were clustered within separate patches. This effect was due to the use of a self-organising map (SOM) architecture implemented within each layer of the model. In this case, the short range excitatory connections between neurons within each layer encouraged nearby neurons to learn to respond to similar stimuli. This was sufficient to lead to separate patches for faces and non-face objects.

Figure 3.39 shows maps of the selectivity of all 4th layer neurons to the faces and non-face objects before and after training. The selectivity measure was calculated for each cell as follows.

First, the average firing rate response $\in (0, 1)$ of the cell to all 150 faces and the average firing rate response to all non-face objects were computed. If the average firing rate response to both categories of stimuli was less than 0.8 then the cell was deemed not responsive enough and the selectivity measure was set to zero. If the average firing rate response of the cell was greater than or equal to 0.8 for at least one of the stimulus categories then the selectivity measure was calculated by subtracting the average firing rate response to the non-face objects from the average firing rate response to the faces. Thus, the selectivity measure has a value near +1 (red) for a cell that is selective to faces and a value near -1 (blue) for a cell that is selective to non-face objects. Figure 3.39 shows that layer 4 developed large distinct patches of neurons that were selective for either faces or non-face objects after training. As mentioned above, this map like structure is reminiscent of the localised regions of face selectivity, called face patches, reported in neurophysiology studies (Gross et al., 1972).

Figure 3.39 shows some evidence of spatially structured selectivity to the two stimulus categories in the untrained network. However, this is simply due to neurons in different regions of layer 4 receiving different amounts of connectivity from different regions of the retina because of the topological feedforward connectivity through the layers. In this case, neurons in the centre of layer 4 may be driven more by the relatively rich visual structure at the centre of non-face objects, while the neurons surrounding the centre of layer 4 may be driven more by peripheral facial features such as the facial outline, etc. This effect is also artifactual because it only arises due to the face and non-face objects not being shown in different retinal locations. However, the strengths of the feedforward connections are extensively modified during visually guided training, leading to the development of selectivity maps with relatively large, contiguous patches of stimulus selectivity, as seen in the primate visual system.

Even though faces and non-face objects are processed in different visual areas, does this mean that the underlying nature of visual processing is different for these two stimulus categories? It was originally suggested by Biederman and Kalocsai (1997) that the processing of faces is unlike that of other objects. These authors argued that the retinal image of an object is decomposed into simple 3D primitives called geons as well as the spatial relationships between these primitives (Biederman, 1987). Such a structural description is viewpoint-independent. On the other hand, the information required for face recognition is proposed to be more holistic, which may be coded as a form of graph with each node representing a particular facial feature and each link representing a relationship between features (Lades et al., 1993; Biederman and Kalocsai, 1997).

However, it has later been argued that the basic visual processing of faces and non-face objects may in fact be similar (Mangini and Biederman, 2004; Yue et al., 2006). In particular, the different psychophysical behaviour towards faces and non-face objects may naturally arise if the representation of faces retains aspects of the original spatial filter representation. In particular, they proposed that larger receptive fields which partially overlap with each other would provide sufficient information to produce the sensitivity to the layout and the spacing of nameable parts of the face (configural effects). A recent simulation study conducted by Xu et al. (2014) supports this hypothesis and concluded that "the configural effect is largely a function of the overlap in the encoding of multiple face features allowed with large receptive fields".

The simulations reported in this chapter use the VisNet architecture originally developed by Wallis and Rolls (1997). This model uses a biologically plausible architecture, with unsupervised learning mediated by local, associative learning rules. Wallis and Rolls (1997) proposed that the learning principles underlying the processing of faces and non-face objects are similar, and used VisNet to model the development of transform invariant representations of both faces and non-face objects. These authors hypothesised that it would be possible for the model to develop detailed representations of the local features of faces if more neurons were incorporated into the model. I have now verified this prediction in a network with 16 times as many neurons as that originally used by Wallis and Rolls (1997).

3.5.3 Limitations and Future Directions

The simulations reported in this chapter used face images constructed using the FaceGen 3D face modelling software. In future work, I plan to replicate these studies using real faces. This will introduce new problems such as how the network achieves correspondence of the same facial features, such as the eyes or nose, across very differently shaped faces. I anticipate that this may require training VisNet on faces in different retinal locations and different scales to achieve representations of facial features that are both location and scale invariant. This will be a significantly more challenging task than with the controlled artificial FaceGen images used in the current study.

The version of the VisNet architecture used in this chapter incorporated only bottom-up (feedforward) connections between successive layers of the network. No top-down connections were included in the model even though these are known to exist in the primate ventral visual pathway and are proposed to have a role in matching incoming inputs with top-down expectations or predictions (Clark, 2013). Nevertheless, the rationale for using this simplified architecture in the current study was that it is sufficient to replicate how neurons in face areas are able to learn to encode the various kinds of face related information.

Another thing to be mentioned is that the current study proposes theories that may explain the development of various neuronal properties that are localized along the ventral visual system, which may be mapped onto the 'core system' in the model proposed by Haxby et al. (2000); however, our model does not explicitly model cognitive processes, which is achieved in the 'extended system' to act in concert with the regions of the core system to extract meaning from faces (Haxby et al., 2000). Therefore, even though our results are compared with physiological data, such as face feature space representations (Freiwald et al., 2009) and global representations of identity and expression (Hasselmo et al., 1989a; Morin et al., 2014), the model does not generate the behaviours associated with the extracted information.

I believe that such behaviour can be implemented with architectural extensions to the current model that may more accurately reflect the known neuroanatomy of the relevant brain areas. For example, Rolls and Treves (1998) have previously hypothesised that pattern association learning may operate in the feedforward connections from area TE at the end of the ventral visual pathway, which represents faces and other visual objects, to areas such as the amygdala and orbitofrontal cortex (OFC). Consistent with this theory, single unit recording studies have shown that neurons in the amygdala and OFC learn associations between visual stimuli (conditioned stimuli) and the corresponding tastes (unconditioned stimuli). Therefore, pattern association appears to operate in these brain areas, which are thought to be involved in the evaluation of the emotional valence of visual stimuli. This would be a useful extension of the model in order to compare the simulated results with human performance on a human categorization task.

Nevertheless, as Wallis (2013) explains, the work presented in this chapter also "serves to explain how such a core system would operate, in terms of its adaptive encoding of objects of expertise, but not how these other systems come to extract information from it to solve specific tasks." For example, Yankouskaya et al. (2014) has recently reported that the level of integration of identity and emotion cues in faces may be determined by life experience and exposure to individuals of different ethnicities. This is consistent with the finding reported in Wallis (2013) that showed that the network trained on Caucasian faces exhibits less sensitivity to changes in appearance of Japanese faces than those of Caucasian faces. I have also shown that after the exposure to 450 faces with randomly generated identities and expressions, many cells became sensitive to changes of identity and expression in a target face. Moreover, the fact that some cells in our model became sensitive to both of these attributes is consistent with the physiological evidence provided by Morin et al. (2014). Such neuronal representations developed in the self-organizing models provide important information to the 'external system' to generate the perceptions and behaviour reported in cognitive experiments (Yankouskaya et al., 2014). Accordingly, this chapter investigated the developmental process of various kinds

of such 'structural codes' (Bruce and Young, 1986), which may set the necessary foundation to achieve face perception in the later stages.

Chapter 4

The Neural Basis of Cognitive Bias Modification as a Clinical Treatment for Depression

Many mental health problems are linked to cognitive biases towards emotionally negative information. For example, depressed patients have a greater tendency to interpret faces as sad and are less able to detect mildly happy expressions. Recently, interest has grown in a new class of psychological treatments for depression, anxiety, and addictive disorders known as Cognitive Bias Modification (CBM), which can eliminate these underlying negative cognitive biases. It is thought that the elimination of negative cognitive biases may help to shift the depressed mood state of a patient. In this chapter, I use computer simulation to investigate the neural and synaptic dynamics underlying two new predicted forms of CBM, which may be able to eliminate the negative biases in the way that depressed patients evaluate facial expressions. The new CBM methodologies utilise two previously established biologically plausible synaptic learning mechanisms, continuous transformation (CT) learning and trace learning (see Section 1.4). These learning mechanisms are able to guide visual development by exploiting either the spatial continuity or temporal continuity between visual stimuli presented during training. Our simulation results show that both of these learning mechanisms, when combined with carefully designed sequences of transforming face images presented to the model, will eliminate negative biases in the interpretation of facial expression. That is, a sub-population of 'sad' output neurons that initially responds to both sad and neutral faces before learning will only respond to the sad faces after CBM training. I first describe simulations with a simplified one-layer neural network architecture in order to test the two hypothesised CBM learning mechanisms in a highly controlled manner. Then I present simulation results in which realistic face stimuli are used to train a more biologically detailed multi-layer neural network computer model, VisNet, of the ventral visual pathway in the primate brain.

4.1 Introduction

Depression is the most common mental health problem, affecting 8 - 12% of the adult population (Ustn et al., 2004). It can lead to a significant reduction in the quality of life for sufferers and in extreme cases may lead to suicide. It has been related to a number of chronic diseases such as coronary heart disease (Rugulies, 2002; Schneider and Moyer, 2010), and has damaging long term effects on health and well-being. Furthermore, depression within the UK population places a significant burden on psychiatric health services and impacts negatively on the economy due to reduced productivity. The economic cost of depression in England in the year 2000 was estimated at £9 billions (Thomas and Morris, 2003), with the biggest impact on workplace

productivity. The latest WHO report shows that anxiety and depression leads to a loss of millions of work days (Jones, 2016); indeed the most common cause of absenteeism from work in the UK is self-reported depression (Almond and Healey, 2003). The economic impact of depression in the United States is estimated to be in the tens of billions of dollars (Wang et al., 2003). Consequently, it is of huge importance to discover new more effective treatments for depression.

A common finding in both clinical depression and anxiety is a link to cognitive biases in processing towards emotionally negative information, with patients tending to pay attention to negative stimuli, interpret events negatively, and recall negative memories (Mathews and MacLeod, 2005; Roiser et al., 2012). These biases therefore have been included within cognitive models of depression (Beck, 2008) and anxiety (Mathews and MacLeod, 2005), leading to a growing interest in exploring the causal relationship between these biases, mood states and clinical symptoms. A series of experimental tools have been developed to assess this relationship. For example, a typical approach has been to present ambiguous information and then ask participants to make a judgement on a positively or negatively valanced probe word (for example, whether or not the word is grammatically legitimate) (Hirsch and Mathews, 1997; MacLeod and Cohen, 1993; Richards and French, 1992). Anxious patients are usually found to have quicker and/or more accurate responses to negative probes.

4.1.1 Cognitive Bias Modification (CBM)

It is thought that the elimination of negative cognitive biases may help to shift the depressed mood state of a patient and reduce anxiety. This led many researchers to recognise the clinical potential of these tools, inspiring the development of a family of potential treatments known as Cognitive Bias Modification (CBM) (MacLeod and Mathews, 2012; MacLeod, 2012). CBM seeks to eliminate these underlying processing biases, with three main varieties of treatment:CBM-Attention (CBM-A), CBM-Interpretation (CBM-I), and CBM-Memory (CBM-M). CBM-A seeks to shift the attention of subjects away from negative stimuli in the environment (MacLeod et al., 2002; Hakamata et al., 2010), CBM-I aims to reduce the tendency for negative interpretation of events (Grey and Mathews, 2009, 2000), and CBM-M seeks to reduce the recall and influence of negative memories (Anderson and Green, 2001; Joormann et al., 2005). In this chapter, the aim is to advance understanding of how CBM might alter biases through the application of neural network modelling, focusing on CBM-I.

One specific example of a negative bias found in clinical disorders has been in the interpretation of facial expressions. For example, depressed patients evaluate facial expressions as being more sad, compared with the evaluations of healthy individuals (Bourke et al., 2010a). Similarly, high-trait anxious individuals are more likely to classify neutral facial expressions as fearful (Richards et al., 2002; Surcinelli et al., 2006). It has been found that CBM intervention can change the bias in emotion recognition of facial expressions, giving a measurable therapeutic outcome (Penton-Voak et al., 2013). In this study, faces were morphed from unambiguously happy to unambiguously angry to give 15 total stimuli. Participants were asked to rate each randomly presented face as either happy or angry, giving a baseline for each participant's emotion recognition along the spectrum of morphs. A balance point was therefore determined, at which participants switched from a categorisation of happy to a categorisation of angry. A CBM training procedure followed in which the previous procedure was repeated, but participants were also given feedback about whether their decision was 'correct' or 'incorrect'. Correct responses were defined as the responses they had previously given in the baseline phase, but with the balance point shifted so that two more faces should now be classified as happy. A final testing phase showed that feedback had shifted participants' balance point in the direction of training. Furthermore, when conducted on adolescents in a youth program who were at high risk for committing a crime, a reduction was also found in self-rated aggression and staff-rated aggression two weeks after training.

Similarly promising results have been found using other CBM-I paradigms, usually involving language-based scenarios, where training a particular interpretation often produces congruent changes in anxiety or affect (e.g. Grey and Mathews, 2000; Mathews and Mackintosh, 2000). Indeed, some studies have found that with variations on this paradigm, CBM can be used as a "cognitive vaccine" to protect against markers of relapse risk in depression (Browning et al., 2012; Holmes et al., 2009). However, CBM is not without controversy. CBM-A in particular has suffered from a number of recent studies failing to find any clinical benefits at all (e.g. Carlbring et al., 2012; Julian et al., 2012), leading some researchers to denounce its potential as a clinical tool entirely (Emmelkamp, 2012). Whilst meta-analyses on CBM-I have shown slightly more promising results than CBM-A, it has remained unclear whether this is due to a time-lag bias - CBM-I is a newer procedure and therefore potentially more likely to have positive results due to studies being less well-controlled (Cristea et al., 2015; Hallion and Ruscio, 2011).

Despite this, many researchers feel it is too early to give up on CBM research just yet. It has been less than two decades since the seminal CBM studies, meaning the field is still in its early stages (Grey and Mathews, 2000; MacLeod et al., 2002). A recent commentary describes the problem with current CBM research as a lack of focus on reliably changing the underlying cognitive biases (Fox et al., 2014). They argue that the theoretical assumption behind CBM is the role of negative biases in maintaining clinical symptoms. Any procedure that does not successfully change a bias cannot be expected to give a clinical benefit. Indeed, a study working from the same premise found that when a bias change is achieved, so is the change in clinical symptom (Clarke et al., 2014). When the study fails to change a bias, then the clinical change is also not found.

Seemingly therefore, future CBM research needs to investigate the mechanisms behind changing cognitive biases, in order to optimise bias-change procedures. Some explanation for this can be given by neuropharmacology, in which it has been found that the action of antidepressants can reduce negative affective biases in depressed patients (Harmer et al., 2009), as well as modify the neural processing of nonconscious threat cues (Harmer et al., 2006). These studies suggest that antidepressants may operate, at least in part, by altering underlying negative biases in information processing in a similar way to psychological treatments such as CBM. The common mechanism shared between these psychological and pharmacological approaches is hypothesised to be the shifting of underlying negative patterns of information processing in the brain. From a theoretical perspective, this implies modifying synaptic connections between neurons in order to adjust the flow of electrical signals through the brain. This must be done through synaptic plasticity, which is either guided by the form of the CBM treatment used or shaped by the action of antidepressants. Effective application of CBM, therefore, could potentially offer a low-cost and non-invasive treatment, particularly if used in combination with other therapies (e.g., Cognitive behavioural therapy (CBT)).

4.1.2 Modelling Study

Computational modelling is one way to investigate the clinical impact of CBM methodologies. For example, Frewen et al. (2008) has replicated the attentional bias towards threatening stimuli, which is a common symptom of highly anxious patients. In the modelling study, the authors used two output units representing whether the network was attending to either the left or right of its attentional field. Furthermore, a number of more abstract modelling studies have investigated the possibility of restructuring the synaptic connectivity within neural network architectures and thereby reshaping memories. For example, modelling studies with attractor neural networks with associatively modifiable recurrent connections (Blumenfeld et al., 2006; Bernacchia and Amit, 2007), as well as psychophysical studies (Preminger et al., 2007, 2009; Wallis and Blthoff, 2001), have shown that different memories can be merged together when a continuum of intermediate stimuli are presented during further training.

The current study investigates the effects of CBM-I through neural network computer mod-

elling in order to understand how CBM might work from a neurobiological perspective. More precisely, I investigated the underlying plasticity mechanisms and emergent neural dynamics using competitive neural networks, which are unsupervised in that no given activity pattern is imposed on the output neurons during training. In other words, the learning in our model is solely guided by the suitable input patterns.

In this chapter, computer simulations are presented to explore two possible CBM-I training methodologies for rewriting previously learned associations. I refer back to the work of Bourke et al. (2010b), aiming to change a negative interpretation of facial expressions into a positive interpretation. The new CBM methodologies utilise two previously established biologically plausible synaptic learning mechanisms known as *continuous transformation (CT) learning* (Stringer et al., 2006) (see Section 1.4.1) and *trace learning* (Foldiak, 1991; Wallis and Rolls, 1997) (see Section 1.4.2). These learning mechanisms are able to guide visual development by exploiting either the spatial continuity or temporal continuity between visual stimuli presented during training. The aim is to explore whether both of these learning mechanisms, when combined with carefully designed sequences of transforming face images presented to the model, will eliminate negative biases in the interpretation of facial expression.

4.2 Hypothesis

4.2.1 Continuous Transformation Learning

It has been reported that people learn to associate visually similar images together. In an experimental study, Preminger et al. (2007) trained subjects to classify faces into two categories: friends (F) and non-friends (NF). Upon reaching good performance, subjects were then trained with a sequence of morphed images from F to NF. The subjects were tested on how they classified the morphed images. Initially, the first half of the morphed image sequence was classified as F, while the second half of the morphed sequence was classified as NF. However, as training progressed, the separation threshold moved towards NF; that is, an increasing number of frames were classified as F. Eventually, all morphed frames were classified as F.

Continuous transformation (CT) learning is an invariance learning mechanism that may provide an insight into the mechanism of such memory reconstruction via ordinary Hebbian learning at the neuronal level (Stringer et al., 2006). It associatively remaps the feedforward connections between successive neural layers while keeping the same initial set of output neurons activated as the input patterns are gradually changed. Consider a set of stimuli that can be arranged into a continuum, in which each successive stimulus in the continuum has a degree of overlap – a number of features in common – with the previous stimulus in the continuum. CT learning can exploit this feature overlap between successive stimuli to form a single percept of all, or at least a large subset, of the stimuli in the stimulus set.

Specifically, when an output neuron responds to one of the input patterns, the feedforward connections from the active input neurons to the active output neuron are strengthened by associative (Hebbian) learning. Then, when the next similar (overlapping) input pattern is presented, the same output neuron is again activated due to the previously strengthened connections. Now the second input pattern is associated with the same output neuron through further associative learning. This process can continue to map a sequence of many gradually transforming input patterns, where each input pattern has a degree of spatial overlap with its neighbours, onto the same output neuron. The standard Hebbian learning rule used to modify the feedforward synaptic connections at each timestep τ is described in Equation (1.7). To prevent the same few neurons always winning the competition, the synaptic weight vector of each output neuron i is renormalized to unit length after each learning update for each training pattern with Equation (1.10) and (1.11).

I hypothesised that this CT learning will eliminate negative biases in the interpretation of

facial expression when combined with carefully designed sequences of transforming face images presented to the model. More precisely, it is hypothesised that this CT learning will eventually shift the critical point of the categorical perception through training so that more neutral faces will be seen as happy faces rather than sad faces. In particular, I will exploit the remapping capabilities of CT learning by morphing very happy faces, which are associated with a positive output representation, into neutral faces during training. This may cause the strong efferent connections from the neutral faces to be remapped to the positive output representation by associative learning operating in the feedforward connections. This should result in positive output neurons firing to both positive (happy) and neutral faces, and negative output neurons only firing to negative (sad) faces.

4.2.2 Trace Learning

Other psychological studies have shown that sequential presentation of the different views of an object, which produces temporal continuity, can facilitate view invariant object learning, where the different views of an object occurring close together in time are bound onto the same output representation (Perry et al., 2006). In contrast, systematically switching the identity of a visual object during such sequential presentation impairs position-invariant representations (Cox et al., 2005). Li and DiCarlo (2008) have reported a neuronal evidence of similar temporal association of visual objects that are presented close together in time. In their study, monkeys were first trained to track an object that was shifted around on a screen. In the experimental condition, the target object was swapped to a different object when the object was at a particular retinal location for the monkeys. As a result, individual neurons in IT that were originally selective to the target object started to respond also to the different object at the specific retinal location. These results show that the temporal statistics of object presentations should play a key role in the development of transform-invariant object representations in the visual brain.

Trace learning is a biologically plausible mechanism to achieve such temporal association by incorporating a memory trace of recent neuronal activity into the learning rule used to modify the feedforward synaptic connections (Foldiak, 1991; Wallis and Rolls, 1997). This encourages output neurons to learn to respond to input patterns that occur close together in time. Stimuli that are experienced close together in time are likely to be strongly related; for instance, successive stimuli could be different views of the same object. If a mechanism exists to associate together stimuli that tend to occur close together in time, then a network will learn that those stimuli form a single percept. Trace learning provides one such mechanism by incorporating a temporal memory trace of postsynaptic cell activity $\bar{r}_i$ into a standard Hebbian learning rule as described in Equation (1.8) and (1.9).For the simulations described below, η was set to 0.8. The synaptic weight vector of each output neuron i is renormalized to unit length according to Equation (1.10) and (1.11) after each learning update for each training pattern.

I propose that such innate trace learning mechanisms may also be exploited to eliminate negative biases in the interpretation of facial expression when combined with carefully designed sequences of transforming face images presented to the model. In particular, if during training with a trace learning rule, a neutral face is presented in temporal proximity with many other very happy faces that are associated with a positive output representation, then this should encourage these positive output neurons to learn to respond to the neutral face as well. When the neutral face is subsequently presented, the positive output representation should suppress the negative output representation by competition mediated by inhibitory interneurons. By implementing a trace learning rule and presenting the network with occasional neutral faces amongst many happy faces, it is expected to see positive output neurons learning to respond to both positive and neutral faces.

4.2.3 Overview of Simulation Studies Carried Out in this chapter

Simulations with a simplified one-layer neural network architecture were first described in order to test the two hypothesised CBM learning mechanisms in a highly controlled manner in Section 4.3 (Studies 1). Then, simulation results in which realistic face stimuli are used to train VisNet (Wallis and Rolls, 1997) are presented in Section 4.4 (Studies 2).

In both sections, I extend these previous modelling studies involving synaptic plasticity and learning to the problem of understanding the neurobiological basis of CBM training by both CT learning (Studies 1a and 2a) and trace learning (Studies 1b and 2b). Specifically, I show that both of these learning mechanisms can be used to eliminate negative biases in the interpretation of facial expression. That is, a subpopulation of 'sad' output neurons that initially responds to both sad and neutral faces before learning will only respond to the sad faces after CBM training. On the other hand, a subpopulation of 'happy' output neurons that initially responds to just happy faces before learning will respond to both happy and neutral faces after training.

4.3 Simulation Studies 1: One-Layer Network

In this section, the aim is to demonstrate how CT learning and trace learning may each be used to carry out CBM within a one-layer competitive neural network. These simulations used a highly idealised network architecture and input stimulus representations in order to provide a very controlled way of investigating and testing the underlying computational hypotheses described in Section 4.2.

In particular, I show how the responses of a one-layer competitive neural network may be remapped, through CBM training, from a negatively biased state to an unbiased state. The remapping using CT learning is demonstrated first in Study 1a in Section 4.3.3, then the remapping using trace learning is demonstrated in Study 1b in Section 4.3.4.

4.3.1 One-Layer Model Description

The network architecture and activation equations are common to the models described in Section 4.3.3 (Study 1a) and in Section 4.3.4 (Study 1b). The network, depicted in Figure 4.1a, comprises a single layer of input cells which drive activity in a layer of two output cells through feedforward synapses. The output neurons compete with each other so that only one such neuron can remain active at a time when an input pattern is presented to the network. In the brain, such competition between neurons within a layer is implemented by inhibitory interneurons.

This architecture is described as a one-layer network because there is only a single layer of synapses in the model. The 1-dimensional layer of input cells provide a highly idealised representation of facial expressions ranging continuously from happy to sad. In the simulations, the input neurons have binarised (0/1) firing rates. Each input neuron responds selectively to a small localised region of the unidimensional space of facial expressions, with the entire space of expressions from happy to sad covered by the input layer. Consequently, the input layer represents each facial expression of a particular emotional valence by the co-activation of a localised cluster of input neurons at the corresponding position within the layer.

At the beginning of the simulation, the feedforward synaptic connection weights are initialised such that the left output cell (happy output cell) responds to happy stimuli, and the right output cell (sad output cell) responds to sad stimuli. A negative cognitive bias can be introduced in the network by initialising the synaptic connections such that the more neutral input stimuli are initially responded to by the sad output neuron rather than the happy output neuron. Then, by modifying the strengths of the feedforward synaptic weights from the input cells to the output cells through CBM training, it is possible to alter the response characteristics of the output neurons in the network. In particular, it will be shown that CBM training

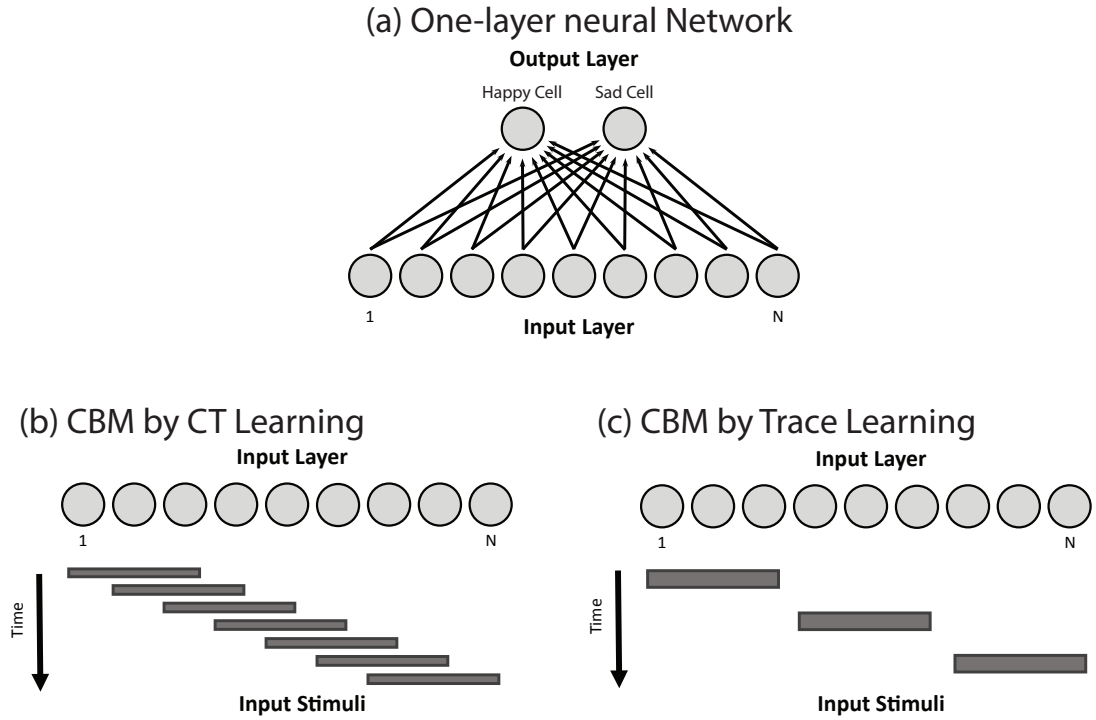

Figure 4.1: **(a)** The one-layer neural network architecture used in the models described in Section 4.3 (Study 1). A single layer of N input cells drives activity in the two output cells through the feedforward synapses (black arrows). The input layer cells respond to stimuli that range from happy to sad, with different simulations requiring different numbers of input cells, as detailed in Section 4.3.3 (Study 1a) and 4.3.4 (Study 1b). There are always two output cells in the network, with the left output cell responding to happy stimuli, and the right output cell responding to sad stimuli. **(b)** The training protocol for the one-layer network trained by CT learning. The input layer contains a total of $N = 600$ neurons. During each training step, the current input stimulus is represented by the firing rates of a contiguous subblock of input cells being set to 1 as illustrated by the horizontal grey lines. The length of the stimulus is referred to as its *stride*, which was set to be 100 input neurons. The firing rates of all other input cells are set to 0. At each successive training step, the input stimulus is advanced by 1 input cell in order to ensure that successive stimuli are varied in a continuous manner, i.e. successive stimuli overlap with each other, which is a requirement of CT learning. During each training epoch, the input stimili are shifted once through the whole continuum from happy to sad. **(c)** The training protocol for the one-layer network trained by trace learning. The input layer contains a total of $N = 900$ neurons. During each training step, the current input stimulus is represented by the firing rates of a contiguous subblock of input cells being set to 1 as illustrated by the horizontal grey lines. The length of each stimulus, its stride, was set to be 100 input neurons. The firing rates of all other input cells are set to 0. In order to prevent CT-like learning effects from occurring, the input stimuli do not overlap with each other. During training, the most happy input stimuli are closely interleaved with more neutral input stimuli from the middle of the stimulus range, while the most sad stimuli are shown without temporal interleaving with the neutral stimuli. This stimulus presentation order enables trace learning to associate together the happy and neutral stimuli onto the same happy output cell.

by either CT learning or trace learning can shift the network away from a negative bias to a situation in which the happy output cell responds to the majority of the input stimuli including both happy and more neutral stimuli.

At each timestep during simulation of the network, an input stimulus of a particular emotional valence is selected to be presented to the network. During CBM training, the input stimuli are presented in accordance with the spatio-temporal statistics required by either CT learning or trace learning, as described in Section 4.3.3 (Study 1a) and in Section 4.3.4 (Study 1b) respectively. Then the input cell firing rates, r_j, are set to be either 0 or 1 according to the training and testing protocols described in Section 4.3.3.1 (Study 1a) and in Section 4.3.4.1 (Study 1b). The output cell firing rates, r_i, are calculated by setting the activation level, h_i, of each output cell i based on the Equation (1.3).

The output cell firing rates are then set by applying winner-take-all inhibition so that the output cell with the highest activation level is given a firing rate of 1 and the other output cell is given a firing rate of 0. During CBM training, after the firing rates of the output cells have been computed, the synaptic weights are then updated by either the Hebbian learning rule (1.7)

in Study 1a or the trace learning rule (1.8) and (1.9) in Study 1b.

4.3.2 Initial Setup of the Network

Before the network undergoes CBM training, the feedforward synaptic weights to the sad and happy output cells are set manually to control whether or not there is a pre-existing cognitive bias.

In order to establish the synaptic connectivity without an initial bias, the synaptic weights to the sad output cell, $w_{\mathrm{SAD}j}$, are set so that

$$w_{\mathrm{SAD}j} = \frac{1}{1 + \exp(-2\beta(\epsilon_j - \alpha))} \tag{4.1}$$

The parameter $\epsilon_j \in [-3, +3]$ represents the preferred stimulus location of input cell j within the sad to happy continuum, with most sad =-3 and most happy=+3. The input neurons are distributed evenly throughout the sad to happy stimulus continuum. The slope β is set to an appropriate value (described in Table 4.1a), and the threshold α is set to 0. The synaptic weights to the happy output cell, $w_{\mathrm{HAPPY}j}$, are set to be

$$w_{\mathrm{HAPPY}j} = 1 - w_{\mathrm{SAD}j}. \tag{4.2}$$

The effect of setting the weights in this manner is that all input cells send feedforward synaptic weights to both of the output cells, but the sad output cell receives stronger synaptic weights from the input cells representing the sad end of the input continuum, and the happy output cell receives stronger synaptic weights from the input cells representing the happy end of the input continuum. In particular, with $\alpha = 0$, the feedforward synaptic connections are unbiased in that the happy output cell and sad output cell receive mirror-symmetric distributions of afferent synaptic connections covering the entire stimulus space. This can be seen in the left plot of Figure 4.2a for the first simulation with CT learning (Study 1a) and Figure 4.2d for the second simulation with Trace learning (Study 1b).

In order to introduce a negative bias in the synaptic weights such that the sad output cell will also respond to most of the middle, more neutral, portion of the input continuum, the synaptic weights from the input cells to the sad output cell are set according to Equation (4.1) with the threshold α set to a negative value (described in Table 4.1a for Study 1a and Table 4.1b for Study 1b). The synaptic weights from the input cells to the happy output cell are then set according to Equation (4.2). As can be seen in the left plot of Figure 4.2b for the first simulation (Study 1a) and Figure 4.2e for the second simulation (Study 1b), this results in the sad output cell receiving stronger synaptic weights from a greater proportion of the input cells than the happy output cell does.

4.3.3 Study 1a: CBM by CT Learning

In this section, CBM in the one-layer network is simulated by the continuous transformation (CT) learning mechanism described in Section 4.2.1. It associatively remaps the feedforward connections between successive neural layers while keeping the same initial set of output neurons activated as the input patterns are gradually changed. This mechanism will be exploited by morphing happy input stimuli, which are strongly associated with the positive output representation, i.e. the happy output neuron, into more neutral stimuli during training. This causes the efferent connections from the neutral stimuli to be remapped to the positive output representation by associative learning operating in the feedforward connections. When the neutral stimuli are presented again after training, the positive output representation should respond and also suppress the negative output representation by competition mediated by inhibitory interneurons.

Table 4.1: Parameters used in the simulations

Parameter	**Value**			
(a) One-Layer Network (CT)				
No. of Input Cells N	600			
Stride	100			
Sigmoid Slope (β)	0.5			
Biased Sigmoid Threshold (α)	-1			
Learning Rate (k)	0.001			
Training Epochs	100			
(b) One-Layer Network (Trace)				
No. of Input Cells N	9			
Stride	100			
Sigmoid Slope (β)	0.5			
Biased Sigmoid Threshold (α)	-1			
Learning Rate	0.01			
Eta (η)	0.8			
Training Epochs	100			
(c) VisNet				
Gabor: Phase shift (ψ)	$0, \pi$			
Gabor: Wavelength(λ)	2			
Gabor: Orientation(θ)	$0, \pi/4, \pi/2, 3\pi/4$			
Gabor: Spatial bandwidth (b)	1.5 octaves			
Gabor: Aspect ratio (γ)	0.5			
No. of Layers	4			
Retina	$256 \times 256 \times 16$			
	1st layer	**2nd layer**	**3rd layer**	**4th layer**
Dimension	128×128	128×128	128×128	128×128
Num. of fan-in connections	201	100	100	100
Fan-in radius	24	24	36	48
Sparseness of activations	1 %	44 %	32 %	25 %
Sigmoid slope (β)	15	99	146	207
Learning rate (k)	1.0	1.0	1.0	1.0
Training Epochs	20	20	20	20
Excitatory Radius (σ_E)	1.4	1.1	0.8	1.2
Excitatory Contrast (δ_E)	5.35	33.15	117.57	120.12
Inhibitory Radius (σ_I)	4.94	13.88	9.72	14.80
Inhibitory Contrast (δ_I)	1.5	1.5	1.6	1.4

4.3.3.1 Method

Figure 4.1b shows the setup for training the one-layer network with CT learning. The input layer contains a total of $N = 600$ neurons. The layer of input cells represent a continuum of facial expressions from happy (left) to sad (right). The input stimulus presented to the network at any given training step is represented by the firing rates of a contiguous subblock of input cells being set to 1, as illustrated by the horizontal grey lines in Figure 4.1b. The length of the stimulus is referred to as its *stride*, which was set to be 100 input neurons. The firing rates of all other input cells are set to 0. In this simulation with CT learning, the Hebb learning rule (1.7) is used. Since the Hebb learning rule does not contain a memory trace of previous neuronal activity, this ensures that any observed bias modification is the result of CT learning and not the result of trace learning.

During training of the network, illustrated in Figure 4.1b, the input stimulus is moved continuously through the layer of input cells, advancing one input cell per learning update of the network. At each stimulus presentation, the activations of the output neurons are first updated according to Equation (1.3), the firing rates of the output neurons are then computed using winner-take-all competition, and then the feedforward synaptic weights are modified according to Equations (1.7), (1.10), and (1.11). One epoch of training is completed after the input stimulus has been shifted through the whole continuum from happy to sad. Upon reaching the specified number of training epochs, the training phase is finished and the testing phase begins, which follows the same protocol as the training phase with the exception that the weight update and normalization equations, Equations (1.7), (1.10), and (1.11), are not simulated. The simulation is then complete. A one-layer neural network model was simulated with the parameters given in Table 4.1a.

4.3.3.2 Results

First, the network was simulated with the synaptic weights initially hardwired to unbiased values according to Equations (4.1) and (4.2) with the threshold α set to 0. Next, the network was simulated with a negative cognitive bias introduced by hardwiring the synaptic weights according to Equations (4.1) and (4.2) with the threshold α set to -1. This ensured that the sad output cell responded not only to very sad stimuli but also to the majority of the more neutral stimuli. In the final simulation, the negative bias in the previous biased network was eliminated by CBM training using CT learning. This had the effect of remapping the feedforward synaptic weights so that the happy output cell took over responding to the majority of the neutral stimuli.

Untrained Network Performance (Before and After Biases are Added) The network was simulated with the synaptic weights initially hardwired to unbiased values. The left plot of Figure 4.2a shows the unbiased weights from the input cells to the output cells. The sad output cell receives the strongest synaptic weights from the input cells representing the sad end of the stimulus continuum, and the happy output cell receives the strongest synaptic weights from the input cells representing the happy end of the stimulus continuum. The two output cells receive equal, albeit mirror symmetric, distributions of synaptic weights from the input cells representing the middle, more neutral, portion of the stimulus continuum. The right plot of Figure 4.2a shows the firing rates of the two output cells in response to presentation of the input stimuli. The happy output cell responds strongly to very happy input stimuli, the sad output cell responds strongly to very sad input stimuli, and most importantly both output cells respond to equal sized regions of the more neutral intermediate input stimuli. These responses are to be expected given the unbiased feedforward synaptic weight profiles between the input cells and the output cells.

The network was simulated with a negative cognitive bias introduced by hardwiring the synaptic weights. The left plot of Figure 4.2b shows the synaptic weights after a bias has been applied. The sad output cell receives stronger synaptic weights from the sad end of the input range and most of the more neutral input cells, and the happy output cell now receives stronger synaptic weights from only the input cells representing the happy end of the input continuum. The effect of this bias is that the sad output cell now responds not only to very sad stimuli but also to the majority of the more neutral stimuli, whereas the happy output cell does not. This can be seen in the right plot of Figure 4.2b.

Learned (Remapped) Network Performance The negative bias in the previous biased network was eliminated by CBM training using CT learning. After CT learning, the synaptic weights should remap such that the happy output cell now receives stronger synaptic weights from the input cells representing a larger portion of the intermediate, more neutral, stimuli than the sad output cell. The effect of this learned remapping is that the happy output cell responds to a greater proportion of the input stimulus space than the sad output cell does. That is, the happy output cell now responds to the majority of the intermediate neutral stimuli. This can be seen in the right plot of Figure 4.2c (c.f. the right plot of Figure 4.2b). This represents CBM, where the bias in the network has been shifted from negative to positive by CT learning.

4.3.4 Study 1b: CBM by Trace Learning

Having shown how CBM may be accomplished through CT learning, I now show how it may also be accomplished using a different learning paradigm: trace learning. In this section, CBM in the one-layer network is simulated by the trace learning mechanism described in Section 4.2.2. Trace learning is an invariance learning mechanism which utilises a trace learning rule, Equation (1.8) and (1.9) with weight vector normalisation Equation (1.10) and (1.11) to modify the feedforward synaptic connections. Trace learning incorporates a memory trace $\bar{r}_i^{\tau-1}$ of recent

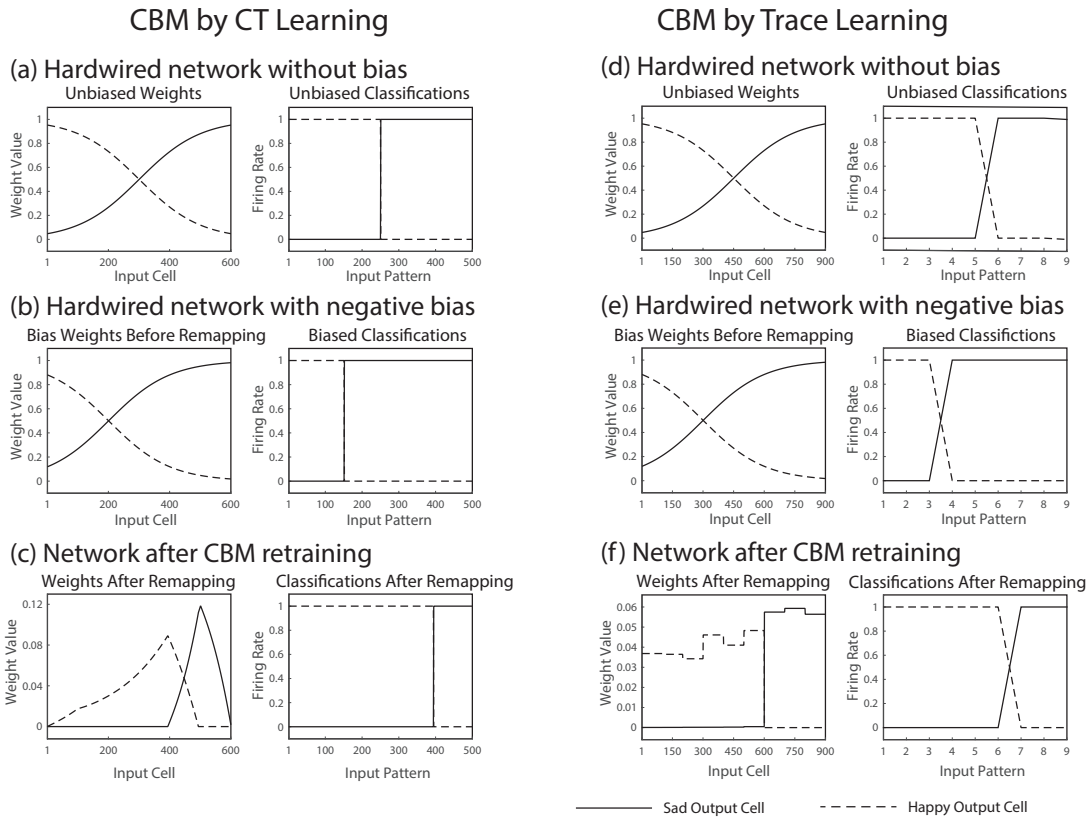

Figure 4.2: Demonstration of CBM in a one-layer network using CT learning (a-c) and trace learning (d-f) to remap the synaptic weights. The figure shows the feedforward synaptic weights (left column) and firing rates of the output cells (right column) at various stages of the simulation. **(a,d)** Results of testing the initial, unbiased hardwired network. The lack of bias in the synaptic weights results in the happy and sad output cells responding to equal numbers of the input patterns. **(b,e)** Results of testing the biased hardwired network. After the negative bias is introduced to the synaptic weights, the sad output cell now responds to the majority of the input patterns. **(c,f)** Results of testing the network after remapping the synaptic weights through CBM training with CT learning (c) and with trace learning (f). The learning has effected a remap in the synaptic weights such that the happy output cell now has stronger synaptic weights from the majority of the input cells. The effect of this remapping is that the happy output cell now responds not only to the most happy stimuli but also to the majority of the more neutral input patterns.

neuronal activity into the learning rule used to modify the feedforward synaptic connections. This encourages output neurons to learn to respond to input patterns that occur close together in time. If, during training, a neutral stimulus is presented in temporal proximity with many other very happy stimuli that are associated with the positive output representation, i.e. the happy output neuron, then this should encourage the positive output representation to respond to the neutral stimulus as well. When the neutral stimulus is subsequently presented, the positive output representation should suppress the negative output representation by competition, which in the brain is mediated by inhibitory interneurons.

4.3.4.1 Method

The setup for training the one-layer network with trace learning is shown in Figure 4.1c. The input layer contains $N = 900$ neurons. The input layer represents a range of facial expressions from happy (left) to sad (right). Each input stimulus shown to the network is represented by the firing rates of a contiguous subblock of input cells being set to 1, as illustrated by the horizontal grey lines in Figure 4.1c. The length of each stimulus presented to the network was set to be 100 input neurons, while the firing rates of all other input cells were set to 0.

In contrast to the training protocol used for the above simulations with CT learning described in Section 4.3.3.1, the input stimuli used for trace learning in this section do not overlap as they advance through the input space. This prevents any CT-like learning effects from oc-

curring, and so ensures that any bias modification that occurs is the result of trace learning and not the result of CT learning. The training protocol with trace learning is shown in Figure 4.1c.

During training of the network, illustrated in Figure 4.1c, the input stimuli are divided into two separate groups: one group containing stimuli from the most happy and more neutral (middle) parts of the input stimulus range; and one group containing only stimuli from the sad end of the input stimulus range. During an epoch of training, one of the two stimulus groups is selected at random. If the stimulus group contains only the sad stimuli, these stimuli are shown to the network in a random order. If the stimulus group contains both the happy and more neutral stimuli, then the happy stimuli are interleaved with the neutral stimuli such that a happy stimulus is shown followed by a neutral stimulus, but with these stimuli paired in a random order. After presentation of the first group of stimuli (happy/neutral, or sad), the second group of stimuli is shown to the network. During the presentation of each stimulus, the activations of the output neurons are updated by Equation (1.3), the firing rates of the output neurons are then computed according to winner-take-all competition, and the synaptic weights are then updated according to the trace learning rule Equation (1.8) and (1.9) with weight vector normalisation Equation (1.10) and (1.11). After all stimuli have been presented, an epoch of training is complete and the next epoch of training begins. The order of the stimulus groups, and the order of stimulus presentation within the group, are randomly selected for each training epoch. Upon reaching the specified number of epochs, the training phase is finished and the testing phase begins, during which the input stimuli are presented one at a time to the network, ranging from happy to sad. The weight update and normalization equations, Equations (1.8), (1.10), and (1.11), are not simulated during the testing phase. After the testing phase, the simulation is complete. A one-layer neural network model was simulated with the parameters given in Table 4.1b.

4.3.4.2 Results

The network was first simulated with the synaptic weights manually set to unbiased values according to Equations (4.1) and (4.2) with $\alpha = 0$. Next, the network was simulated with a negative bias introduced by hardwiring the synaptic weights according to Equations (4.1) and (4.2) with $\alpha = -1$. This caused the sad output neuron to respond to most of the more neutral stimuli in addition to the sad stimuli. Lastly, the negative bias in the previous network was eliminated by CBM training using trace learning. This resulted in the happy output neuron now responding to most of the neutral stimuli as well as the happy stimuli.

Untrained Network Performance (Before and After Biases are Added) The network was simulated with unbiased hardwired synaptic weights. Figure 4.2d (left) shows the unbiased synaptic weights. The sad output cell receives the strongest synaptic weights from the sad end of the stimulus range, while the happy output cell receives the strongest synaptic weights from the happy end of the stimulus range. The two output cells receive equal, albeit mirror symmetric, distributions of synaptic weights from the intermediate neutral portion of the stimulus continuum. Figure 4.2d (right) shows the firing rate responses of the two output cells to the full range of input stimuli. The happy output cell responds to happy stimuli, the sad output cell responds to sad stimuli, while both output cells respond to equal numbers of the more neutral intermediate stimuli.

The network was then simulated with a negative cognitive bias introduced by hardwiring the synaptic weights. Figure 4.2e (left) shows the synaptic weights. The sad output cell receives stronger synaptic weights from the sad end of the input range and most of the more neutral input cells, while the happy output cell receives stronger synaptic weights from only the happy end of the input range. Figure 4.2e (right) shows the firing rate responses of the two output neurons to the the full range of input stimuli. Due to the biased synaptic weights, the sad

output cell responds to the majority of the more neutral stimuli in addition to the sad stimuli, whereas the happy output cell only responds to the more happy stimuli.

Learned (Remapped) Network Performance The negative bias in the previous biased network was eliminated by CBM training using trace learning. After trace learning, the feedforward synaptic weights remap so that the happy output neuron receives stronger synaptic weights from input neurons representing the happy stimuli and the majority of the more neutral stimuli, while the sad output cell receives strong synaptic weights only from the sad end of the input stimulus range. This can be seen in the left plot of Figure 4.2f. The effect of this remapping is that the happy output cell now responds to stimuli from the happy to middle neutral region of the input stimulus range, while the sad output cell responds only to stimuli from the sad end of the input stimulus range, which can be seen in the right plot of Figure 4.2f. Thus, trace learning has produced CBM, where the bias in the network has been shifted from negative to positive.

4.4 Simulation Studies 2: VisNet Simulation

In this section, computational hypotheses described in Section 4.1 is tested using realistic face stimuli presented to VisNet (Wallis and Rolls, 1997; Stringer et al., 2006). The simulations with VisNet were carried out in two stages of training as follows.

In the first training stage, VisNet was trained on a set of randomised computer generated face images, where the identity and expression of each face was chosen randomly. In Chapter 3, it was reported that this led to the development of separate sub-populations of output neurons that responded selectively to either facial identity or expression (Eguchi et al., 2016). Such neurons have been experimentally observed in single unit recording neurophysiology studies on the primate brain (Hasselmo et al., 1989b).

The second stage of training involved CBM by either CT learning or trace learning, similar to that described above for the one-layer network in Section 4.3.3.1 and in Section 4.3.4.1. Specifically, I tested whether the initial negative bias in the synaptic connectivity developed in the pretraining could be shifted from sad to happy after CBM retraining on new, specially designed sequences of face images. In these second stage simulations, the sequences of face images used for CBM retraining were constructed in accordance with the spatio-temporal stimulus statistics required by either the CT learning (Study 2a) or trace learning hypotheses (Study 2b).

4.4.1 VisNet Model Description

The simulation studies presented below are conducted with a hierarchical neural network model of the primate ventral visual pathway, VisNet, which was originally developed by Wallis and Rolls (1997) (see Section 1.3). The values used in the current studies are given in Table 4.1c. The gradual increase in the receptive field of cells in successive layers reflects the known physiology of the primate ventral visual pathway (Freeman and Simoncelli, 2011; Pasupathy, 2006; Pettet and Gilbert, 1992). During training with visual objects, the strengths of the feedforward synaptic connections between successive neuronal layers are modified by biologically plausible local learning rules, where the change in the strength of a synapse depends on the current or recent activities of the pre- and post-synaptic neurons. A variety of such learning rules, in this case both Hebbian learning (Equation (1.7)) and trace learning (Equation (1.8) and (1.9)), may be implemented with different learning properties.

In the experiments conducted in this chapter, an array of Gabor filters (see Section 1.3.1) is generated at each of 256×256 retinal locations with the parameters given in Table 4.1c. The outputs of the Gabor filters are passed to the neurons in layer 1 of VisNet according to

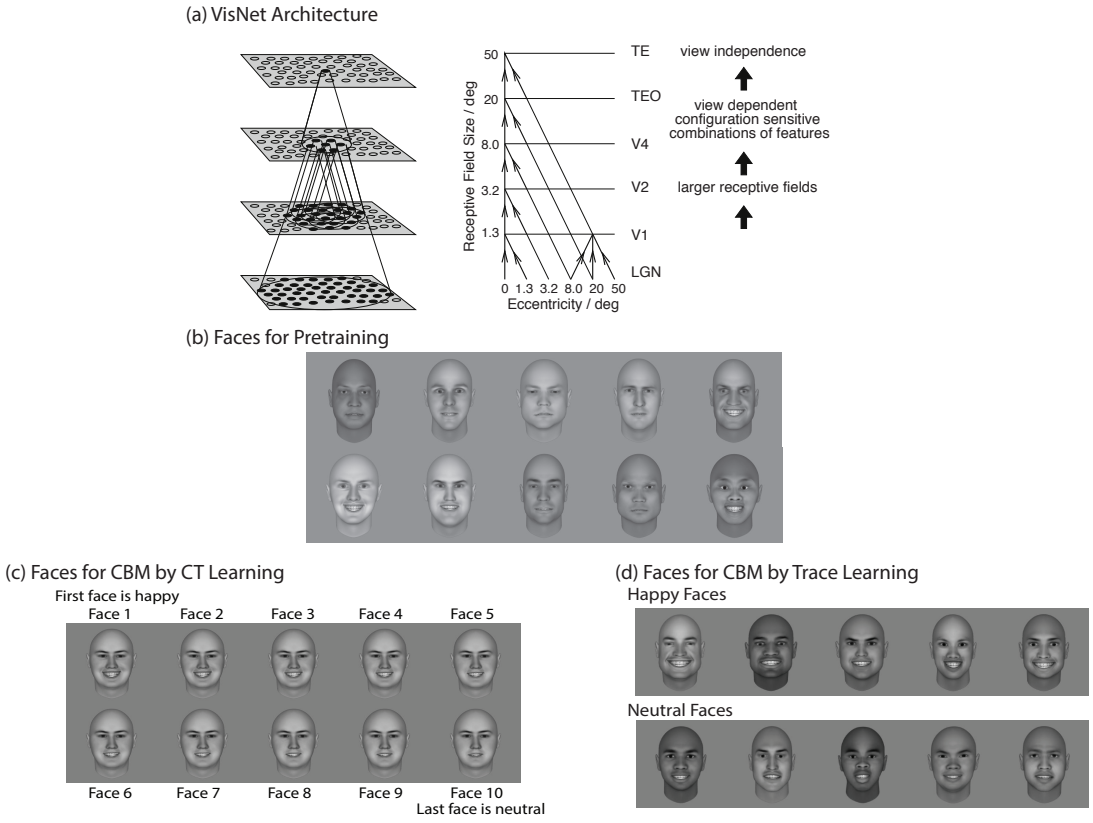

Figure 4.3: **(a)** Left: Stylised image of the four layer VisNet architecture. Convergence through the network is designed to provide fourth layer neurons with information from across the entire input retina. Right: Convergence in the visual system V1: visual cortex area V1; TEO posterior inferior temporal cortex, TE inferior temporal cortex (IT) **(b)** Examples of the face stimuli used to pretrain VisNet. 100 realistic human faces were randomly generated with different identities, and the expressions of individual faces were also randomly set along a continuous dimension between happy and sad. **(c)** Examples of the face stimuli used to perform CBM retraining on VisNet through CT learning with the Hebbian learning rule, Equation (1.7) and weight vector renormalisation, Equation (1.10) and (1.11). The image set is constructed from 5 different facial identities. For each of these facial identities, ten face images were constructed by sampling ten evenly-spaced expressions between happy and neutral. This gave a total of 50 face images used to retrain VisNet. The figure shows a subset of these images corresponding to one particular facial identity morphed through 10 equispaced expressions from happy (top left) to neutral (bottom right) **(d)** Examples of the face stimuli used to perform CBM retraining on VisNet through trace learning with the trace learning rule, Equation (1.8) and (1.9) and weight vector renormalisation, Equation (1.10) and (1.11). The image set consisted of 25 faces with a happy expression and 25 faces with a neutral expression. Each of these 50 faces had a different randomly generated identity. The figure presents some examples of these images. The top row shows a selection of 5 happy faces, while the bottom row shows 5 neutral faces. Faces with happy and neutral expressions were interleaved during CBM retraining by trace learning.

the synaptic connectivity given in Table 4.1c. That is, each layer 1 neuron receives connections from 201 randomly chosen Gabor filters localised within a topologically corresponding region of the retina.

In this chapter, simulations with a self-organising map (SOM) (von der Malsburg, 1973; Kohonen, 1982) implemented within each layer were conducted. In the SOM architecture, short-range excitation and long-range inhibition are combined to form a Mexican-hat spatial profile and is constructed as a difference of two Gaussians as described in Section 1.3.3.2. The lateral inhibition and excitation parameters used in the SOM architecture are given in Table 4.1c.

The parameters for the sigmoid activation function (see Section 1.3.4) are shown in Table 4.1c. These are general robust values found to operate well. They are similar to the standard VisNet sigmoid parameter values that were previously optimised to provide reliable performance (Stringer et al., 2006, 2007; Stringer and Rolls, 2008).

4.4.1.1 Information analysis

A single cell information measure was applied to the trained network of simulation study in Chapter 3 in order to identify the different subpopulations of output (4th layer) neurons that responded selectively to either happy faces or sad faces regardless of facial identity (Eguchi et al., 2016). Full details on the application of this measure to VisNet are given by Rolls and Milward (2000). In particular, the magnitude of the information measure reflects the extent to which a neuron responds selectively to a particular stimulus category such as a happy or sad expression, but also responds invariantly to different examples from that category such as different face identities.

The single cell information measure is applied to individual cells in layer 4, and measures how much information is available from the response of a single cell about which stimulus category, i.e. a happy expression or a sad expression, was shown. For each cell, the single cell information measure used was the maximum amount of information a cell conveyed about any one stimulus category. The stimulus-specific information $I(s, R)$ is the amount of information the set of responses R has about a specific stimulus category s, and is given by equation (1.12) in Section 1.5.2.1.

The maximum amount of information that can be attained is $\log_2(N)$ bits, where N is the number of stimulus categories. For the case of two stimulus categories, i.e. happy and sad expressions, the maximum amount of information is 1 bit.

4.4.2 Pretraining VisNet

In the first stage of the simulations, VisNet was pretrained on a set of 100 randomised computer generated face images, which were created using the software package FaceGen (FaceGen, 2013). FaceGen allows for controlled production of realistic face stimuli, developed from a series of photographs of real people. The faces were randomly generated with different identities, and the expressions of individual faces were also randomly set along a continuous dimension between happy and sad. Examples of these face images are shown in Figure 4.3b.

The pretraining stage was carried out using the Hebbian learning rule (1.7) with weight vector normalisation (1.10) and (1.11). The presentation of the 100 randomised faces constituted one epoch of training, and the network was trained for a total of 20 training epochs during this stage.

The network was then tested by presenting 100 happy faces all with different facial identities, and then presenting 100 sad faces with different facial identities. For each presentation of a face, the firing rates of all of the output neurons were recorded. Information analysis was then used to identify whether any output neurons carried high levels of information about facial expression. That is, whether these neurons had learned to respond to either happy expressions regardless of identity, or sad expressions regardless of identity.

Figure 4.4b shows the single cell information carried by all output (4th layer) neurons before and after pretraining on the randomised face images. The plot shows the information carried by the 4th layer neurons about either happy or sad expressions, where the neurons are plotted in rank order along the abscissa. The maximum amount of information possible for the simulation is $\log_2(N)$ bits where N is the number of categories (Happy or Sad), that is 1 bit. The dashed line represents the untrained network while the solid line represents the trained network. The result shows that pretraining VisNet on many randomly generated faces has significantly increased the amount of single cell information carried by 4th layer neurons about the facial expression as originally reported in Chapter 3 (Eguchi et al., 2016).

These computed information values enabled us to identify two different subpopulations of output neurons that had learned to respond to either happy or sad expressions regardless of facial identity. Figure 4.4c shows the response profiles of five Happy output neurons and five Sad output neurons recorded in response to the matrix of test faces shown in Figure 4.4a

directly after the initial stage of pretraining (solid line). The plots show the average firing rate of the cells in response to 20 different facial expressions ranging from very happy (1) to very sad (20). For each facial expression, the firing rates are averaged over the 20 different facial identities. These neurons have approximately monotonic response profiles, with Happy neurons (top row) responding maximally to the most happy faces and Sad neurons (bottom row) responding maximally to Sad faces, as previously reported in the simulation study in Chapter 3 (Eguchi et al., 2016). Importantly, these neurons were shown to in fact encode particular spatial relationships between the facial features that correlated with facial expression. For a more detailed analysis of the neuronal firing properties that developed during the pretraining stage, please refer to Chapter 3.

In the next sections, I show how to remap the feedforward synaptic connections to these two subpopulations of output neurons by either CT learning or trace learning in order to shift the cognitive bias from negative to positive.

4.4.3 Study 2a: CBM by CT Learning

4.4.3.1 Method

After pretraining VisNet on 100 randomised faces as described in Section 4.4.2, VisNet then underwent a stage of CBM retraining by CT learning. During this, the network was retrained on continuously transforming face images with the Hebbian learning rule (1.7) with weight vector renormalisation (1.10) and (1.11). Figure 4.3c shows examples of the face stimuli used to perform CBM retraining by CT learning. The image set is constructed from 5 different facial identities. For each of these facial identities, ten face images were constructed by sampling ten evenly-spaced expressions between happy and neutral. Figure 4.3c shows a subset of these images corresponding to one particular facial identity morphed through 10 equispaced expressions from happy (top left) to neutral (bottom right). During CBM retraining, the first facial identity was presented transforming continuously through the 10 expressions from happy to neutral. Then the second facial identity was similarly presented transforming continuously through the 10 expressions from happy to neutral. This was repeated for all 5 facial identities in turn. This constituted one epoch of training. The network underwent a total of 50 training epochs.

In this situation, CT learning (Stringer et al., 2006) will begin to remap the feedforward synaptic connections through successive neuronal layers within the network according to the computational hypothesis described in Section 4.2. That is, when the happy face is presented, this stimulates the happy output (4th layer) neurons to respond. Then, as the face is gradually morphed from happy to neutral, the happy output cells continue to respond due to the CT learning mechanism operating in the feedforward synaptic connections between successive layers. At the same time, the later more neutral faces are remapped onto the happy output neurons through the Hebbian learning rule (1.7) with weight vector renormalisation (1.10) and (1.11). This retraining is carried out for each of the 5 different facial identities over 100 training epochs. In this way, the low-level features representing more neutral faces in the lower layers of the network become remapped onto the more happy output representations. Thus, CBM occurs.

I wanted to assess how well CBM retraining remapped the more neutral faces away from the sad output neurons and onto the happy output neurons. In order to do this, it was begun by reanalysing the amount of information that individual output neurons carried about either happy or sad expressions directly before the CBM retraining stage. Specifically, the subset of 1,000 neurons that carried the most information about the presence of a happy expression, and another subset of 1,000 neurons that carried the most information about the presence of a sad expression were identified . In this way, two separate subsets of output neurons (i.e. Happy vs Sad subpopulations) were identified. The performance of the CBM retraining was assessed by recording and analysing the firing rates of the Happy and Sad subpopulations of output neurons in response to the set of test faces shown in Figure 4.4a directly before and after CBM retraining.

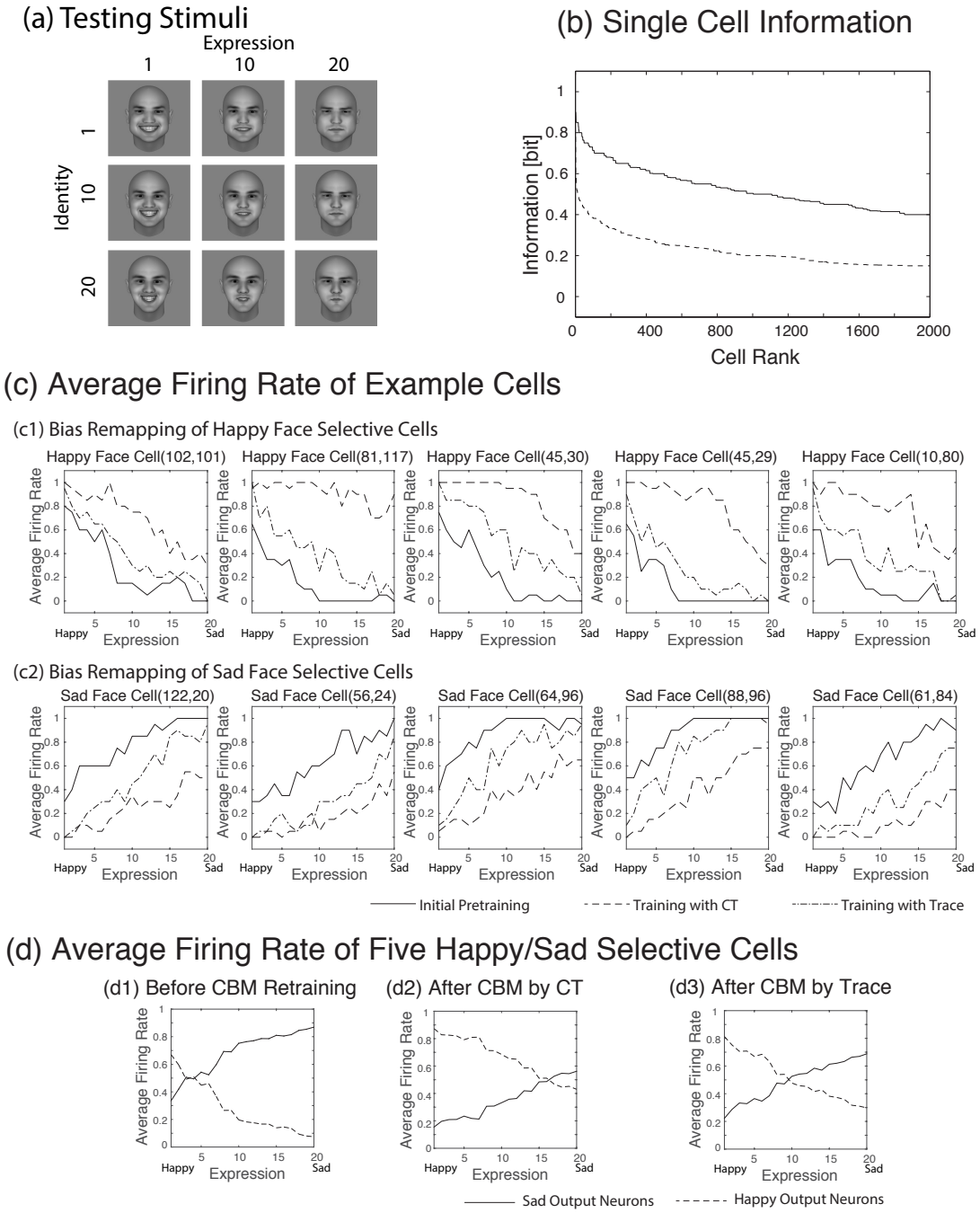

Figure 4.4: **(a)** The face stimuli used to test VisNet. A 1-dimensional space of 20 different facial identities, which varied gradually from Identity A to Identity B, were constructed. Then each of these identities was varied over a 1-dimensional space of 20 different expressions which varied gradually from sad to happy. **(b)**The amount of information carried by output (4th layer) neurons after pretraining VisNet. The plot shows the information carried by all of the 4th layer neurons about either happy or sad expressions, where the neurons are plotted in rank order along the abscissa. **(c)** Demonstration of CBM by CT learning (c1) and trace learning (c2) in VisNet. The firing rates of five Happy output neurons and five Sad output neurons are recorded in response to the matrix of test faces shown in (a) directly before and after CBM retraining. The plots show the average firing rate of the cells in response to 20 different facial expressions ranging from very happy (1) to very sad (20). For each facial expression, the firing rates are averaged over the 20 different facial identities. **(d)** The plots show the average firing rate of all the Happy output cells (dashed line) and all the Sad output cells (solid line) in response to 20 different facial expressions ranging from very happy (1) to very sad (20). For each facial expression, the firing rates are averaged over the 20 different facial identities. The subplot (d1) shows the output of the network directly before CBM retraining, and the subplot (d2) and (d3) shows the output of the network after CBM retraining with CT learning and with trace learning, respectively.

This is the same set of face images as used in the simulation study conducted in Chapter 3 (Eguchi et al., 2016). In particular, a 1-dimensional space of 20 different facial identities, which varied gradually from one Identity A to another Identity B, was constructed. Each of these facial identities was then varied over a 1-dimensional space of 20 different expressions which varied gradually from sad to happy. This produced a matrix of 400 face stimuli constructed from 20 identities $\times$ 20 expressions. By recording the responses of the Happy and Sad subsets of output neurons to these test faces directly before and after CBM retraining, it was now able to assess how well the CBM retraining had remapped the more neutral faces away from the Sad neurons and onto the Happy neurons.

4.4.3.2 Results

After pretraining the network on the set of 100 randomly generated faces (Figure 4.3b), the subset of five output neurons that carried the most information about a happy expression, and another subset of five output neurons that carried the most information about a sad expression were identified. Figure 4.4c shows the average firing rates of the five Happy output neurons (top row) and five Sad output neurons (bottom row) recorded in response to the matrix of test faces shown in Figure 4.4a directly before and after CBM retraining. The plots show the average firing rate of the cells after the initial pretraining (solid line), after the remapping with CT learning (dashed line) in response to 20 different facial expressions ranging from very happy (1) to very sad (20). For each facial expression, the firing rates are averaged over the 20 different facial identities. It can be seen that the Happy output neurons respond with a greater average firing rate across the space of expressions after CBM training by CT learning. In particular, CBM retraining has remapped the more neutral faces away from the Sad output neurons and onto the Happy output neurons.

Furthermore, the subset of 1,000 neurons that carried the most information about a happy expression, and another subset of 1,000 neurons that carried the most information about a sad expression were identified. The firing rates of the subpopulation of Happy output neurons and subpopulation of Sad output neurons were then recorded in response to the matrix of test faces shown in Figure 4.4a directly before and after CBM retraining. Figure 4.4d shows the average firing rate of all the Happy output cells (dashed line) and all the Sad output cells (solid line) in response to 20 different facial expressions ranging from very happy (1) to very sad (20). The left plot shows the output of the network directly before CBM retraining, and the right plot shows the output of the network after CBM retraining with CT learning. It can be seen that directly before CBM retraining, the subpopulation of Sad output neurons respond more strongly on average than the Happy output neurons to all facial expressions greater than 4 on the happiness scale (1-20) represented along the abscissa. However, after CBM retraining, the Sad output neurons respond more strongly than the Happy output neurons only to facial expressions greater than 16 on the happiness scale. Thus, CBM retraining has remapped the more neutral faces away from the sad output neurons and onto the happy output neurons. In particular, CBM retraining is able to shift the bias in the network from negative to positive using a biologically plausible Hebbian learning rule (1.7) with weight vector renormalisation (1.10) and (1.11) when the faces are presented transforming continuously from happy to sad as shown in Figure 4.3c.

4.4.4 Study 2b: CBM by Trace Learning

4.4.4.1 Method

In this section, VisNet underwent a stage of CBM retraining by trace learning after the initial stage of pretraining VisNet on 100 randomised faces as described in Section 4.4.2. During this, the network was retrained on faces with either happy or neutral expressions, with the synapses

modified using the trace learning rule (1.8) and (1.9) with weight vector renormalisation (1.10) and (1.11). Figure 4.3d shows examples of the face stimuli used to perform CBM retraining by trace learning. The image set consisted of 25 faces with a happy expression and 25 faces with a neutral expression. Each of these 50 faces had a different randomly generated identity. Figure 4.3d shows some examples of these images. The top row shows a selection of 5 happy faces, while the bottom row shows 5 neutral faces. During CBM retraining, faces with happy or neutral expressions were shown alternately in an interleaved fashion. That is, the presentation order was happy face 1, neutral face 1, happy face 2, neutral face 2, and so on until eventually happy face 25, neutral face 25. The ordered presentation of all 50 faces constituted one epoch of training. The network underwent a total of 50 training epochs. In this situation, trace learning (Foldiak, 1991; Wallis and Rolls, 1997) will encourage the happy output neurons to learn to respond to both the happy faces and more neutral faces that are presented in temporal proximity. That is, the neurons that are originally selective to only happy faces may start to respond also to the more neutral faces based on temporal associations. In this way, the low-level features representing more neutral faces in the lower layers of the network become remapped onto the more happy output representations. Hence, CBM takes place.

4.4.4.2 Results

The network performance was assessed in a similar manner to that described above for CT learning in Section 4.4.3.2. After pretraining the network on the set of 100 randomly generated faces (Figure 4.3b), the subset of five neurons that carried the most information about a happy expression, and another subset of five neurons that carried the most information about a sad expression were identified. Figure 4.4c shows the average firing rates of the five Happy output neurons (top row) and five Sad output neurons (bottom row) recorded in response to the matrix of test faces shown in Figure 4.4a directly before and after CBM retraining. The plots show the average firing rates of the cells after the initial training (solid line), and after the remapping with trace learning (dash-dot line), in response to 20 different facial expressions ranging from very happy (1) to very sad (20). For each facial expression, the firing rates are averaged over the 20 different facial identities. It can be seen that the Happy output neurons respond with a greater average firing rate across the space of expressions after CBM training by trace learning. In particular, CBM retraining has remapped the more neutral faces away from the Sad output neurons and onto the Happy output neurons.

Also, the subset of 1,000 neurons that carried the most information about a happy expression, and another subset of 1,000 neurons that carried the most information about a sad expression were identified. These were exactly the same subsets of Happy and Sad output cells that were identified for the CT learning simulation described in Section 4.4.3.2 (Study 2a). The firing rates of the subpopulation of Happy output neurons and subpopulation of Sad output neurons were then recorded in response to the matrix of test faces shown in Figure 4.4a directly before and after CBM retraining. Figure 4.4 shows the average firing rate of all the Happy output cells (dashed line) and all the Sad output cells (solid line) in response to 20 different facial expressions ranging from very happy (1) to very sad (20). The subplot (d1) shows the output of the network directly before CBM retraining, and the subplot (d3) shows the output of the network after CBM retraining with trace learning. It can be seen that directly before CBM retraining, the subpopulation of Sad output neurons respond more strongly on average than the Happy output neurons to all facial expressions greater than 3 on the happiness scale (1-20) represented along the abscissa. However, after CBM retraining, the Sad output neurons respond more strongly than the Happy output neurons only to facial expressions greater than 18 on the happiness scale. Hence, the more neutral faces have been remapped away from the sad output neurons and onto the happy output neurons by the CBM retraining. In particular, CBM retraining has shifted the bias in the network from negative to positive using a biologically plausible trace learning rule (1.8) and (1.9) with weight vector renormalisation (1.10) and

(1.11) when the faces are presented with the happy and neutral expressions shown in Figure 4.3d interleaved.

4.5 Discussion

In this chapter, two alternative CBM training mechanisms, continuous transformation (CT) learning (Stringer et al., 2006) and trace learning (Foldiak, 1991; Wallis and Rolls, 1997), are described and modelled. These learning mechanisms were previously used to model how the primate ventral visual pathway learns to perform transform invariant visual object recognition. CT learning binds together input stimuli onto the same categorical output representation using spatial continuity, while trace learning binds together stimuli using temporal continuity. Experimental support for these two learning mechanisms has been provided by previous psychophysical studies, which have confirmed that human subjects bind together different images onto a single categorical representation using a mixture of both spatial continuity (CT learning) and temporal continuity (trace learning) (Perry et al., 2006). Our current simulations have shown that these same learning mechanisms may be implemented in neural network computer models to rewire the synaptic connectivity in order to eliminate the kind of negative cognitive biases associated with clinical depression.

The results of these simulations are highly informative for the development of experimental protocols to develop optimal CBM training methodologies with human participants. One such suggestion is to have a clearer focus on the way in which a bias might be altered. This chapter presents a bias change through the exploitation of two visual learning rules. For a bias change to occur, some sort of learning must occur, so it follows to use what we know about learning to inform CBM procedures. Here the stimuli are optimised for use with trace and CT learning, but future work could look at other types of learning. For example, findings from reinforcement learning research could be used to optimise CBM procedures using feedback.

The first of the CBM retraining mechanisms, CT learning, utilizes a Hebbian learning rule (1.7) with weight vector renormalisation (1.10) and (1.11). During training, the face stimuli gradually transform from happy to sad. The initial presentation of the happy face stimulates the happy output representation, which then stays active while the faces morph continuously through more neutral to sad faces. The continual application of Hebbian learning at each face presentation then remaps the more neutral faces onto the active happy output representation. In this way, the positive output neurons that originally fire only to very happy faces are remapped to also fire to the more neutral faces.

The second of the CBM retraining mechanisms, trace learning, utilizes a trace learning rule (1.8) and (1.9) with weight vector renormalisation (1.10) and (1.11). Trace learning encourages output neurons to respond to input patterns that tend to occur close together in time. During training, the sad and neutral faces are presented to the network in an interleaved manner. The application of trace learning at each face presentation then binds the happy and neutral faces together onto the same happy output representation. In this way, positive output neurons are remapped to also respond to neutral faces.

The two CBM training mechanisms, CT learning and trace learning, were first tested in a simplified one-layer neural network model in order to investigate the operation of these learning mechanisms in a highly controlled way. These computer simulations allowed us to explore the neural and synaptic dynamics underpinning the two CBM training mechanisms. It was found that both CT learning and trace learning were able to remap the synaptic connectivity such that the happy output cell responded to the happy to neutral portion of the stimulus range, while the sad output cell responded only to the sad end of the stimulus range. Thus CBM retraining by either CT learning or trace learning produced successful CBM, where the bias in the network was shifted from negative to positive.

Next the CBM training methodologies in a much more biologically detailed multi-layer

model, VisNet, with realistic face images generated using the FaceGen 3D face modelling software package were tested. The network was first pretrained on 100 randomly generated faces with a variety of facial identities and expressions ranging from happy to sad, as described in Section 4.4.2. During this pretraining stage, the network developed output neurons responding preferentially to either happy or sad facial expressions, as previously shown in Chapter 3 (Eguchi et al., 2016). Then the network underwent CBM retraining using either CT learning or trace learning as described in Section 4.4.3 (Study 2a) and in Section 4.4.4 (Study 2b), respectively. It was found that both CT learning and trace learning were able to remap the more neutral faces away from the sad output neurons and onto the happy output neurons, thus shifting the cognitive bias in the network connectivity from negative to positive.

To the authors' knowledge, this is the first study that has modelled the application of the CT learning and trace learning mechanisms to CBM-Interpretation. Previous experimental studies have found that CBM-Interpretation can reduce negative cognitive biases in human participants (e.g. Grey and Mathews, 2000; Mathews and Mackintosh, 2000), which in turn can reduce the risk for depression recurrence (Holmes et al., 2009). Furthermore, considering the fact that neither CT nor trace learning require any feedback signal, this can potentially be a very powerful tool for the treatment. The work reported in this chapter provides potential explanations at the neuronal and synaptic level for how such a shift in interpretational bias might occur through CBM training. Understanding the way in which biases can be shifted is crucial at present, given the mixed results seen in CBM research so far (Fox et al., 2014).

4.5.1 Future Work

The development of well specified computational models may help to guide future research aimed at optimising the effectiveness of CBM treatments. For example, the simulations presented in this chapter utilized either CT learning or trace learning, but not both together, to effect a shift in the cognitive bias from negative to positive. However, psychophysical studies have shown that human subjects bind together different images onto a single categorical representation using a mixture of both spatial continuity and temporal continuity (Perry et al., 2006). Wallis and Blthoff (2001) have also shown that both spatial and temporal continuity seem to play a key role for modifying recognition memory. The researchers presented subjects with sequences of rotating faces, in which the identity of the face changed during rotation. It was found that observers tended to bind together the different views into a single identity if the faces transformed with both spatial and temporal continuity during training. That is, the sequence of face images had to correspond to a continuous rotation of the head with the identity gradually transformed between consecutive head orientations. Furthermore, a recent modelling study has predicted that invariance learning in the primate ventral visual pathway may be most effective when CT learning and trace learning are combined together simultaneously (Spoerer et al., 2016). Therefore, in future work I may investigate CBM training methodologies that combine together both CT learning and trace learning simultaneously for maximum therapeutic effect.

Besides CBM-interpretation training explored in this chapter, there has been argument that antidepressants operate by initially shifting negative cognitive biases early in treatment, which then leads to the later improvement of mood (Harmer, 2012). In support of this hypothesis, Harmer and colleagues have found that the action of antidepressants can reduce negative affective biases in depressed patients (Harmer et al., 2009), as well as modify the neural processing of nonconscious threat cues (Harmer et al., 2006). In the future, I plan to develop computer simulations aimed at shedding light on the operation of antidepressants in the brain as an extension to the work presented in this chapter. The mechanisms by which antidepressants shift information processing from negative to positive may be similar to what has been proposed above for the CBM training methodologies. However, the initial shift in cognitive bias caused by antidepressants must now be achieved by some form of pharmaceutically driven global neuromodulation. Such neuromodulation may affect the processing of every day sensory

experiences, which in turn may drive further synaptic modification resulting in reduced innate negative biases. Therefore, I will explore how global changes to neuronal and synaptic model parameters may lead to shifts in cognitive bias similar to those described above for the CBM training methodologies.

Chapter 5

The Neural Basis of Border Ownership Representations

Our visual perception tends to assign luminance contrast borders to one or other of the adjacent image regions. Experimental evidence for the neuronal coding of such border-ownership in the primate visual system has been reported in neurophysiology. I investigated exactly how such neural circuits may develop through visually-guided learning. More specifically, I investigated through computer simulation how top-down connections may play a fundamental role in the development of border ownership representations in the early cortical visual layers V1/V2. Our model consists of a hierarchy of competitive neuronal layers, with both bottom-up and top-down synaptic connections between successive layers, and the synaptic connections are self-organised by a biologically plausible, temporal trace learning rule during training on differently shaped visual objects. The simulations reported in this chapter have demonstrated that top-down connections may help to guide competitive learning in lower layers, thus driving the formation of lower level (border ownership) visual representations in V1/V2 that are modulated by higher level (object boundary element) representations in V4. Lastly I investigate the limitations of our model in the more general situation where multiple objects are presented to the network simultaneously.

5.1 Introduction

As Rubin's famous vase (Figure 5.1) demonstrates, our visual perception tends to assign luminance contrast borders to one or other of the adjacent image regions, as if they serve as occluding contours (von der Heydt et al., 2003). This is an example of *feature binding* in vision, in this case binding a luminance contrast border to a particular object. Representing such binding relationships between visual features is essential to the ability of the visual system to interpret and *make sense* of complex visual scenes. Experimental evidence for the neuronal coding of such border-ownership in the primate visual system has arisen in a neurophysiology study carried out by Zhou et al. (2000).

Zhou et al. (2000) have shown that the responses of simple cells in earlier cortical stages of visual processing such as V1 and V2, which respond preferentially to oriented edges, are also modulated by which side of an object or figure the edge occurs on. This is the case even when the figure/background cues lie well outside the classical receptive field of the neuron, which in area V1 is approximately 1 degree in size. Such neurons are referred to as *border ownership cells*. Sugihara et al. (2011) later reported that the border ownership signal emerges with a latency of 61 ms, which is about 13 ms later than the onset of orientation selectivity. This suggests that the global image context specifying border ownership modulates the activity of these neurons. In other words, there must be a mechanism that enables the contextual information to be conveyed to these early stage visual neurons in V1 and V2. It has been proposed that these

Figure 5.1: Rubin's Vase (Rubin, 1915)

kinds of border ownership responses in area V1 represent a form of feature binding, and so may be important for understanding how primate vision may solve the problem of feature binding more generally.

Some theoreticians have suggested that the context integration required for border ownership representations in V1 and V2 can be achieved via lateral propagation of signals within a layer via horizontal fibres (Zhaoping, 2005; Baek and Sajda, 2005; Nishimura and Sakai, 2004). However, Sugihara et al. (2011) have argued that the conduction velocity of horizontal fibres is too slow (most of them being between 0.1 and 0.4 m/s (Angelucci and Bullier, 2003)) to produce the border ownership signals within the short latency observed in neurophysiology studies. Furthermore, Sugihara et al. (2011) showed that varying the distance between the target border and the visual features that carry contextual information about the 'owner' of the border does not in fact influence the latency before the border ownership signals arise. Therefore, they concluded that context influence by horizontal signal propagation alone is highly unlikely.

On the other hand, the feedforward (bottom-up) and feedback (top-down) connections between successive visual stages have fast-conducting axons, with conduction velocities of between 2 and 6 m/s, which is about ten times faster than cortical horizontal fibres (Angelucci and Bullier, 2003). Accordingly, both Craft et al. (2007) and Jehee et al. (2007) have proposed models that involve hypothetical 'grouping circuits' within a higher cortical layer that capture the contextual information about local boundary elements, and these contextual signals are then relayed down through feedback connections to modulate responses in an earlier layer. They proposed that the larger receptive fields in the higher layer allow the network to employ 'grouping circuits' without having to rely on slow lateral propagation of signals. Nevertheless, it still remains a challenge to understand exactly how such neural circuits may be learned. The objective of the current study is to investigate the learning mechanisms that underpin the development of border ownership cells in the primate visual brain, in terms of synaptic modification guided by visual experience and consequent neural adaptation throughout a hierarchy of cortical stages. Moreover, given the proposed role of border ownership cells in feature binding, which is essential for integrating the visual features within a scene, the simulations described below provide a step towards understanding how the brain learns to make sense of the visual world.

One higher visual area that might provide appropriate top-down modulatory signals is V4, which contains neurons that represent the localised boundary contour elements of objects (Layton et al., 2012). The responses of these neurons are sensitive to both the shape of the boundary element and where the element is with respect to the centre of mass of the object (Pasupathy and Connor, 2001, 2002). Hence each of the neurons encodes that a specific border element belongs to a particular object - i.e. a kind of border ownership representation. A subpopulation of these neurons will provide a distributed representation of the entire boundary of the object. Furthermore, the neurons are able to respond invariantly as the object is shifted across different

locations on the retina over a modest range.

In Chapter 2, the visually-guided development of such V4 cells has been investigated in a computational modelling study with an established neural network model, VisNet, of the primate ventral visual pathway (Eguchi et al., 2015). The network architecture consisted of a hierarchy of cortical visual layers, with each layer modelled as a competitive neural network (Wallis and Rolls, 1997). Whenever an image was presented to the network, visual signals propagated through feedforward plastic synaptic connections between successive layers. Within each competitive layer, the excitatory cells competed with each other to respond to the current visual stimulus. In the brain, competition between excitatory cells is implemented via inhibitory interneurons. Although to save computational expense in VisNet, competition between excitatory neurons is modelled more directly using local filters. During an initial period of training with visual objects, the feedforward synaptic connections between successive layers of the network are continually modified using local, biologically plausible, associative learning rules. The competition within each layer then forces individual neurons to learn to respond selectively to a particular stimulus class, with different neurons responding to different kinds of stimulus. Competitive learning is a very simple unsupervised learning paradigm that allows neurons to discover important features of the stimulus input patterns (Rumelhart and Zipser, 1985). In Chapter 2, it was shown that the gradual increase in the receptive field size of neurons through successive layers of the visual system (Gross et al., 1969; Pettet and Gilbert, 1992) allows V4 neurons access to local image information specifying how localised luminance contrast contours belong to adjacent object regions (Eguchi et al., 2015). As a result, cells in the higher layer of their hierarchical competitive neural network model developed neuronal response properties similar to those reported by Pasupathy and Connor (2001, 2002) when the model was trained on a number of real world objects.

In this chapter, the previous purely feedforward model investigated in Chapter 2 (Eguchi et al., 2015) is extended by incorporating both feedforward (bottom-up) and feedback (top-down) connections. This extended model architecture is used to investigate how the edge-detecting simple cells in the earliest layer of the network, which corresponds to visual areas V1/V2 in the primate brain, may develop border ownership representations via top-down modulation from neurons in the output layer, which corresponds to visual area V4. The necessary feedforward and feedback synaptic connectivity within the network is set up by visually-guided learning using a biologically plausible, local, trace learning rule (Foldiak, 1991) as the network is trained on a collection of differently shaped visual object stimuli. I go on to show how these border ownership signals in the earliest layer evolve dynamically during the 300ms time course of a stimulus presentation, as reported by Sugihara et al. (2011) and Jehee et al. (2007). The limitations of the model in the more general situation where multiple objects are presented to the network simultaneously are then investigated.

5.2 Hypothesis

In Chapter 2, it was shown that when an established hierarchical neural network model of the primate ventral visual pathway, VisNet (Wallis and Rolls, 1997), is trained on 177 images of real world objects, which rotated in plane through 360 degrees and shifted across a 3×3 grid of nine different retinal locations, the neurons in the higher layers of the model learn to represent local boundary contour elements (Eguchi et al., 2015). Individual neurons are tuned to boundary elements with a specific curvature at a particular location with respect to the centre of mass of the object. Moreover, the neurons respond invariantly as an object is translated across different retinal locations. These are the same neuronal response properties as observed in area V4 of the primate visual system by Pasupathy and Connor (2002). Although they have reported that the translation invariant responses of V4 neurons are only over a modest range, we can simply suppose that the size of simulated retina in the model matches to the covered range.

The version of the VisNet architecture used in the previous study reported in Chapter 2 incorporated only feedforward (bottom-up) connections between successive layers of the network (Eguchi et al., 2015). No feedback (top-down) connections were included in the model even though these are known to exist in the primate ventral visual pathway. It has previously been suggested that the top-down connections might implement attention to objects during visual search (Deco and Lee, 2002; Wagatsuma et al., 2013) and were incorporated into a variant of VisNet model to simulate top-down biasing effects (Deco and Rolls, 2004). However, in this previous study the top-down connections were only implemented after training, and so did not influence the visual representations that developed during visually-guided learning. In contrast, in our present chapter the top-down connections are also present during training, and thus play a key role in the development of border ownership representations in the early layers. In particular, I propose that the global image context specifying border ownership is conveyed to the early stage visual neurons by top-down connections between layers in order to drive the development of border ownership cells in the early cortical areas as reported by Zhou et al. (2000).

Accordingly, I hypothesised that learning in the extended VisNet architecture introduced in this chapter would operate as follows. First, during visually-guided learning in which VisNet is trained on images of differently shaped objects, neurons in the later stages of visual processing such as V4 will learn to encode boundary contour elements through learning in the feedforward connections as previously demonstrated in Chapter 2 (Eguchi et al., 2015). Next, with continued visually-guided training on the same object images, I expect that strong polysynaptic feedback connections may subsequently develop from those neurons in the later stages of visual processing to neurons in earlier stages such as V1 and V2. These strengthened top-down connections might then modulate the responses of neurons in V1 and V2 according to where their preferred edge element occurs within an object.

More precisely, let us consider a subset Φ^{V4}_{Left} of neurons in V4 that have learned, by the visually-guided competitive learning mechanisms, to encode a vertical straight contour on the left of an object across different retinal locations. This subset of V4 neurons may also develop strengthened top-down polysynaptic connections to a subset of simple cells in V1 and V2 that originally signal the presence of any vertical straight contour within their small classical receptive field. This will force the subset of V1/V2 neurons to preferentially respond when the vertical straight contour is part of the left boundary of an object (top-down signals) at a particular retinal location (bottom-up signals).

Figure 5.2(a) shows a case example in which an object with a straight vertical border on its *left* is presented with this border positioned at retinal location *1*. The figure illustrates how the subset Φ^{V4}_{Left} of V4 neurons, which represent a vertical straight edge on the left of an object, may modulate the responses of a subset of V1/V2 simple cells $\Phi^{V1/V2}_{Left,Loc1}$ that represent the presence of a vertical contour at retinal location 1. Figure 5.2(b)-(d) shows similar case examples in which the vertical straight edge may occur on either the left or right boundary of the object, with the vertical straight edge positioned in either retinal location 1 or location 2.

In summary, I hypothesise that the observations of Zhou et al. (2000), in which the responses of V1 and V2 neurons are modulated by which side of a figure the edge occurs on, may be replicated by incorporating both bottom-up and top-down associatively modifiable connections within VisNet. This will allow neurons in the early layers to develop their firing responses through visually-guided competitive learning driven by a combination of both bottom-up and top-down visual signals. The neural circuits developed after visually-guided learning in VisNet are expected to be similar to the hypothetical 'grouping circuits' proposed in a previous modelling study of border ownership representation with top-down connections carried out by Craft et al. (2007). However, the focus of our current study is to investigate exactly how such neural circuits may be learned when the network is trained on visual images of differently shaped objects.

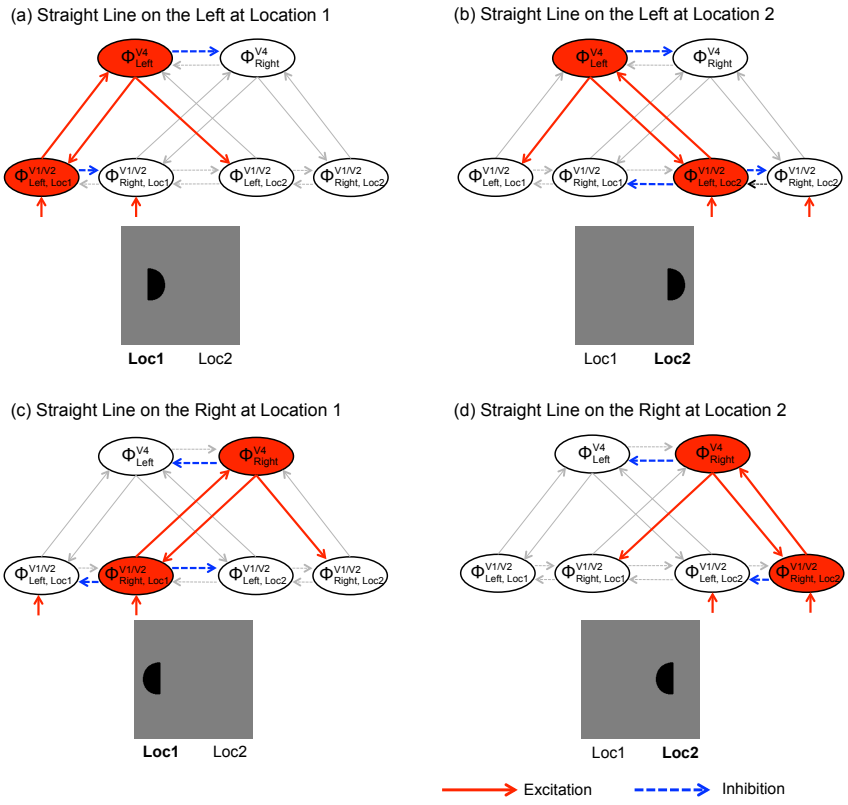

Figure 5.2: Hypothesised modulation of edge detecting simple cells in lower layers V1/V2 by top-down signals from higher layer V4 neurons representing boundary contour elements. The figure shows the steady state activations of all neurons after sufficient time (e.g. $\geq$ 61ms) has elapsed after stimulus presentation to allow visual signals to propagate from the retina up to V4 and then back down to modulate V1/V2 responses. The following four cases are shown. (a) An object with a straight vertical border on its *left* is presented with this border positioned at retinal location *1*. Ascending visual input initially stimulates both subsets of V1/V2 neurons, $\Phi^{V1/V2}_{Left,Loc1}$ and $\Phi^{V1/V2}_{Right,Loc1}$, representing a vertical straight edge at retinal location 1. However, in layer V4, only those V4 neurons Φ^{V4}_{Left} representing a vertical straight edge on the left of an object are preferentially stimulated by the current visual input. Note that these V4 neurons receive additional feedforward (bottom-up) input signals from other V1/V2 neurons (not shown in the figure) which represent local image context, and these additional context signals are required to guide the selective responses of the V4 neurons. How V4 neurons may develop such response properties through self-organisation of the feedforward connections has been previously modelled in Chapter 2 (Eguchi et al., 2015). The subset of V4 neurons Φ^{V4}_{Left} then stimulates via feedback (top-down) connections those two subsets of V1/V2 neurons $\Phi^{V1/V2}_{Left,Loc1}$ and $\Phi^{V1/V2}_{Left,Loc2}$ which receive strengthened connections from Φ^{V4}_{Left} and are consequently modulated by a straight vertical edge on the left of an object. However, only the particular subset of V1/V2 cells $\Phi^{V1/V2}_{Left,Loc1}$, which represent a vertical bar at retinal location 1 where the vertical bar forms the left hand border of an object, receive the greatest combination of bottom-up and top-down input. Consequently, these V1/V2 neurons fire maximally, representing the border ownership of the vertical edge at this location. (b) An object with a straight vertical border on its *left* is presented with this border positioned at retinal location *2*. In this case, the subset of V1/V2 cells $\Phi^{V1/V2}_{Left,Loc2}$, which represent a vertical bar at retinal location 2 where the vertical bar forms the left hand border of an object, receive the greatest combination of bottom-up and top-down input and fire maximally. (c) An object with a straight vertical border on its *right* is presented with this border positioned at retinal location *1*. This time the subset of V1/V2 cells $\Phi^{V1/V2}_{Right,Loc1}$, which represent a vertical bar at retinal location 1 where the vertical bar forms the right hand border of an object, receive the greatest combination of bottom-up and top-down input and fire maximally. (d) An object with a straight vertical border on its *right* is presented with this border positioned at retinal location *2*. Now the subset of V1/V2 cells $\Phi^{V1/V2}_{Right,Loc2}$, which represent a vertical bar at retinal location 2 where the vertical bar forms the right hand border of an object, receive the greatest combination of bottom-up and top-down input and fire maximally.

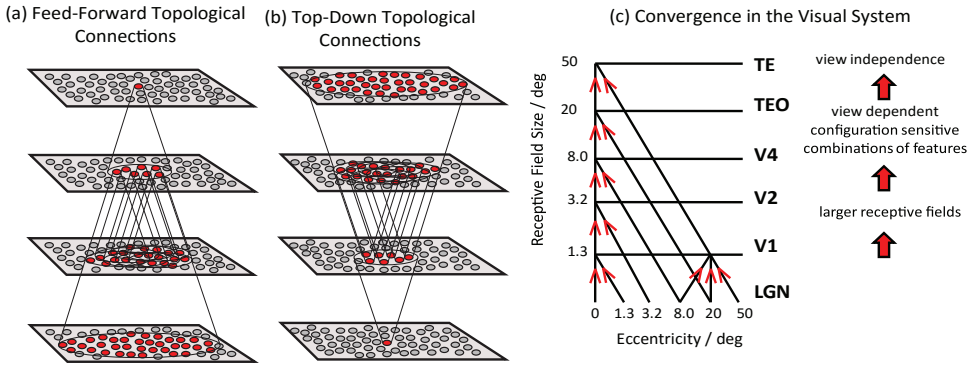

Figure 5.3: (a) The original four-layer feedforward (bottom-up) version of the VisNet architecture. The figure shows the feedforward connectivity, where each neuron receives connections from a topologically corresponding region of the preceding layer. The convergence of feedforward connections through the network is designed to provide fourth layer neurons with information from across the entire input retina. The new VisNet architecture implemented in this chapter was extended to incorporate additional feedback (top-down) connections, which have the similar topological connectivity as the feedforward connections except in the opposite direction as shown in (b). (c) Convergence in the visual system V1: visual cortex area V1;TEO: posterior inferior temporal cortex, TE: anterior inferior temporal cortex (IT).

5.3 Materials & Methods

5.3.1 VisNet Model

The simulation studies presented in this chapter are conducted with a modified version of an established neural network model, VisNet, of the primate ventral visual pathway, which was originally developed by Wallis and Rolls (1997). A detailed description of the original model is provided in Section 1.3. In the current simulations reported below, the number of the layers has been reduced to three since a large number of border ownership neurons were found to develop in the third layer of VisNet, which corresponds to TEO in the earlier study reported in Chapter 2 (Eguchi et al., 2015).

In the simulations described in this chapter, the VisNet architecture was extended to incorporate additional feedback (top-down) connections, which have the similar topological connectivity as the feedforward connections except in the opposite direction (Figure 5.3(b)). Both the feedforward and feedback 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. The values used in the current studies are given in Table 5.1. The gradual increase in the receptive field of cells in successive layers 1 to 3 reflects the known physiology of the primate ventral visual pathway (Freeman and Simoncelli, 2011; Pasupathy, 2006; Pettet and Gilbert, 1992).

Furthermore, in order to investigate the precise temporal dynamics of the top-down modulation, the original discrete time model, which has been used for past VisNet studies, was converted into a time-continuous model with differential equations that are given below.

5.3.1.1 Pre-processing of the visual input by Gabor filters

Before the visual images are presented to the VisNet's input layer 1, they are preprocessed by a set of Gabor filters, previously implemented by Deco and Rolls (2004), which accord with the general tuning profiles of simple cells in V1 (Jones and Palmer, 1987; Cumming and Parker, 1999; Lades et al., 1993). The filters provide a unique pattern of filter outputs for each transform of each visual object, which is passed through to the first layer of VisNet. These filters are known to provide a good fit to the firing properties of V1 simple cells, which respond to local oriented bars and edges within the visual field (Jones and Palmer, 1987; Cumming and Parker, 1999).

Table 5.1: Parameters used for simulations with VisNet

(a) Parameters for VisNet Model			
Layer	**1**	**2**	**3**
Dimensions	64×64	64×64	64×64
Number of feedforward fan-in connections	201	100	100
Fan-in Radius (feedforward)	12	12	18
Number of feedback fan-in connections	5	5	-
Fan-in Radius (feedback)	12	12	-
Sparseness of activations (set by adjusting sigmoid threshold α)	33 %	33 %	50 %
Sigmoid slope (β)	31.5	46.1	1.48
Learning rate (k)	1.0	1.0	1.0
Excitatory Radius (σ_E)	1.4	1.1	0.8
Excitatory Contrast (δ_E)	5.35	33.15	117.57
Inhibitory Radius (σ_I)	2.76	5.4	8.0
Inhibitory Contrast (δ_I)	1.6	1.5	1.5
(b) Parameters for Gabor Filtering			
Phase shift (ψ)	0, π, $-\pi/2$, $\pi/2$		
Wavelength (λ)	2		
Orientation (θ)	$0, \pi/4, \pi/2, 3\pi/4$		
Spatial bandwidth (b)	1.5 octaves		
Aspect ratio (γ)	0.5		
(c) Parameters for Differential Model			
Activation time constant (τ_h) [s]	0.1		
Trace time constant (τ_t) [s]	0.5		
Presentation time per stimulus transform [s]	1.0		
Numerical step size (Δt) [s]	0.01		

The input filters used are computed by the Equation (1.1) and (1.2). In the experiments in this chapter, an array of Gabor filters is generated at each of 256 $\times$ 256 retinal locations with the parameters given in Table 5.1.

The outputs of the Gabor filters are passed to the neurons in layer 1 of VisNet according to the synaptic connectivity given in Table 5.1. That is, each layer 1 neuron receives connections from 201 randomly chosen Gabor filters localised within a topologically corresponding region of the retina (this number has been used to be consistent with the original VisNet study (Wallis and Rolls, 1997)). These distributions are defined by a radius shown in table 5.1.

5.3.1.2 Activations of neurons and competition within the network

Within each of the neural layers 1 to 3 of the network, the activation h_i of each neuron i is governed by the following differential equation:

$$\tau_h \frac{dh_i(t)}{dt} = -h_i(t) + \sum_j w_{ij}(t) r_j(t) \tag{5.1}$$

where τ_h is the time constant, r_j is the firing rate of presynaptic neuron j, and w_{ij} is the strength of the synapse from neuron j to neuron i. The value of τ_h used in the simulations is 0.1, which is larger than the typical values used for spiking network, 0.01. However, since spikes of the neurons are not implemented, and the synaptic learning rule does not depend on the precise timing like STDP, the larger time constant was used for this particular model for the speed of its computation. In this chapter, the full differential model, which comprises equation (5.1) and equations (5.2) and (5.3) given below, is numerically simulated using a Forward Euler finite difference scheme with a fixed numerical timestep Δt given in Table 5.1.

In this chapter, simulations with a self-organising map (SOM) (von der Malsburg, 1973; Kohonen, 1982) implemented within each layer were conducted. In the SOM architecture, short-range excitation and long-range inhibition are combined to form a Mexican-hat spatial profile and is constructed as a difference of two Gaussians as Equation (1.5). The lateral inhibition and excitation parameters used in the SOM architecture are given in Table 5.1. These values were previously found to optimize the performance of the VisNet model (Rolls, 2000; Tromans

et al., 2011).

Next, the contrast between the activations of neurons within each layer is enhanced by passing the activations of the neurons through a sigmoid transfer function as Equation (1.6). The parameters for the sigmoid activation function are shown in Table 5.1.

5.3.1.3 Modification of synaptic weights during training

During training with visual objects, while the connectivity pattern is fixed, the strengths of the feedforward and feedback synaptic connections between successive neuronal layers are modified by a trace learning rule (see Section 1.4.2) (Foldiak, 1991; Wallis and Rolls, 1997), which incorporates a memory trace of recent neuronal activity:

$$\frac{dw_{ij}(t)}{dt} = k\overline{r_i}(t)r_j(t) \tag{5.2}$$

where $r_j(t)$ is the firing rate of pre-synaptic neuron j, $\overline{r_i}(t)$ is the memory trace value of the firing rate of post-synaptic neuron i, w_{ij} is the synaptic weight from pre-synaptic neuron j to post-synaptic neuron i, and k is the learning rate constant. The memory trace value $\overline{r_i}(t)$ is updated according to the equation:

$$\tau_t \frac{\overline{r_i}(t)}{dt} = -\overline{r_i}(t) + r_i(t) \tag{5.3}$$

where $r_i(t)$ is the firing rate of post-synaptic neuron i, and τ_t is a trace time constant which is given in Table 5.1. The effect of the trace learning rule (5.2) is to encourage neurons to learn to respond to visual input patterns that tend to occur close together in time. The utility of this temporal binding is as follows. If, during training, each object is presented to the network in a sequence of different retinal locations clustered together in time before switching to the next object, then this enables neurons in higher layers to learn to respond to their preferred visual stimulus with shift invariance across different retinal locations as described in the earlier simulation study reported in Chapter 2 (Eguchi et al., 2015).

During the numerical simulation, to prevent the same few neurons always winning the competition, the synaptic weight vector $\mathbf{w}_i$ for each neuron i is normalised to unit length after each learning update for each training image by Equation (1.10) and (1.11).

In the original discrete-time version of VisNet, the synaptic weights are trained layer by layer (Wallis and Rolls, 1997). However, it is important to note that in the current time-continuous version of VisNet, all the synapses across the layers are trained simultaneously. This means that every time step, each neuron calculates the weighted sum of the pre-synaptic activations, at both feed-forward and top-down synapses, to update the activation h (Equation (5.1)). Next the neuronal firing rates within each layer are simultaneously determined by applying the SOM filter (Equation (1.5)) and then the contrast enhancement (Equation (1.6)). The trace learning rule (Equation (5.2) and (5.3)) is then applied at all of the synapses simultaneously to update the synaptic weights. In other words, in the current VisNet model, the training of the backprojections starts at the same time as the forward projections, with the bottom-up and top-down afferent connections to all of the layers being trained simultaneously.

5.3.2 Analysis Techniques

Information theory is used to quantify how selective neurons are for members of a particular stimulus category. If a neuron responds invariantly to the members of a particular stimulus category but not to stimuli from other stimulus categories, then the neuron carries a high level of information about the presence of its preferred stimulus category.

For example, information theory has been used to quantify how well neurons have learned to respond selectively to a particular visual stimulus with translation invariance across different

retinal locations. If the responses r of a neuron carry a high level of information about the presence of a particular stimulus s across different retinal locations, then this implies that the neuron will respond selectively to the presence of that stimulus regardless of where the stimulus is presented on the retina. In this way, information theory can provide a direct measure of both the selectivity of a neuron for a particular stimulus, as well as how translation-invariant the neuronal responses are as the stimulus is shifted across the retina.

In this chapter, I continue to use information theory to assess the stimulus selectivity and translation invariance of neurons in the layer 3 that have learned to respond to localised object boundary elements with translation invariance, as previously investigated in Chapter 2 (Eguchi et al., 2015). However, in this new study information theory was also applied to assess how well simple cells in layer 1 have learned to represent border ownership through top-down modulation. Therefore information theory was used to assess whether some layer 1 simple cells learn to respond selectively to a vertical straight edge on the *left* boundary of an object, while other simple cells learn to respond to a vertical straight edge on the *right* boundary of an object, regardless of the overall object shape. The simple cells in layer 1 have a small fan-in from the retina and are tuned to specific retinal locations, and consequently do not respond invariantly over different retinal locations. Instead, the simple cells should ideally respond invariantly over different global object shapes, as long as there is a straight vertical edge in the correct location on the object boundary.

Two information measures were used to assess network performance (see Rolls et al. (1997); Rolls and Milward (2000)). These two measure use the responses from either individual neurons (single-cell information analysis) or small ensembles of neurons (multiple-cell information analysis). A detailed description of the process is explained in Section 1.5.2.

5.4 Simulation Studies

5.4.1 Study 1: simulation of the visually-guided development of border ownership representations

In this simulation study, VisNet was initially trained and tested on the same abstract visual object shapes (familiar objects) shown in Figure 5.4(a). The model was then also cross-validated by testing the same trained network on the novel visual objects (novel objects) shown in Figure 5.4(b), which were not presented to the network during initial training. The familiar objects were hexagons and semicircles, which were either black or light grey. Black objects were presented against a light grey background, while light grey objects were presented against a black background. Each object had a vertical straight edge either on its left boundary (Figure 5.4(a2,a4)) or right boundary (Figure 5.4(a1,a3)). Although in the natural environment, the object does not normally jump from one location to the other instantaneously, the region activated on the retina does constantly shifts around due to the rapid eye movement called saccades. To simulate this effect, during training and testing, each object was presented in two locations on the left (Location 1) and right (Location 2) of the 256×256 retina.

Whenever an object was presented on the left of the retina, the vertical straight edge on its (left or right) boundary was precisely aligned with retinal Location 1 (Figure 5.4(a1,a2)). This enabled us to explore the top-down modulation of the subpopulation of simple cells in Layer 1 tuned to vertical straight edges at this specific retinal location. In a similar manner, whenever the object was presented on the right of the retina, the vertical straight edge on its (left or right) boundary was aligned with retinal Location 2 (Figure 5.4(a3,a4)). Again, this permitted us to explore the top-down modulation of simple cells in Layer 1 tuned to vertical straight edges at this retinal location.

During training, the familiar objects shown in Figure 5.4(a) were presented to the network one at a time shifting across the two retinal locations while the feedforward and feedback

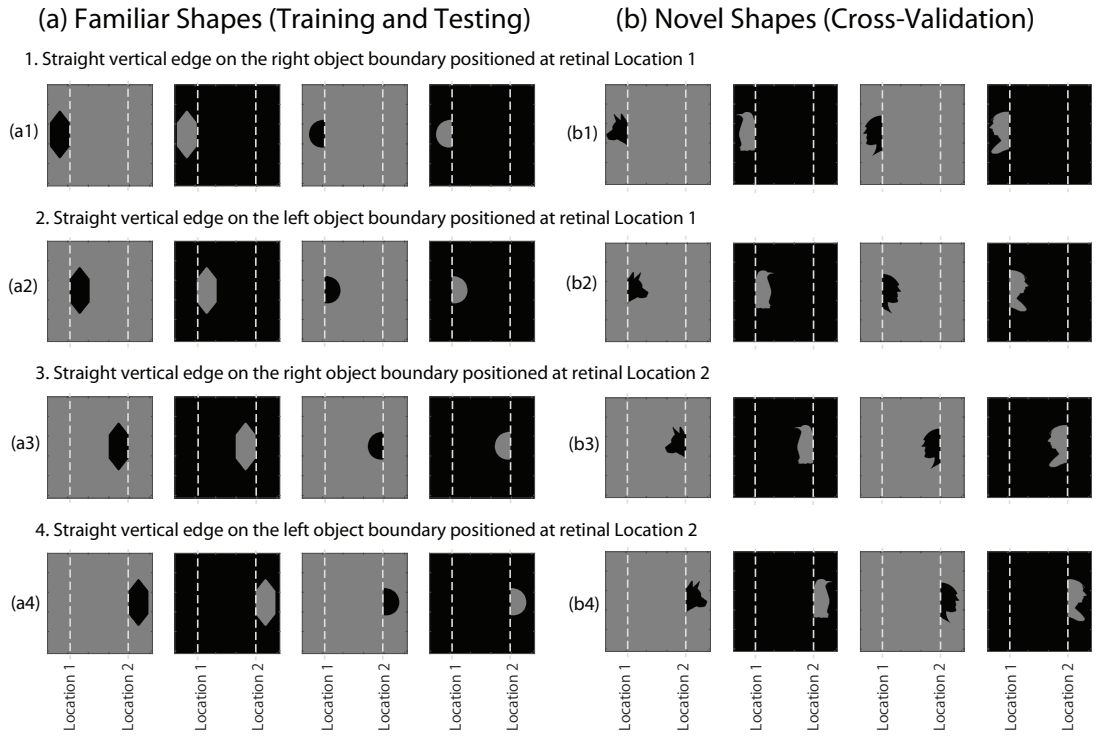

Figure 5.4: The visual object stimuli used for the simulation study. **(a)** A set of abstract familiar shape stimuli used to both train and test the network model (shaded hexagons and semicircles). The objects were black when presented on a light grey background or light grey when presented on a black background. Each object had a vertical straight edge either on its left boundary (a1,a3) or right boundary (a2,a4). During training and testing, each object was presented in two locations on the left and right of the retina. Whenever an object was presented on the left of the retina, the vertical straight edge on its (left or right) boundary was precisely aligned with retinal Location 1 (a1,a3). Similarly, whenever the object was presented on the right of the retina, the vertical straight edge on its (left or right) boundary was aligned with retinal Location 2 (a2,a4). **(b)** A set of novel stimuli used to cross-validate the performance of the network after it had been trained on the familiar set of stimuli (a). The four novel stimuli were a dog's head, a penguin, and two differently shaped human heads. Each novel stimulus has a vertical straight edge on one side. The four novel objects are each presented in two retinal locations in a similar manner to the familiar shapes (a). This gives a total of eight novel stimulus presentations.

synaptic connections between successive layers were modified using the trace learning rule (5.2) and (5.3). The trace learning rule in the feedforward connections drives the development of neuronal responses in the higher layers that are translation invariant across different retinal locations by encouraging postsynaptic neurons to learn to respond to subsets of input patterns that tend to occur close together in time. As long as, during training, each object is presented across different retinal locations in temporal proximity, then the trace learning rule will produce output neurons that have learned to respond selectively to a particular object feature in a translation invariant manner. Therefore, during training, each object was selected in turn and presented in the two different retinal locations before moving on to the next object.

5.4.1.1 Steady state firing properties of cells in layers 1 and 3 at the end of each stimulus presentation

In this section the *steady state* firing responses of Layer 1 and Layer 3 neurons at the end of each stimulus presentation before and after training were analysed with the same object stimuli used for training (familiar objects) shown in Figure 5.4(a) as well as with the novel object stimuli shown in Figure 5.4(b) to cross-validate the developed response properties.

The firing properties of the output (Layer 3) neurons was first tested to investigate whether these neurons had learned to respond selectively to the presence of a vertical straight edge on either the left boundary or right boundary of an object. Such neurons had to respond

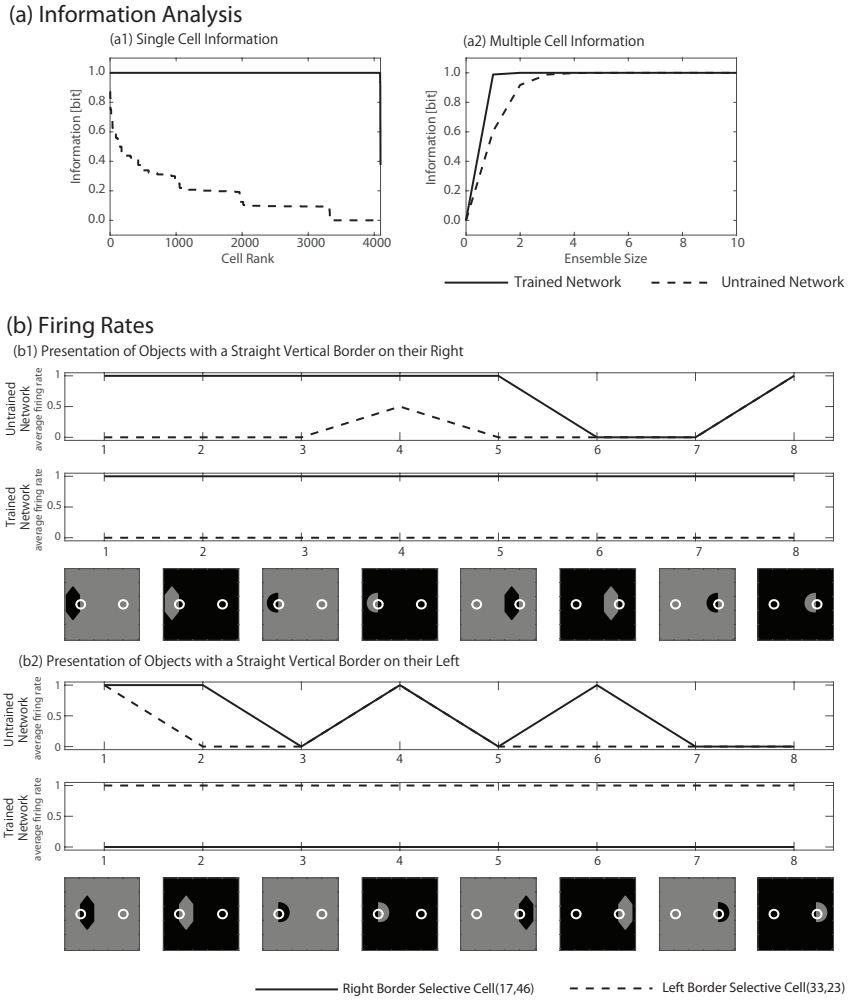

Figure 5.5: The steady state response properties of Layer 3 neurons at the end of each stimulus presentation of familiar shapes that were used to train the network (shown in Figure 5.4(a)). **(a) Information analysis:** Information carried by the output (3rd layer) neurons about whether the vertical straight edge was on the left or right boundary of each object presented to the network before and after training. Plot (a1) shows the maximum single cell information carried by each of the 4096 neurons in Layer 3 about which one of the two stimulus categories was presented, where all of the neurons in Layer 3 are plotted along the abscissa in rank order. The result shows that nearly all of the Layer 3 neurons learned to respond selectively to a vertical straight edge either on the left or on the right of an object boundary, regardless of the global shape, shading or retinal location of the object. Plot (a2) shows the multiple cell information carried by different sized (i.e. up to ten neurons) random ensembles of Layer 3 neurons that individually had high levels of single cell information. It is evident that training has led to an increase in the multiple cell information, which after training asymptotes to the maximum level of 1 bit with only one neuron included in the analysis. **(b) Firing rate responses of two Layer 3 neurons that has maximum single cell information:** plot (b1) shows the responses of two Layer 3 neurons to all eight objects with a vertical straight edge on their right boundary, and plot (b2) shows the responses of the same two Layer 3 neurons to all eight objects with a vertical straight edge on their left boundary. These results show that neuron (17,46) learned to respond selectively to all objects with a vertical straight edge on the right, while neuron (33,23) learned to respond to all objects with a vertical straight edge on the left.

invariantly across different global object shapes (i.e. hexagon or semicircle), different kinds of object shading (i.e. black or light grey), and different trained retinal locations (i.e. Location 1 or Location 2). The same set of stimuli used to train the network shown in Figure 5.4(a) was presented to VisNet during testing, and the firing rate of each neuron in the output layer of the network was recorded. In order to quantify the performance, information analysis was conducted as described in Section 5.3.2.

In this analysis, there are two different stimulus categories ($n = 2$) as explained in Section 5.3.2. In Figure 5.4(a), stimuli from the first category with a vertical straight edge on the left are shown in rows (b) and (d), while stimuli from the second category with a vertical straight

edge on the right are shown in rows (a) and (c). Since each category member was defined by its shape (hexagon or semicircle), shading (black or light grey), and retinal location (Location 1 or Location 2), there were $2^3 = 8$ members of transforms of each of the two stimulus categories. Individual Layer 3 neurons had to respond invariantly over the eight transforms of its preferred stimulus category, and not respond to any members of the other stimulus category, in order to carry maximum information about its preferred category.

Figure 5.5 shows the information analysis of the steady state response properties of Layer 3 neurons at the end of each stimulus presentation. Results are presented before and after training. Plot (a) shows the single cell information analysis. The maximum amount of information possible for the simulation is $log_2(n)$ where n is the number of stimulus categories = 2, that is 1 bit. Before training, no neurons reached 1 bit of information and in fact most neurons carried much less than 1 bit. However, after training, nearly all the neurons carried 1 bit of information. This result confirms that nearly all of the Layer 3 neurons had successfully learned to respond selectively to a vertical straight edge either on the left or on the right of an object boundary, regardless of the global shape, shading or retinal location of the object.

Plot (b) shows the multiple-cell information analysis. It is evident that training has led to an increase in the multiple cell information, which after training asymptotes to the maximum level of 1 bit with only one neuron included in the analysis. This is possible because, in the case of just two stimulus categories, the low or high firing responses of a single perfectly discriminating neuron will provide 1 bit of information about both stimulus categories. However, further inspection of the responses of Layer 3 neurons confirmed that some neurons had learned to respond selectively to objects with a straight vertical edge on the left boundary, while other neurons had learned to respond to a straight vertical edge on the right object boundary. This confirmed that the population of Layer 3 neurons learned to represent both of these stimulus categories.

Figure 5.5(b) shows the steady state firing rate responses of two typical Layer 3 neurons (17,46) and (33,23) at the end of each stimulus presentation. The firing rate responses are plotted before and after training. Plot (b1) shows the responses of the two Layer 3 neurons to all eight object stimuli from the second stimulus category, i.e. objects with a vertical straight edge on their right boundary. While plot (b2) shows the responses of the same two Layer 3 neurons to all eight object stimuli from the first stimulus category, i.e. objects with a vertical straight edge on their left boundary. The white circle plotted on each stimulus gives the idea of the size of the fan-in radius of the neurons in the input layer of the network. The results show that, after training, neuron (17,46) had learned to respond selectively to all objects with a vertical straight edge on the right, while neuron (33,23) had learned to respond to all objects with a vertical straight edge on the left. These observed firing rate responses in Layer 3 were similar to those experimentally observed in area V4 of the primate visual system (Pasupathy and Connor, 2001, 2002) and demonstrated in the previous simulation study reported in Chapter 2 (Eguchi et al., 2015). These are the kind of neuronal response characteristics needed to provide top-down modulation of border ownership neurons in Layer 1 (corresponding to V1/V2).

Since the study above uses exactly the same two shapes (hexagon and semicircle) for training and testing the network, there is a possibility that the responses of the developed cells are specific to the set of actual trained objects and might not generalise to novel objects not encountered during training. Therefore, in order to cross-validate the response characteristics of these neurons, the four novel shapes shown in Figure 5.4(b) are presented to the same trained network and the firing rates are recorded. In other words, the network was trained with the objects shown in Figure 5.4(a) and then tested with a set of four different novel shapes shown in Figure 5.4(b).

Figure 5.6 shows the firing rate responses of the two Layer 3 neurons, which were previously tested on familiar objects in Figure 5.5(b), at the end of each novel stimulus presentation before and after training. Similar to the original set of shapes used to train the network, each shape

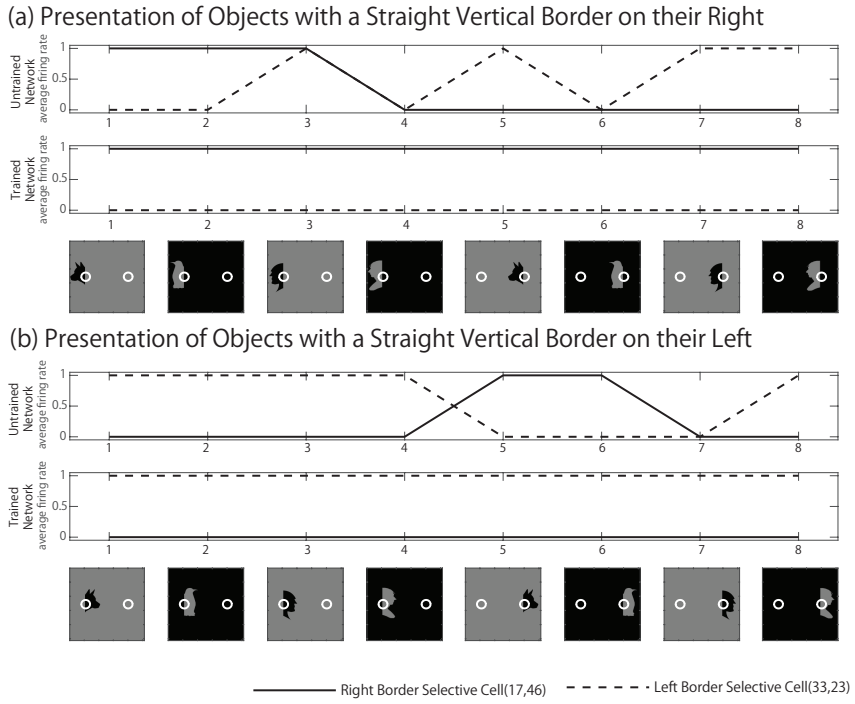

Figure 5.6: Cross-validation of the developed firing properties of neurons in Layer 3 with the set of novel shapes not presented during training as shown in Figure 5.4(b). Each of the four novel objects is presented in two retinal locations giving a total of eight novel stimulus presentations. This figure shows the firing rate responses of the same two Layer 3 neurons that were previously tested on familiar objects in Figure 5.5(b). Plot (a) shows the responses of the two Layer 3 neurons to all eight novel stimulus presentations with a vertical straight edge on their right boundary, and plot (b) shows the responses of the same two Layer 3 neurons to all eight novel stimulus presentations with a vertical straight edge on their left boundary. These results show that neuron (17,46) learned to respond selectively to all objects with a vertical straight edge on the right, while neuron (33,23) learned to respond to all objects with a vertical straight edge on the left. These results confirm that the developed firing properties of the cells are not specific to the set of trained objects and generalise to the set of novel objects not presented during training.

contains a vertical straight edge on either the right or left and is presented at two different retinal locations (i.e., Location 1 or Location 2). Figure 5.6(a) shows the responses of two Layer 3 neurons to all eight novel object stimuli from the second stimulus category, i.e. objects with a vertical straight edge on their right boundary. Before training, neuron (17,46) and neuron (33,23) both responded quite erratically to the different object stimuli. However, after training, neuron (17,46) responded to all of the objects with a vertical straight edge on their right, while neuron (33,23) did not respond to any of these stimuli. Plot (b) shows the responses of the same two Layer 3 neurons to all eight novel object stimuli from the first stimulus category, i.e. objects with a vertical straight edge on their left boundary. Before training, neurons (17,46) and (33,23) responded quite erratically to the different object stimuli. However, after training, neuron (33,23) responded to all of the objects with a vertical straight edge on their left, while neuron (17,46) did not respond to any of these stimuli. Taken together, these results show that neuron (17,46) learned to respond selectively to all novel objects with a vertical straight edge on the right, while neuron (33,23) learned to respond to all novel objects with a vertical straight edge on the left. Thus, the neurons continued to respond selectively to the presence of a vertical straight edge on either the left or the right of an object even if the objects are novel. This result confirms that the representations developed in the output layer of VisNet are not specific to the set of trained objects, but are in fact more generally selective to the presence of a vertical straight edge on either the left boundary or right boundary of an object.

I next tested whether Layer 1 neurons had developed the kind of border ownership representations reported by Zhou et al. (2000). In other words, I tested whether the feedback (top-down) connections newly implemented in VisNet enabled the activity in Layer 3 (corresponding to the

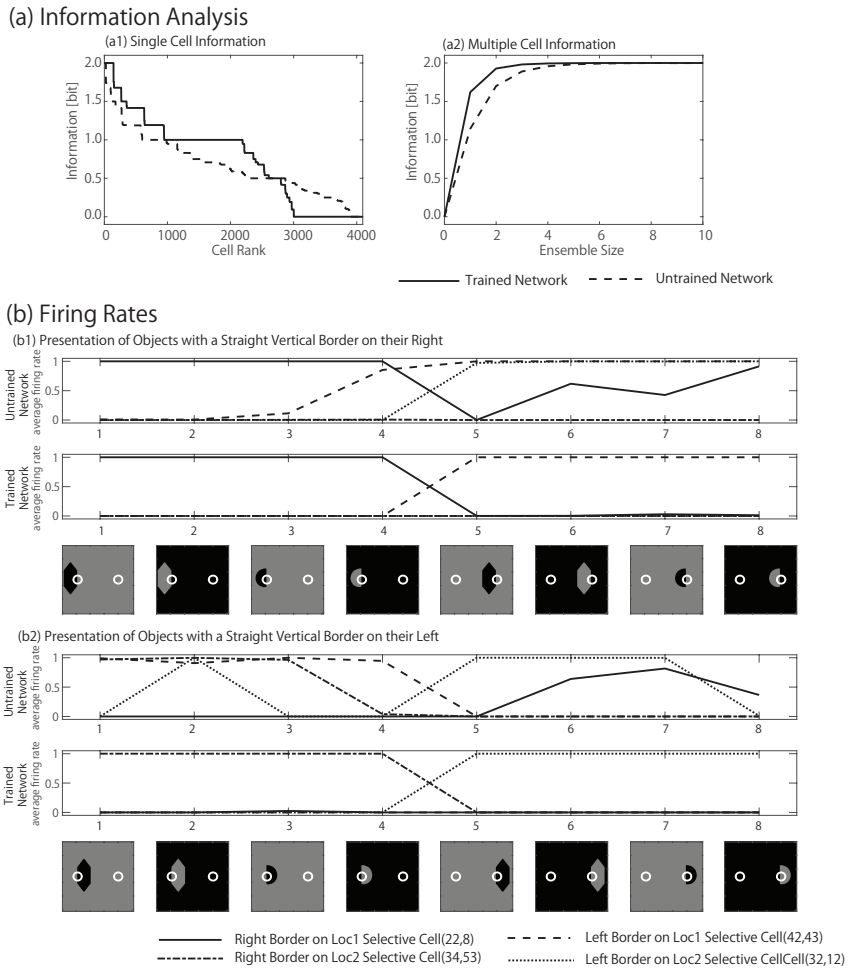

Figure 5.7: Steady state response properties of Layer 1 neurons at the end of each stimulus presentation of the familiar shapes that were used to train the network (Figure 5.4(a)). **(a) Information analysis:** Since Layer 1 neurons are not expected to develop translation invariance across different retinal locations, I computed the information carried by these neurons about whether the vertical straight edge in the object stimulus was from one of four stimulus categories: (i) Location 1 / left boundary, (ii) Location 1 / right boundary, (iii) Location 2 / left boundary, and (iv) Location 2 / right boundary. Since there are n = four stimulus categories, perfectly discriminating neurons carry a maximum of 2 bits of information. Plot (a1) shows the maximum single cell information carried by each of the 4096 neurons in Layer 1 about which one of the four stimulus categories was presented, where all of the neurons in Layer 1 are plotted along the abscissa in rank order. The result shows that these Layer 1 neurons have learned to respond with perfect selectivity to one of the four stimulus categories, thus providing the kind of border ownership representations experimentally observed in cortical visual area V1 by Zhou et al. (2000). Plot (a2) shows the multiple cell information carried by different sized (i.e. up to ten neurons) random ensembles of Layer 1 neurons that individually had high levels of single cell information. It is evident that training has led to an increase in the multiple cell information, which after training asymptotes to the maximum level of 2 bits with only two neurons included in the analysis. **(b) The firing rate responses of four Layer 1 neurons with maximum single cell information:** plot (b1) shows the responses of the four Layer 1 neurons to all eight object stimuli with a vertical straight edge on their right boundary. The first four stimuli 1-4 shown along the abscissa have the object presented in retinal Location 1, while the next four stimuli 5-8 have the object presented in retinal Location 2. Plot (b2) shows the responses of the same four Layer 1 neurons to all eight object stimuli with a vertical straight edge on their left boundary. The first four stimuli 1-4 shown along the abscissa have the object presented in retinal Location 1, while the next four stimuli 5-8 have the object presented in retinal Location 2. These results show that different Layer 1 neurons had learned to respond selectively to each of the four stimulus categories. These are the same kinds of border ownership representations found experimentally in primate visual area V1 by Zhou et al. (2000).

experimentally observed neural responses in primate visual area V4) to successfully modulate the responses of neurons in Layer 1 (corresponding to visual areas V1/V2) such that the Layer 1 simple cells representing vertical straight edges at either retinal Location 1 or 2 responded selectively depending on whether the vertical straight edge was on the left or right boundary of the object.

In order to quantify the performance of Layer 1 neurons, the information carried by the

steady state responses of these cells at the end of each stimulus presentation was computed. The results of this analysis are presented in Figure 5.7(a), where the information carried by Layer 1 neurons before and after training is shown. Layer 1 neurons are not expected to develop translation invariance across different retinal locations due to the small fan-in of connections from the retina. Therefore, information that was specific to either retinal Location 1 or Location 2 is computed. Specifically, the analysis calculated the information carried by the Layer 1 neurons about whether the vertical straight edge in the object stimulus presented to the network was an example from one of four stimulus categories: (i) the vertical straight edge is positioned at retinal Location 1 and is on the left boundary of the object presented there, (ii) the vertical straight edge is positioned at retinal Location 1 and is on the right boundary of the object presented there, (iii) the vertical straight edge is positioned at retinal Location 2 and is on the left boundary of the object presented there, and (iv) the vertical straight edge is positioned at retinal Location 2 and is on the right boundary of the object presented there. Since there are $n = 4$ stimulus categories, perfectly discriminating neurons carry a maximum of $log_2(n) = 2$ bits of information.

Figure 5.7(a1) shows the single cell information analysis. The plot shows the maximum information carried by each of the 4096 neurons in Layer 1 about which one of the four stimulus categories was presented. It can be seen that training the network has led to a large increase in the number of neurons carrying the maximum 2 bits of information. After training, 145 cells learned to carry the maximum single cell information. These Layer 1 neurons thus provide the kind of border ownership representations experimentally observed in cortical visual area V1 by Zhou et al. (2000). Plot (a2) shows the multiple-cell information analysis.

Although one may be confused with the unexpectedly good decoding performance even in the untrained network, this can be explained by the topologically established synaptic connections and the feedback connections from the neurons in the higher layer which has larger size of receptive field. However, as long as there is some statistical correlations between the input pattern and the output, the multiple-cell information analysis can better capture the information than the single-cell information analysis. Therefore, it is important to see whether the performances improved after the training or not. In this case, although the change is not as obvious as the case of the layer 3, it is still evident that training has led to an increase in the multiple cell information, which after training asymptotes to the maximum level of 2 bits with only two neurons included in the analysis.

Figure 5.7(b) shows the steady state firing rate responses of four typical Layer 1 neurons at the end of each stimulus presentation before and after training. Plot (b1) shows the responses of the four Layer 1 neurons to all eight object stimuli with a vertical straight edge on their right boundary. After training, neuron (22,8) responded selectively to all of the objects with a vertical straight edge on their right boundary aligned with retinal Location 1, while neuron (42,43) responded to all of the objects with a vertical straight edge on their right boundary aligned with retinal Location 2. Plot (b2) shows the responses of the same four Layer 1 neurons to all eight object stimuli with a vertical straight edge on their left boundary. After training, neuron (34,53) responded selectively to all of the objects with a vertical straight edge on their left boundary aligned with retinal Location 1, while neuron (32,12) responded to all of the objects with a vertical straight edge on their left boundary aligned with retinal Location 2.

The developed firing properties were next cross-validated by testing the same trained network on the novel set of shapes shown in Figure 5.4(b). Figure 5.8(b) shows the firing rate responses of the same four Layer 1 neurons that were previously tested on familiar objects in Figure 5.7(b). Plot (b1) shows the responses of the four Layer 1 neurons to all eight object stimuli with a vertical straight edge on their right boundary. The first four stimuli 1-4 shown along the abscissa have the object presented in retinal Location 1, while the next four stimuli 5-8 have the object presented in retinal Location 2. Plot (b2) shows the responses of the same four Layer 1 neurons to all eight object stimuli with a vertical straight edge on their left boundary.

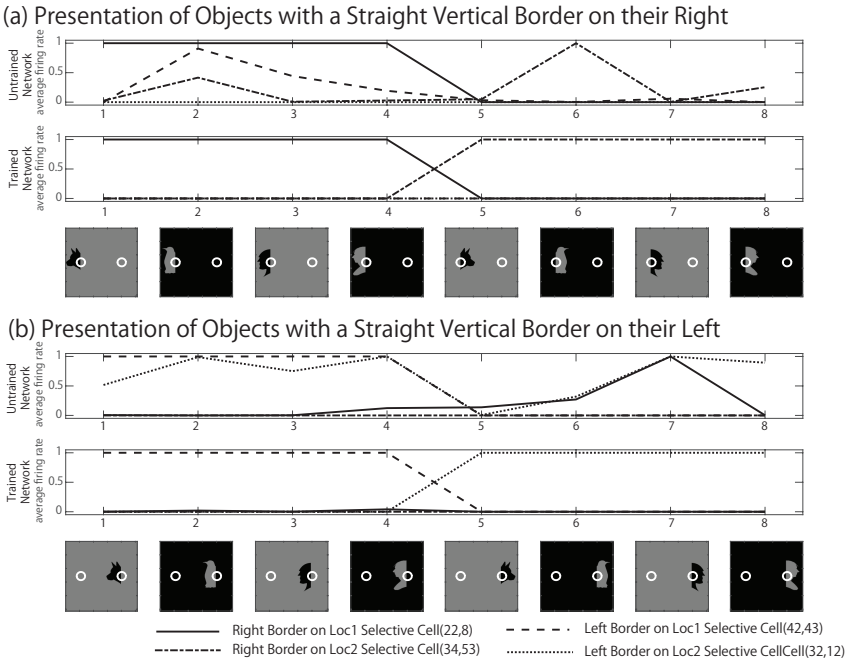

Figure 5.8: Cross-validation of the developed firing properties of neurons in Layer 1 with the set of novel shapes not presented during training as shown in Figure 5.4(b). Each of the four novel objects is presented in two retinal locations giving a total of eight novel stimulus presentations. This figure shows the firing rate responses of the same four Layer 1 neurons that were previously tested on familiar objects in Figure 5.7(b). Plot (a) shows the responses of the four Layer 1 neurons to all eight novel object stimuli with a vertical straight edge on their right boundary. The first four stimuli 1-4 shown along the abscissa have the object presented in retinal Location 1, while the next four stimuli 5-8 have the object presented in retinal Location 2. Plot (b) shows the responses of the same four Layer 1 neurons to all eight novel object stimuli with a vertical straight edge on their left boundary. The first four stimuli 1-4 shown along the abscissa have the object presented in retinal Location 1, while the next four stimuli 5-8 have the object presented in retinal Location 2. It is evident that each of the four Layer 1 neurons has learned to respond to one of the four stimulus categories: (i) Location 1 / left boundary, (ii) Location 1 / right boundary, (iii) Location 2 / left boundary, and (iv) Location 2 / right boundary. These results confirm that the border ownership representations developed in Layer 1 were not specific to the set of trained objects and generalise to the set of novel objects not presented during training.

The first four stimuli 1-4 shown along the abscissa have the object presented in retinal Location 1, while the next four stimuli 5-8 have the object presented in retinal Location 2. It can be seen that each of the four Layer 1 neurons responds selectively to one of the four stimulus categories: (i) Location 1 / left boundary, (ii) Location 1 / right boundary, (iii) Location 2 / left boundary, and (iv) Location 2 / right boundary. These results confirm that the border ownership representations developed in Layer 1 are not specific to the set of trained objects, and are in fact able to generalise to the set of novel object shapes. Thus, different Layer 1 neurons had learned to respond selectively to the presence of a vertical straight edge on either the left boundary or right boundary of an object when the edge is aligned with a particular retinal location. These are the same kinds of border ownership representations reported in the neurophysiology study of primate visual area V1 carried out by Zhou et al. (2000).

5.4.1.2 Dynamical firing properties of cells in layer 1 during each stimulus presentation: time course of the emergence of border ownership signals

Sugihara et al. (2011) reported that the representation of border ownership in primate visual area V1, i.e. the selective modulation of the responses of V1 neurons that encode vertical straight edges by whether the edge appears on the left or right boundary of an object, begins to appear at around 61ms after the presentation of the visual stimulus. I hypothesise that this gradual emergence of the border ownership signal in area V1 is due to the time it takes for visual signals to propagate up to higher visual areas such as V4, where neurons may represent a vertical

straight edge on either the left or right of an object boundary across different retinal locations, and then to propagate back down to modulate the activities of neurons in area V1. This proposal was investigated computationally by recording the temporal evolution of the responses of border ownership neurons in Layer 1 of the trained VisNet model during 300ms stimulus presentations.

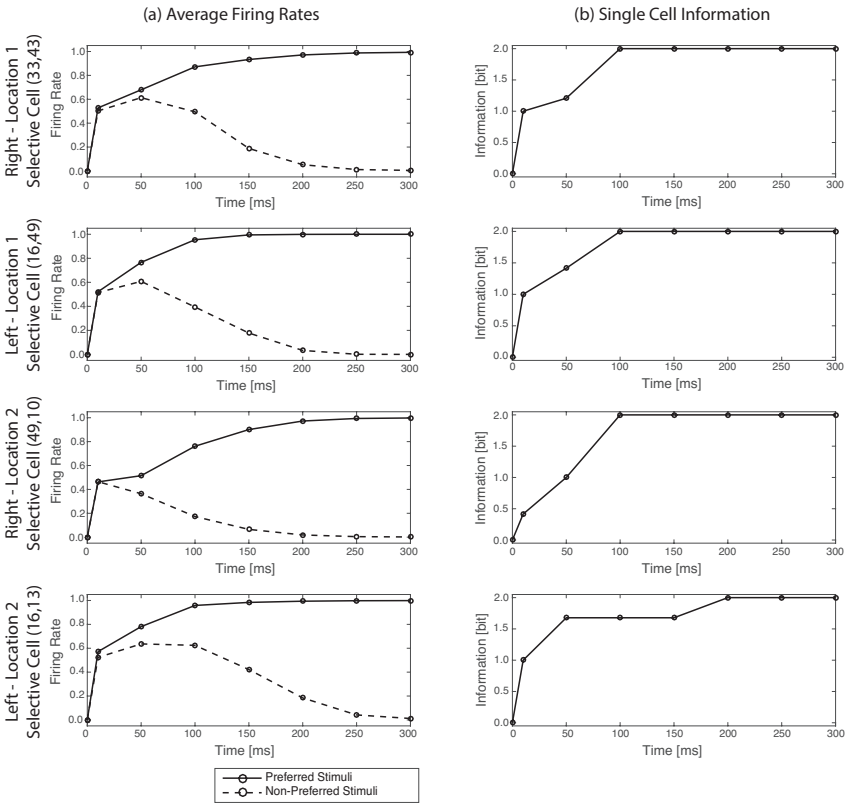

Figure 5.9: The temporal evolution of border ownership representations conveyed by four typical Layer 1 neurons during the 300ms time course of stimulus presentations. The network was trained and tested with the objects shown in Figure 5.4. Results are shown after training. Each row shows results for one of the four neurons, where each neuron is tuned to a different border ownership category as follows. Row 1 (top row): a neuron tuned to a vertical straight edge on the right object boundary aligned with retinal Location 1, Row 2: a neuron tuned to a vertical straight edge on the left object boundary aligned with retinal Location 1, Row 3: a neuron tuned to a vertical straight edge on the right object boundary aligned with retinal Location 2, and Row 4 (bottom row): a neuron tuned to a vertical straight edge on the left object boundary aligned with retinal Location 2. Column (a) shows the average firing rates of the four neurons plotted over the 300ms time courses of the stimulus presentations. Each subplot shows the average responses of the neuron to the members of its preferred stimulus category (solid line) and the members of its three non-preferred stimulus categories (dashed line). For all four neurons, it can be seen that their firing responses begin to strongly differentiate between the preferred and non-preferred stimulus categories at around 50ms. By the end of the stimulus presentation at 300ms, the neurons show complete differentiation between the preferred and non-preferred stimulus categories. Column (b) shows the average single cell information carried by the four neurons about their preferred stimulus category plotted over the 300ms time courses of the stimulus presentations.

Figure 5.9 shows the dynamical evolution through time of the border ownership representations conveyed by four typical neurons in Layer 1 during the 300ms time course of stimulus presentations. The results are shown after training has established border ownership representations in Layer 1. Each row shows results for one of the four neurons, where each neuron is tuned to a different border ownership category as follows: (Row 1) the neuron is tuned to a vertical straight edge on the right object boundary aligned with retinal Location 1, (Row 2) the neuron is tuned to a vertical straight edge on the left object boundary aligned with retinal Location 1, (Row 3) the neuron is tuned to a vertical straight edge on the right object boundary aligned with retinal Location 2, and (Row 4) the neuron is tuned to a vertical straight edge on the left object boundary aligned with retinal Location 2. Column (a) shows the average

responses of each neuron to the members of its preferred stimulus category (solid line) and the members of its three non-preferred stimulus categories (dashed line) plotted over the 300ms time courses of the stimulus presentations. It can be seen that the firing responses of all four neurons begin to strongly differentiate between their preferred and non-preferred stimulus categories by about 50ms after the start of stimulus presentation. By the end of the stimulus presentation at 300ms, the responses of the neurons fully differentiate between their preferred and non-preferred stimulus categories. Column (b) shows the average single cell information carried by the four neurons about their preferred stimulus category plotted over the 300ms time courses of the stimulus presentations. Consistent with the firing rate plots, there is a monotonic increase in the information carried by each of the four neurons during the 300ms time course of stimulus presentation.

The simulation results show how the border ownership representations gradually emerge in Layer 1 over the time course of 300ms during stimulus presentation. Near the beginning of the stimulus presentation, the Layer 1 neurons merely represent the presence of a straight vertical edge at a particular retinal Location 1 or 2. The Layer 1 neurons have not begun to carry information about border ownership at this point. However, as the visual signals propagate up to Layer 3 and back down again to Layer 1, these top down signals from Layer 3 begin to strongly modulate the activities of Layer 1 neurons at around 50ms. The effect of this top down modulation is to drive the activity of the Layer 1 neurons to represent the border ownership categories. These simulation results are qualitatively similar to the temporal evolution of border ownership representations reported by Sugihara et al. (2011) and Jehee et al. (2007).

5.4.2 Study 2: failure of the model under more general stimulus conditions

In the above simulations, the model was tested by presenting a single object to the network at a time. However, the primate visual system is usually presented with multiple objects simultaneously in real world scenes. This more realistic situation actually exposes a weakness in our current rate-coded model. As explained earlier, Pasupathy and Connor (2002) have reported that the local boundary representations observed in area V4 such as Φ^{V4}_{Left} and Φ^{V4}_{Right} are translation invariant across different retinal positions over a modest range. This may lead to a lack of specificity with respect to retinal location in the contextual information that is back-projected to the earlier layers of the network. This will be problematic, for example, when two objects that contain a straight vertical contour on different object sides (left or right) are presented to the network simultaneously. In this case, both Φ^{V4}_{Left} and Φ^{V4}_{Right} will be activated in the higher V4 layer. However, Φ^{V4}_{Left} and Φ^{V4}_{Right} will top-down modulate V1/V2 simple cells representing a vertical straight edge on the left and right object boundaries, respectively, across *all* trained retinal locations. Thus, the top-down modulation of V1/V2 neuronal firing rates is not specific to retinal location. This effectively destroys the local border ownership (binding) information carried by the V1/V2 neurons. This important argument is elaborated in more detail next.

5.4.2.1 Proposed mechanism by which border ownership information carried by V1/V2 neurons in the rate-coded model may be lost when the network is presented with multiple visual objects

Suppose that, during testing of the model, an object that contains a straight vertical contour on the left is presented with that contour positioned at a retinal Location 1, and another object that contains a straight vertical contour on the right is presented with that contour at a retinal Location 2 as shown in Figure 5.10. In this case, as explained in the figure, both Φ^{V4}_{Left} and Φ^{V4}_{Right} should become highly activated at the same time. However, during training, the subpopulations Φ^{V4}_{Left} and Φ^{V4}_{Right} are trained to respond invariantly as an object is translated across different retinal locations. This means that both Φ^{V4}_{Left} and Φ^{V4}_{Right} each end up with strong bi-directional

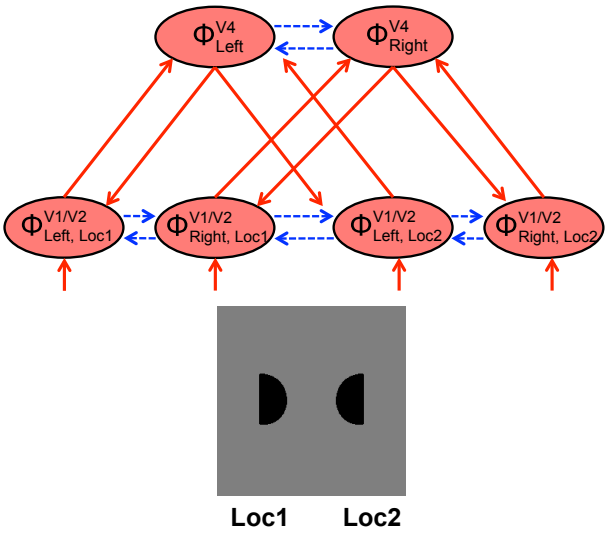

Figure 5.10: Hypothesised modulation of edge detecting simple cells in lower layers V1/V2 of the rate-coded model by top-down signals from higher layer V4 neurons representing boundary contour elements when two visual object stimuli are presented simultaneously. Assume that during testing of the model, an object with a straight vertical border on its *left* is presented with this border positioned at retinal location *1*, and another object with a straight vertical border on its *right* is presented with this border positioned at retinal location *2*. Ascending visual input initially stimulates all subsets of V1/V2 neurons, which represent a vertical straight edge at retinal location 1 ($\Phi^{V1/V2}_{Left,Loc1}$ and $\Phi^{V1/V2}_{Right,Loc1}$), and a vertical straight edge at retinal location 2 ($\Phi^{V1/V2}_{Left,Loc2}$ and $\Phi^{V1/V2}_{Right,Loc2}$). In layer V4, V4 neurons that represent a vertical straight edge on the left of an object (Φ^{V4}_{Left}) are stimulated by ascending visual signals from the object in retinal Location 1, while V4 neurons that represent a vertical straight edge on the right of an object (Φ^{V4}_{Right}) are stimulated by ascending visual signals from the object in retinal Location 2. However, the subpopulations Φ^{V4}_{Left} and Φ^{V4}_{Right} have each been trained to respond with translation invariance across all trained retinal locations, and so have developed strong bi-directional (i.e. bottom-up and top-down) polysynaptic connections with subpopulations of V1/V2 simple cells representing all retinal locations. Consequently, Φ^{V4}_{Left} and Φ^{V4}_{Right} will top-down modulate V1/V2 simple cells representing a vertical straight edge on the left and right object boundaries, respectively, across all trained retinal locations. In this case, all of the V1/V2 cells shown in the figure end up receiving a similar amount of bottom-up and top-down excitatory input. Both subpopulations $\Phi^{V1/V2}_{Left,Loc1}$ and $\Phi^{V1/V2}_{Right,Loc1}$ will be active at retinal Location 1, and both subpopulations $\Phi^{V1/V2}_{Left,Loc2}$ and $\Phi^{V1/V2}_{Right,Loc2}$ will be active at retinal Location 2. Thus, when more than one visual object is presented to the model, the V1/V2 neurons $\Phi^{V1/V2}_{Left,Loc1}$, $\Phi^{V1/V2}_{Right,Loc1}$, $\Phi^{V1/V2}_{Left,Loc2}$ and $\Phi^{V1/V2}_{Right,Loc2}$ may fail to represent the border ownership (binding) information.

polysynaptic connections with subpopulations of V1/V2 simple cells representing a straight vertical contour at all trained retinal locations. In this case, the top-down signals from Φ^{V4}_{Left} and Φ^{V4}_{Right} each modulate the responses of V1/V2 simple cells across both retinal Locations 1 and 2.

In this situation, as explained earlier, $\Phi^{V1/V2}_{Left,Loc1}$ will become strongly activated by receiving both the feedforward signals that indicate that a straight vertical contour is present at retinal Location 1 and the feedback signals from Φ^{V4}_{Left} that indicate that the straight vertical contour is on the left side of the object. Similarly, $\Phi^{V1/V2}_{Right,Loc2}$ will become strongly activated by receiving both the feedforward signals that indicate that a straight vertical contour is present at retinal Location 2 and the feedback signals from Φ^{V4}_{Right} that indicate that the straight vertical contour is on the right side of the object.

However, the problem is that the other sets of neurons, $\Phi^{V1/V2}_{Left,Loc2}$ and $\Phi^{V1/V2}_{Right,Loc1}$ may also be strongly activated. This is because both Φ^{V4}_{Left} and Φ^{V4}_{Right} have strong bi-directional polysynaptic connections with subpopulations of V1/V2 simple cells representing a straight vertical contour at both trained retinal Locations 1 and 2. More specifically, $\Phi^{V1/V2}_{Left,Loc2}$ may receive not only the feedforward signals that indicate that a straight vertical contour is present at the retinal location 2, but also the feedback signals from Φ^{V4}_{Left} which are actually activated by the presence of the other object with a straight vertical contour on the left at retinal Location

1. As a result, $\Phi^{V1/V2}_{Left,Loc2}$ may become activated even though no object with a straight vertical contour on the left is ever presented at retinal Location 2. Similarly, $\Phi^{V1/V2}_{Right,Loc1}$ may receive not only the feedforward signals that indicate that the straight vertical contour is present at retinal Location 1, but also the feedback signals from Φ^{V4}_{Right} which are activated by the presence of the other object with a straight vertical contour on the right at retinal Location 2. As a result, $\Phi^{V1/V2}_{Right,Loc1}$ may become activated even though no object with a straight vertical contour on the right is ever presented at retinal Location 1.

The upshot of this is that when the two objects are presented to the model simultaneously, all of the V1/V2 subpopulations $\Phi^{V1/V2}_{Left,Loc1}$, $\Phi^{V1/V2}_{Right,Loc1}$, $\Phi^{V1/V2}_{Left,Loc2}$ and $\Phi^{V1/V2}_{Right,Loc2}$ may become active. In this case, these subpopulations of V1/V2 neurons will fail to represent the border ownership (binding) information. This will be a general problem for the current rate-coded formulation of the model when presented with visual input from more realistic scenes containing multiple objects.

5.4.2.2 Results

In this section, the model was trained with the set of objects shown in Figure 5.4, where these objects were presented to the network one at a time during training as described in the simulations above. However, the network was then tested with *two* objects shown together during each visual presentation, where the set of test images shown in Figure 5.11 was used . The steady state firing responses of Layer 1 neurons at the end of each such visual presentation was analysed. These results were compared with those reported above in which only a single object was presented to the network at a time during testing. In order to facilitate comparison of the results for the two test situations, in each case I analysed how much information Layer 1 neurons carried about border ownership stimulus categories (straight vertical edges on the left or right object boundaries) that were associated with retinal Location 1.

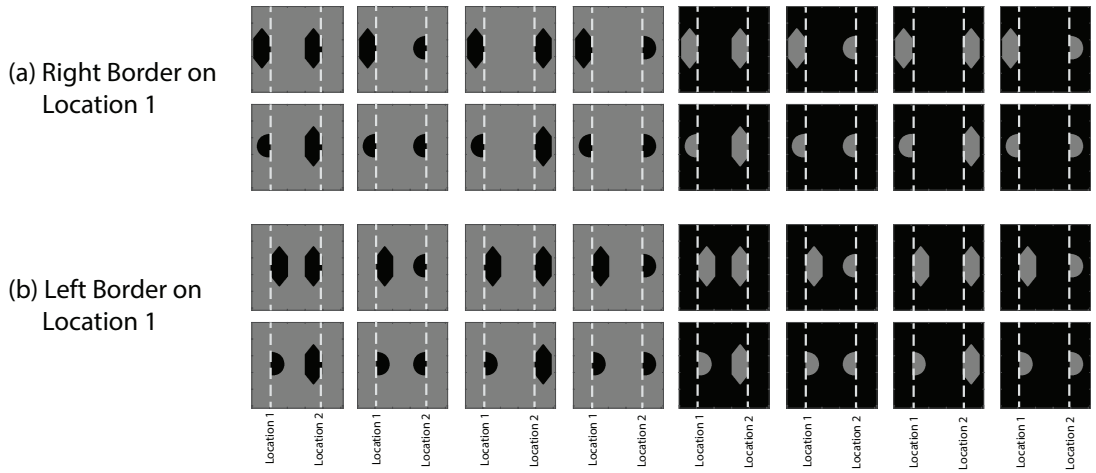

Figure 5.11: The set of visual stimuli used to test the performance of the network when two objects are presented simultaneously during testing. There are two categories of visual stimuli. The first stimulus category consists of all possible combinations two objects where one of the objects has a vertical straight edge on its *right* boundary which is positioned at retinal Location 1. The second stimulus category consists of all possible combinations two objects where one of the objects has a vertical straight edge on its *left* boundary which is positioned at retinal Location 1.

The set of images used for testing the network with two objects at a time are shown in Figure 5.11. There are two different stimulus categories. The first stimulus category, shown in Figure 5.11(a), consists of all possible combinations two objects where one of the objects has a vertical straight edge on its *right* boundary which is positioned at retinal Location 1. On the other hand, the second stimulus category, shown in Figure 5.11(b), consists of all possible combinations two

objects where one of the objects has a vertical straight edge on its *left* boundary which is positioned at retinal Location 1. Each of the two stimulus categories undergoes 16 transforms, which are due to variations in the following four stimulus features: 2 different shapes (semicircle or hexagon) at retinal Location 1 $\times$ 2 different shapes (semicircle or hexagon) at retinal Location 2 $\times$ 2 sides of an object (left or right) on which a straight vertical edge may occur at retinal Location 2 $\times$ 2 kinds of shading contrast between objects and background.

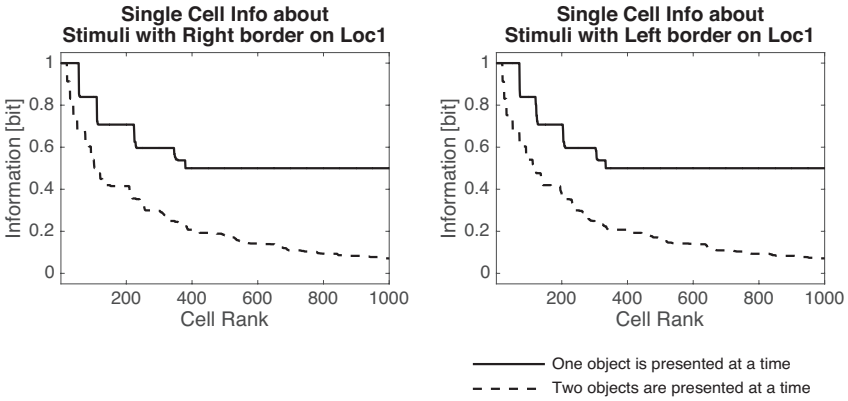

Figure 5.12: A quantitative comparison of the border ownership information carried by Layer 1 neurons when the network is tested with objects shown individually (solid line) or when tested on two objects presented together (dashed line). The network performance is assessed using single-cell information analysis. The information analysis is applied to the steady state firing responses of Layer 1 neurons at the end of each stimulus presentation. In both test situations, the model was initially trained with the individual objects shown in Figure 5.4, as described earlier in this chapter. **Solid lines:** the performance of the model when tested with the objects shown in Figure 5.4 presented one at a time during testing. The single cell information carried by each Layer 1 neuron about the four stimulus categories previously described in Figure 5.7(a) was computed. In this figure the maximum single-cell information ($log_2(4) = 2$) has been rescaled to 1. The information carried by Layer 1 neurons about stimulus category (i) is shown in the right plot, while information carried by Layer 1 neurons about the stimulus category (ii) is shown in the left plot. It can be seen that when the network is tested on a single object at a time, many Layer 1 neurons reach the theoretical maximum level of information about border ownership. **Dashed lines:** the performance of the model when tested with *two* objects shown together during each visual presentation using the test images shown in Figure 5.11. Here the single cell information carried by each Layer 1 neuron about the two stimulus categories described in Figure 5.11 was computed. Since there are two stimulus categories, neurons may carry up to a maximum of 1 bit of information. The information carried by Layer 1 neurons about the first stimulus category is shown in the left plot, while information carried by Layer 1 neurons about the second stimulus category is shown in the right plot. It can be seen that when the network is tested on two objects at a time, there is a large drop in the levels of single cell information carried by Layer 1 neurons about border ownership compared to when the network is tested on individual objects, with far fewer Layer 1 neurons reaching the theoretical maximum of 1 bit for this test case.

Figure 5.12 compares the border ownership information carried by Layer 1 neurons (corresponding to visual areas V1/V2) when the network is tested with objects shown individually (solid line) or when tested on two objects presented together (dashed line). The performance of the network using single-cell information analysis is assessed. The information analysis is applied to the steady state firing responses of Layer 1 neurons at the end of each stimulus presentation.

The solid lines in Figure 5.12 show the performance of the model when tested with the objects shown in Figure 5.4 presented one at a time during testing. I computed the single cell information carried by each Layer 1 neuron about one of the four stimulus categories separately while I previously plotted them altogether in Figure 5.7(a). That is, the information about whether the vertical straight edge in the object stimulus was from one of the following four stimulus categories was computed: (i) a vertical straight edge on the left object boundary positioned at retinal Location 1, (ii) a vertical straight edge on the right object boundary positioned at retinal Location 1, (iii) a vertical straight edge on the left object boundary positioned at retinal Location 2, and (iv) a vertical straight edge on the right object boundary positioned at retinal Location 2. Since there are four stimulus categories, perfectly discriminating neurons carry a maximum of 2 bits of information. However, in this figure the maximum single-cell information has been rescaled to 1. Results for the two stimulus categories (i) and (ii), which

are associated with retinal Location 1, are plotted. The information carried by Layer 1 neurons about stimulus category (i) is shown in the right plot, while information carried by Layer 1 neurons about the stimulus category (ii) is shown in the left plot. It can be seen that when the network is tested on a single object at a time, a large number of Layer 1 neurons (i.e. 55 neurons and 71 neurons for the stimulus category (i) and (ii), respectively) reach the theoretical maximum level of information about border ownership.

The dashed lines in Figure 5.12 show the performance of the model when tested with *two* objects shown together during each visual presentation using the test images shown in Figure 5.11. Here the single cell information carried by the Layer 1 neurons about the two stimulus categories described in Figure 5.11, which are both associated with retinal Location 1, was computed. The first stimulus category includes all combinations of two objects where one of the objects has a vertical straight edge on its *right* boundary which is positioned at retinal Location 1, while the second stimulus category includes all combinations of two objects where one of the objects has a vertical straight edge on its *left* boundary positioned at retinal Location 1. Since there are two stimulus categories, neurons may carry up to a maximum of 1 bit of information. The information carried by Layer 1 neurons about the first stimulus category is shown in the left plot, while information carried by Layer 1 neurons about the second stimulus category is shown in the right plot. It can be seen that when the network is tested on two objects at a time, there is a large drop in the levels of single cell information carried by Layer 1 neurons about border ownership compared to when the network is tested on individual objects, with far fewer Layer 1 neurons reaching the theoretical maximum of 1 bit for this test case. This result supported our above prediction that border ownership information carried by Layer 1 (V1/V2) neurons in the rate-coded model may be lost when the network is presented with multiple visual objects during testing.

It can be seen in Figure 5.12, however, that a small number of Layer 1 neurons did still reach the maximum of 1 bit of information (19 neurons for both stimulus categories (i) and (ii)). How might this happen if both of the Layer 3 subpopulations Φ^{V4}_{Left} and Φ^{V4}_{Right} were completely translation invariant with strong bi-directional (bottom-up and top-down) polysynaptic connections with subpopulations of Layer 1 (V1/V2) simple cells representing a straight vertical contour at both trained retinal Locations 1 and 2? To understand this, the firing properties of neurons in Layer 2 after training were investigated. Although not shown here, it was found that, due to the limited feedforward fan-in of synaptic connections from the input 'retina', some of the Layer 2 neurons had learned to respond to a vertical straight edge either on the left object boundary or right object boundary at only a single retinal location. These location-specific Layer 2 neurons were then able to directly modulate the Layer 1 neurons representing that particular retinal location. This would allow these Layer 1 neurons to continue to respond selectively to whether a vertical straight edge was on either the left or right boundary of an object presented at that retinal location regardless of the presence of another object simultaneously presented elsewhere. However, this effect was rather minor given that the great majority of Layer 1 neurons lost their border ownership selectivity when two objects were presented during testing.

5.5 Discussion

I investigated through computer simulation how top-down connections may play a fundamental role in the development of border ownership representations in the early cortical visual layers V1/V2. In terms of the novelty, this work is different from previous modelling studies that have already proposed hypothetical neural circuits for such coding in that I investigated how such circuits may develop using a biologically plausible, local, trace learning rule to modify the synaptic connectivity during visual experience.

A number of modelling studies have previously considered the role of top-down signals in visual information processing. For example, as discussed in section 5.2, some authors have

proposed that top-down connections might implement attention to objects during visual search (Deco and Lee, 2002; Deco and Rolls, 2004). However, in these previous modelling studies the top-down connections were only introduced after the initial training phase was completed, and hence the self-organisation of the synaptic connections throughout the network relied on purely feedforward visual processing. Consequently, the top-down connections did not affect the visual representations that developed in the network during visually-guided learning. In another modelling study carried out by Renart et al. (1999), top-down connections were able to influence the recall of visual representations in a linked attractor network comprised of multiple cortical modules (Rolls, 2008). However, the representations in this attractor network were hand specified during an initial stage of supervised learning, and did not self-organise using unsupervised competitive learning. Thus, again, the top-down connections were not able to influence the nature of the visual representations that developed. In our own model presented in this chapter, the top-down connections are present during both training and testing. Consequently, the top-down connections played a critical role in the self-organisation of border ownership representations in Layer 1 during the initial unsupervised competitive learning. In this case, each neuron receives signals from both afferent bottom-up and top-down connections, which self-organise simultaneously during learning. This allows the network to develop representations that depend on a precise learned combination of bottom-up and top-down signals.

The simulations reported in this chapter have demonstrated how top-down connections may help to guide competitive learning in lower layers, thus driving the formation of lower level (border ownership) visual representations in V1/V2 that are modulated by higher level (object boundary element) representations in V4. More precisely, it has been shown that simple cells in area V1 representing a vertical straight edge at a particular retinal location can learn to be modulated by top-down connections from higher level representations of object shape in, for example, area V4 (Pasupathy and Connor, 2001, 2002). However, more importantly, I also identified the limitation of the mechanism within a rate-coded model when trying to simulate the results of the neurophysiological studies that have shown that border-ownership selective neurons for single-figure displays generally are so also for multi-figure displays (Qiu et al., 2007; Martin and von der Heydt, 2015). In the second half of the simulation studies, I investigated how the rate-coded model presented in this chapter fails under more general stimulus conditions, in which more than one object stimulus is presented to the network at the same time after training.

The result suggests that the incorporation of additional top-down connections, although necessary, is not sufficient by itself to allow the network to develop robust border ownership representations in the early layers and thus solve this kind of feature binding problem. Our model failed because the current model of the network is not able to specify which features are part of which objects. Therefore, I propose that it is important to have a form of binding neuron (e.g., border ownership neuron in V1/V2) that responds if and only if the neurons representing the low-level feature such as simple oriented bars are actually participating in driving the neurons representing the high-level feature. The binding neuron should not respond if the neurons representing the low-level feature and the neurons representing the high-level feature just happen to be co-active, where the former are not actually driving the latter. Such unrelated co-activation of low and high-level features might occur, for example, because of the presence of multiple similar objects within a complex natural scene. Then, the question is what further biological details is needed to be incorporated into the model to allow it to form such robust border ownership representations under more general stimulus conditions.

A biological detail that is not implemented in the current model is cortical magnification. It is known that mammalian brains process visual input in a highly non-uniform manner. Specifically, the Ganglion cells in the retina sample the visual input at a higher resolution in the fovea than the periphery (Wassle et al., 1990), which gives rise to a distorted visual field representation in V1 where the fovea has a higher "cortical magnification factor", i.e. more V1

neurons processing foveal input than the peripheral visual field (Daniel and Whitteridge, 1961; Cowey and Rolls, 1974). Subsequent neural processing, with a simple Gaussian sampling of the representation that is laid out across the surface of area V1, results in an asymmetry of central V4 receptive fields as well (Motter, 2009). The question is whether cortical magnification may play any role in the development of border ownership representations.

Our laboratory has previously investigated the effects of implementing a cortical magnification factor within a purely feedforward neural network model of primate visual object recognition (Trappenberg et al., 2002). It was found that when the objects were presented against a simple blank background then neurons in the upper cortical layer responded to their preferred objects across a wide region of the retina. In this scenario, trace learning can continue to operate normally as an object translates across different locations on the retina. Neurophysiological evidence for trace learning has been reported by Cox et al. (2005). Moreover, past simulation studies have found that the trace learning mechanism is quite robust to the way in which the eyes saccade around the visual scene, and is in fact enhanced by more randomised exploration of a scene (Rolls and Milward, 2000). Consequently, I would not expect the introduction of a cortical magnification factor into the border ownership simulations reported in this chapter to prevent the model from operating in the same qualitative manner as described above. However, in the simulation study of Trappenberg et al. (2002), it was also found that, with a cortical magnification factor, if the objects were presented against cluttered backgrounds then the receptive fields of neurons in the upper layer shrunk down around the fovea due to competition from the background features. These simulation results reflected what had been previously observed in a primate neurophysiology study carried out by Rolls et al. (2003), in which the receptive fields of object-selective neurons in the primate temporal visual cortex reduced down to approximately the size of the object when it was presented against a natural scene. It should also be noted that the neurophysiology studies of V4 shape selective neurons (Pasupathy and Connor, 2001) investigated the responses of these neurons to shapes that were presented in isolation. Our border ownership model sought to replicate the development of these V4 shape selective firing properties in Layer 3 under similar viewing conditions - that is, the network was trained on one shape at a time presented against a blank background. It remains to be seen how the firing properties of these shape selective neurons in area V4 of the primate brain might be affected when the shapes are presented within natural scenes. The cortical magnification factor may play an important role in this situation. Addressing these issues will require a combination of further neurophysiology and modelling studies.

Another biological detail that is not implemented in the current model is the spike dynamics of neurons, which will be investigated in detail in Chapter 6 and Chapter 7. I hypothesise that extending the model with spiking neural network would solve the issue. The current rate-coded model only represents the average firing rate of each neuron, and not the actual timings of the electrical pulses emitted by neurons in the brain. The architecture and operation of neural tissue in the visual cortex of primates differs from the VisNet model implemented in this chapter in the following important ways. Firstly, real neurons in the brain communicate by emitting and receiving electrical pulses called action potentials or 'spikes'. Secondly, the way in which synapses are strengthened and weakened during learning is dependent on the timings of the spikes emitted by the pre- and post-synaptic neurons (Bi and Poo, 1998; Markram et al., 1997). For example, in the brain, a synapse may be strengthened if the pre-synaptic spike occurs about 20ms before the post-synaptic spike, but weakened if the pre-synaptic spike occurs about 20 ms after the post-synaptic spike. This is known as spike-timeing-dependent plasticity (STDP). Thirdly, the electrical pulses can take several milliseconds to travel along an axon from one neuron to the next, with different axonal connections having different time delays.

Physiological studies have shown that neural synchrony is unrelated, or at best weakly related, to contour grouping (Roelfsema et al., 2004; Martin and von der Heydt, 2015). On the other hand, if distributions of axonal delays between neurons are incorporated into a model,

then this can give rise to a phenomenon known as 'polychronization' (Izhikevich, 2006). This phenomenon involves the network learning many memory patterns, each of which takes the form of a repeating temporal loop of neuronal firings. These temporal memory loops self-organise automatically when STDP is used to modify the strengths of synapses in a recurrently connected spiking network with randomised distributions of axonal conduction delays between neurons. Polychronization can dramatically increase the selectivity of neurons and increase the memory capacity of a network. I hypothesise that such a spiking model may develop border ownership neurons in layer 1 (corresponding to V1/V2) that respond selectively to a vertical straight edge on either the left or right boundary of an object at the neuron's preferred retinal location, regardless of the presence of other objects at different retinal locations. More generally, I propose that these biological elements will be needed to model how the primate visual system solves 'the binding problem' in vision. Consequently, in future work I will explore how border ownership representations may develop in a new spiking neural network version of the VisNet model, which incorporates bottom-up and top-down connections, distributions of axonal transmission delays, and spike-timeing-dependent plasticity (STDP).

Chapter 6

Polychronization and Feature Binding in a Spiking Neural Network Model

In this chapter, I present a hierarchical neural network model, in which subpopulations of neurons develop fixed and regularly repeating temporal chains of spikes (polychronization), which respond specifically to randomised Poisson spike trains representing the input training images. The performance is improved by including top-down and lateral synaptic connections, as well as introducing multiple synaptic contacts between each pair of pre- and postsynaptic neurons, with different synaptic contacts having different axonal delays. Spike-Timing-Dependent Plasticity (STDP) thus allows the model to select the most effective axonal transmission delay between neurons. Furthermore, neurons representing the binding relationship between low-level and high-level visual features emerge through visually-guided learning. This provides a solution to the classic feature binding problem in visual neuroscience and leads to a new hypothesis concerning how information about visual features at every spatial scale may be projected upwards through successive neuronal layers. We name this hypothetical upward projection of information the holographic principle.

6.1 Introduction

Many early neural network models of brain function including VisNet assumed that neurons transmit information exclusively through modulation of their mean firing rates. These 'rate-coded' models represented only the current average firing rate of each neuron, and did not explicitly represent the timings of the action potentials or 'spikes' emitted by cells. However, in modern literature the precise timing of spikes has been proposed to strongly contribute to encoding in the brain (Fujii et al., 1996; Akolkar et al., 2015; Nikoli et al., 2013). Consistent with this view, there is growing evidence from neurophysiological studies supporting the importance of spike-timing dynamics in the brain (Softky, 1995; Lindsey et al., 1997; Prut et al., 1998; Mao et al., 2001). In the current study, the behaviour of a biologically realistic hierarchical neural network model of the primate ventral visual system is investigated. In particular, I explore how the network model develops stimulus representations in the form of fixed and regularly repeating temporal chains of spikes emitted by subpopulations of neurons even when the input images are represented by randomised Poisson spike trains during training. Also, a mechanism of synaptic delay selection with a biologically plausible learning mechanism, Spike-Timing-Dependent Plasticity (STDP), is explored. Perhaps most importantly, a potential solution to the classic feature *binding problem* in visual neuroscience, which concerns how the brain represents the relationships between visual features within a scene, is also investigated. I propose that this can be

provided in the form of what we have named a 'binding neuron', which represents the binding relationship between low-level and high-level visual features; such neurons were originally proposed by von der Malsburg (1999). In our simulations, the aim is to show that binding neurons encode binding relationships between visual features across the entire visual field and at every spatial scale. These binding neurons, which developed automatically through self-organization of the fixed and regularly repeating temporal chains of spikes during visual training, thus will provide a solution to feature binding. Lastly, I show how our proposed mechanism for solving the feature binding problem automatically leads to the bottom-up (feedforward) projection of visual information about lower level visual features, and indeed visual features at every level, through successive neuronal layers to the highest (output) layer of the network. We refer to this as the *holographic principle*. This may be important if subsequent brain areas that guide behaviour are only able to read out visual information from the highest stages of the visual system.

6.1.1 Temporal Coding and Polychronization

In the brain, neurons represent information and communicate with each other by pulses in their somatic membrane potential, called action potentials or 'spikes'. The activity of a somatic spike propagates down the axon of the neuron, causing neurotransmitters to be released from multiple presynaptic axon terminals into their corresponding synaptic clefts. Binding of the neurotransmitters to the receptors of the postsynaptic dendrites causes a change in the electrical activity of the postsynaptic neurons, constituting a communication of information from the presynaptic neuron to the postsynaptic neuron. This neuron also spikes if the excitation of this postsynaptic neuron from its afferent synapses increases the membrane potential above its firing threshold potential. Raising the membrane potential of the postsynaptic neuron above the firing threshold generally requires the activation of afferent synapses within a brief temporal window, as the membrane potential naturally decays quickly back to a resting potential without further afferent excitatory activation.

The relative timings of the spikes emitted by a pair of pre- and postsynaptic neurons has also been shown to affect learning through spike time dependent changes in synaptic efficacy (Bi and Poo, 1998; Markram et al., 1997), and hence how information and representations are stored and propagated in the network. If a presynaptic neuron fires in a short time period (up to tens of ms) prior to the postsynaptic neuron firing, the synaptic efficacy increases. An increase in synaptic efficacy is known as Long Term Potentiation (LTP). If the presynaptic neuron instead fires in a short period of time following the firing of a postsynaptic neuron, the efficacy of the synapse is reduced. This reduction in synaptic efficacy is known as Long Term Depression (LTD). These forms of LTP and LTD, which depend on the relative timings of the pre- and postsynaptic neurons, are known as Spike-Timing-Dependent Plasticity (STDP). Compared to firing rate based synaptic learning rules employed in rate-coded models, an STDP learning rule can result in very different self-organisation of the synaptic connectivity in the network when trained on visual scenes containing multiple objects (Evans and Stringer, 2012, 2013).

In a spiking neural network, individual neurons may operate as 'coincidence detectors' (Abeles, 1991; Jeanson, 2011). That is, a postsynaptic neuron will fire if spikes from a number of presynaptic neurons arrive within a relatively brief time window of the order of a few milliseconds. This will be the case if the neuronal and synaptic time constants of the postsynaptic neuron are relatively brief, allowing for a fast decay in the cell membrane potential between incoming presynaptic spikes. In this situation, the presynaptic spikes must arrive close together in time in order to combine together to drive up the postsynaptic cell membrane potential to reach its firing threshold. A simple example of a coincidence detecting neuron is shown in Figure 6.1a. In the figure, neurons 1 and 2 represent low-level features such as horizontal and vertical bars respectively, while neuron 3 is a coincidence detecting neuron that represents a high-level feature or object such as the alphabetic letter T. Neuron 3 only fires if the spikes emitted by

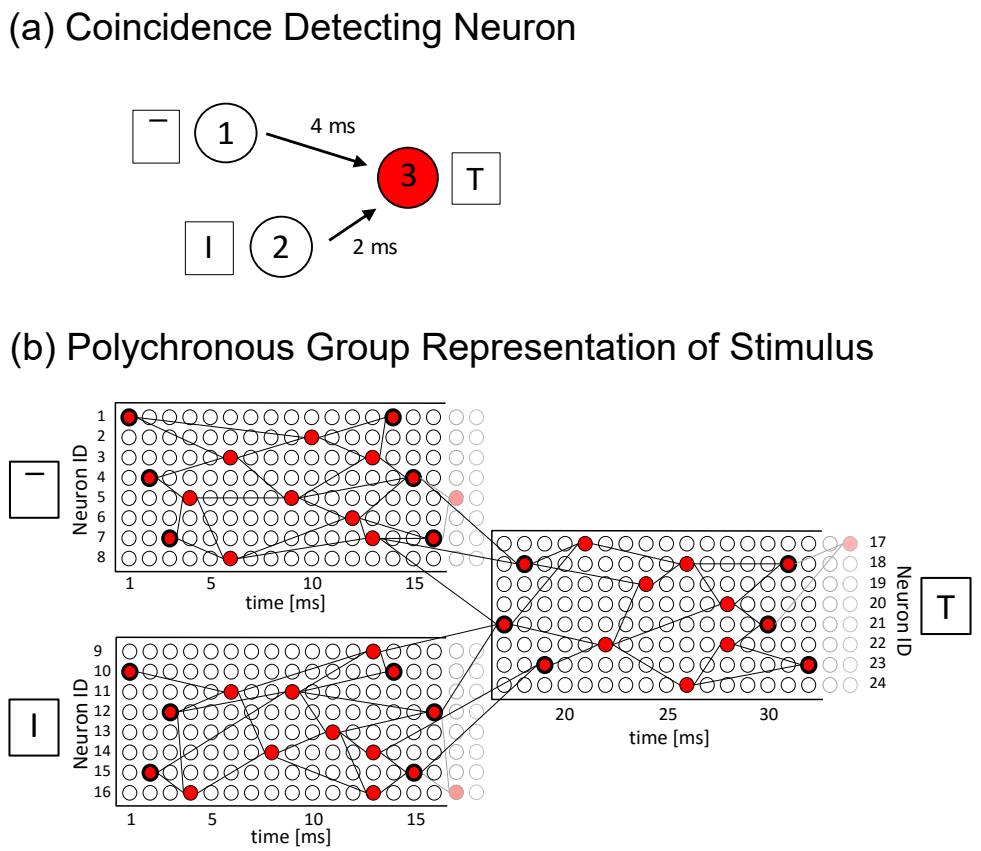

Figure 6.1: **(a) Example of coincidence detecting neuron arrangement.** Neurons 1 and 2 represent two different low-level features, a horizontal bar and a vertical bar respectively. Neuron 3 is a coincidence detecting neuron that represents a high-level feature, namely the alphabetic letter T. Neuron 3 only fires if the spikes emitted by neurons 1 and 2 arrive at neuron 3 close together in time. The action potentials of neurons 1 and 2 propagate activity to neuron 3 with delays 4ms and 2ms respectively. If the action potential of neuron 1 occurs approximately 2ms before the action potential of neuron 2, their propagating activity will arrive simultaneously at neuron 3 and cause it to spike. Neurons 1 and 2 represent the component vertical and horizontal bars comprising a letter T. In reality, the horizontal and vertical bars, as well as the letter T, would each be represented by a unique polychronous groups (PG) of neurons. **(b) Example of PG representation of Stimulus.** A horizontal bar is represented with a PG consisting of neurons 1-8, a vertical bar is represented with a PG consisting of neurons 9-16, and a character T is represented with a PG consisting of neurons 17-24.

neurons 1 and 2 arrive at neuron 3 close together in time. This means that the response of neuron 3 is sensitive not only to which presynaptic neurons are firing, but also to the precise timings of their spikes. As can be seen from this example, such a coincidence detecting neuron can provide a way of constructing higher level symbols through combination of elementary features, and do this in a way that utilises temporal coding that depends on the timings of spikes.

Simulation studies have shown that if the synaptic connections within a large population of neurons have axonal transmission delays that are drawn from a random distribution of variable magnitudes, from say a few milliseconds to several tens of milliseconds, then groups of coincidence detecting cells emerge through STDP (Izhikevich et al., 2004). Furthermore, the network develops repeating temporal chains of spiking activity distributed across subgroups of coincidence detecting neurons, i.e. neurons firing in a well defined temporal sequence. This is referred to as 'polychronization' (Izhikevich, 2006). Each subgroup of coincidence detecting neurons that comes together to form a regularly repeating temporal chain of activity is known as a 'polychronous group' (PG). It has been hypothesised that each PG could represent a particular sensory (e.g. visual) stimulus such as a letter T or perhaps episodic memory (Izhikevich, 2006). Figure 6.1b illustrates an example where a horizontal bar, a vertical bar, and a character T are represented by different PGs. In theory, polychronization in spiking networks can offer a dramatic increase in representational capacity compared to rate-coded models that do not

exploit the timings of spikes (Izhikevich, 2006).

Paugam-Moisy et al. (2008) have examined how PGs selectively respond to artificial input patterns after training with STDP and shed light on the potential of utilising PGs for real-life machine learning tasks such as handwritten digit recognition. However, the study carried out by these authors did not address three key issues as follows. Firstly, the study carried out by Paugam-Moisy et al. (2008) used carefully ordered spike trains to represent input images, which is not biologically plausible. What would happen if the input spike trains contained much more random variation as would be expected in the brain? Secondly, their model did not incorporate multiple synaptic connections with different randomised axonal transmission delays between each pair of pre- and postsynaptic neurons. This meant that the axonal transmission delay between any pair of neurons was fixed to a single value, and could not be effectively selected from a number of alternatives by STDP learning. Consequently, the set of possible PGs that a neuron could participate in was limited before learning. Thirdly, and perhaps most importantly, the study of Paugam-Moisy et al. (2008) did not investigate how feature binding representations, which explicitly encode the binding relations between low and high-level features, might develop through polychronization within a hierarchical model of visual processing. In the simulations presented below, each of these three issues is investigated in a hierarchical spiking neural network model of the primate ventral visual pathway, which is tasked with learning representations of the shapes of 2-dimensional visual objects.

6.1.2 The Binding Problem and a Limitation of Rate Coding

Descriptions of the binding problem vary but generally address the same question: how does the visual system represent which elementary features are bound together to form an object? For example, if the two letters T and L are seen together, how does the visual system represents which horizontal and vertical bars are part of which letter? In traditional hierarchical rate-coded visual processing models (e.g., Fukushima (1980), Wallis and Rolls (1997), and Riesenhuber and Poggio (1999)), simple features (such as horizontal and vertical bars) are represented in the lower visual layers while more complex features (such as letters) are represented in the higher visual layers. However, without a solution to the feature binding problem, there is no way of reading off which bars are part of which letters and hence where the object's constituent components are in space.

The underlying weakness of rate coding is well illustrated in the classical example of Rosenblatt (1961), which was further explained by von der Malsburg (1999). As Figure 6.2 illustrates, the example supposes there is a neural network with four output neurons. Output neurons A and B represent the triangle and square respectively, invariant to retinal position (top or bottom). Output neurons C and D are instead location specific, responding to both objects in either the top or bottom location respectively. When a single object is presented to the network, the responses of the four neurons provide sufficient information to decode both the shape and position of the object. On the other hand, when both objects are presented together, each at a different location, all of the output neurons become highly active; it is no longer clear whether the triangle or the square is in the top retinal location. Thus, the co-activation results in a merging of representations and a loss of information which could have been used to divide the scene into its components. This breakdown is referred to as the "superposition catastrophe" (von der Malsburg, 1999). Similar problems were reported in a study modelling the development of border ownership representations in the early visual cortex, driven by top down modulation from higher layers in Chapter 5 (Eguchi and Stringer, 2016). This rate coded model produced neuron responses characterizing border ownership cells in V1. However, these representations catastrophically failed upon the presentation of multiple visual stimuli due to the inability of the rate coded model to provide spatially selective top down modulation.

In short, the crucial problem with rate coding is the lack of means to represent information regarding which specific low-level/elementary features have been combined to construct higher

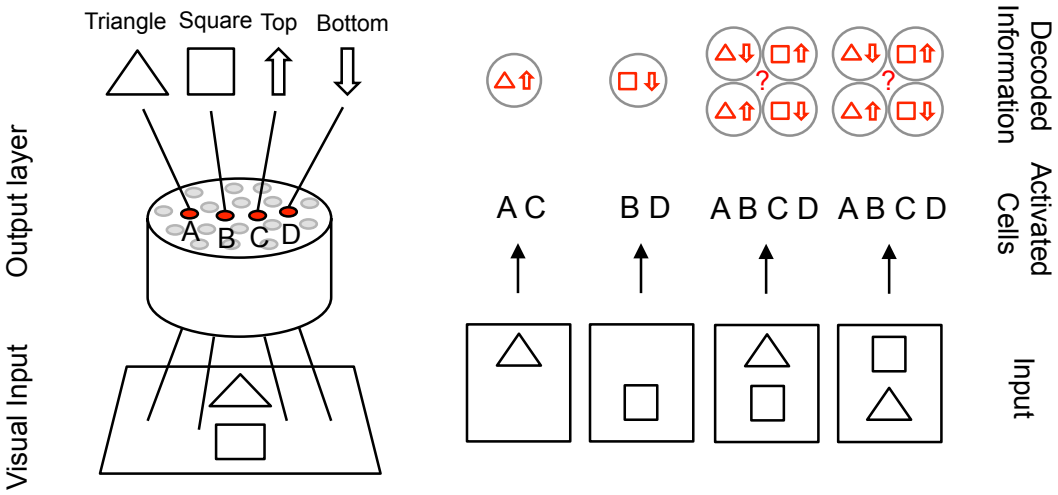

Figure 6.2: **Rosenblatt (1961)'s example of a binding problem in a rate-coded network.** Left: the input from a visual scene is presented to a neuronal population including a set of four output neurons A, B, C and D. The firing rate responses from neurons A-D respectively represent the presence of the following: a triangle, a square, an object in the "top" position, and an object in the "bottom" position. Right: the responses of output neurons A-D when four different scenes are presented to the network. It can be seen that when only a single object is presented, the network can represent both the object and its position. However, when both objects are presented together, although the network is able to represent that both objects are present, it fails to represent the actual position of each object.

level features or objects. Moreover, binding of visual features must operate across the entire visual field and at all spatial scales within a visual scene. How features are bound together underpins how we segment a visual scene into objects and parts of objects, and thus how we make sense of the visual world.

Thus, solving the binding problem is essential to understanding the ability of the primate visual brain to make sense of complex visual scenes, and to developing a next generation of far more powerful computer vision systems with the ability to understand what they are looking at in the same way as the brain. Our simulation results suggest that binding is a much richer phenomenon than traditionally described by visual psychologists. Indeed, the binding mechanism proposed here is potentially so rich that it would be difficult to describe the process at a high psychological level; it requires a description at the neuronal level as presented in this chapter.

6.2 Hypotheses

I investigate the behaviour of a biologically realistic hierarchical neural network model of the primate ventral visual system that incorporates the following key aspects of cortical dynamics and architecture:

- (i) The model implements spiking neural dynamics in which the timings of action potentials or 'spikes' are simulated explicitly.
- (ii) Spike-Timing-Dependent Plasticity (STDP) is used to modify the synaptic connections during visually-guided learning. If a spike from a presynaptic neuron arrives at a postsynaptic neuron just before the postsynaptic neuron emits a spike, then the synapse is strengthened (LTP). Otherwise, if the spike from the presynaptic neuron arrives at the postsynaptic neuron just after the postsynaptic neuron emits a spike, then the synapse is weakened (LTD).
- (iii) The network architecture incorporates bottom-up, top-down, and lateral synaptic connections reflecting the known architecture of the visual cortex.

- (iv) The synaptic connectivity between neurons incorporates distributions of axonal conduction delays of varying durations, from a few milliseconds to tens of milliseconds.
- (v) In some simulations, network performance is enhanced by incorporating multiple synaptic connections between each pair of pre- and postsynaptic neurons, where these connections have different axonal transmission delays. This permits STDP to strengthen just one (or a subset) of these connections in order to effectively select the functional transmission delay between the two neurons (Fares and Stepanyants, 2009; Deger et al., 2012).

Using this underlying model architecture, the current study investigates the following hypotheses: emergence of polychronization, emergence of binding neurons, and 'holographic principle' in the brain.

6.2.0.1 Emergence of Polychronization

During the initial period of visually-guided learning, the network is trained on a set of visual stimuli that are encoded in the input layer by spiking neurons with *randomised* Poisson distributions of spikes. That is, the spike patterns representing the stimuli in the input layer have no special temporal structure, except that the average firing rates of the input neurons are set in accordance with the outputs of Gabor filters that simulate the responses of simple cells in visual area V1. Nevertheless, it was hypothesised that the initial period of visually-guided learning with STDP would lead to the development of large numbers of regularly repeating PGs in the higher layers of the network, where individual PGs respond selectively to particular stimuli. Moreover, it was hypothesised that the emergence of large numbers of stimulus-selective PGs would increase the representational capacity of the network beyond that offered by a localist rate-coded representation in that, after training, the number of stimulus-specific PGs would be significantly greater than the number of single cells that responded selectively to a particular stimulus. The representational capacity is thus increased if the network encodes visual stimuli using temporal spike trains distributed over PGs of neurons rather than relying on the average firing rate responses of individual neurons.

6.2.0.2 Emergence of Binding Neurons

It was hypothesised that the emergence of PGs in the higher layers of the network during visually-guided training with STDP could provide a solution to the classic feature *binding problem* in visual neuroscience. That is, how may the network learn to represent the hierarchical binding relations between low-level features such as horizontal or vertical bars and high-level features or objects such as the alphabetic letters T and L? Specifically, I hypothesised that some cells within PGs, which we will call 'binding neurons', will become tuned through STDP learning to respond if a neuron or subset of neurons representing a specific low-level feature are participating in driving neurons representing a particular high-level feature or object, which may be represented in a higher layer. In this case, the binding neuron carries measurable information that the low-level feature (such as a vertical bar at a particular retinal location) is part of the higher level feature or object (such as the letter T). Such binding neurons were originally proposed by von der Malsburg (1999), but without an explanation of how they might emerge naturally during visual development. I now propose, and demonstrate in the simulations presented below, that such binding neurons may develop automatically within the PGs that emerge during visually-guided learning with STDP.

Here, a simple explanation for how such binding neurons may develop is presented. An actual example is given in Figure 6.3a. Consider a linked set of three neurons at different stages of the ventral visual pathway: (i) neuron 1 (in a lower visual layer) represents a low-level visual feature, (ii) neuron 2 (in a higher visual layer) represents a high-level visual feature, and (iii)

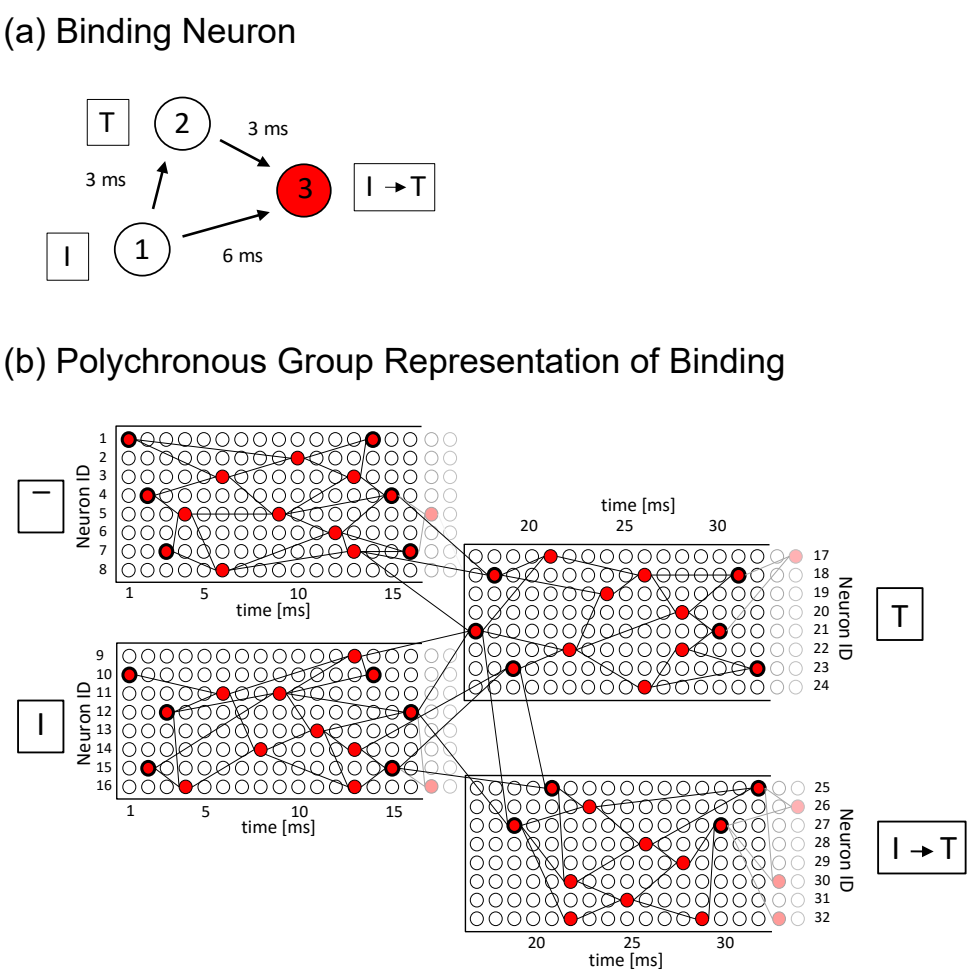

Figure 6.3: **(a) Example of a hypothetical binding neuron.** Consider a linked set of three neurons at different stages of the ventral visual pathway: neuron 1 (in a lower visual layer) represents a low-level visual feature such as a vertical bar, neuron 2 (in a higher visual layer) represents a higher level visual feature such as the letter T, and neuron 3 is a binding neuron within a local layer, say the same layer as either neuron 1 or 2. Importantly, there are non-zero axonal transmission delays in the connections between these three neurons. In this example the delays are as follows: the delay from neuron 1 to neuron 2 is 3ms, the delay from neuron 2 to neuron 3 is 3ms, and the delay from neuron 1 to neuron 3 is 6ms. It is assumed that neurons in the network behave as 'coincidence detectors' in that they require a volley of spikes from presynaptic cells to arrive simultaneously at the postsynaptic cell in order for the postsynaptic cell to fire itself. In the brain, this would be due to neurons having fast neuronal and synaptic time constants allowing neuronal activation to decay quickly between incoming spikes. In the example shown, it is assumed that neuron 3 needs incoming spikes from neurons 1 and 2 to arrive close together in time, e.g. within one or two milliseconds, in order for neuron 3 to fire itself. Then, with the particular set of axonal delays given in this example, neuron 3 will fire, if and only if, neuron 1 is participating in driving neuron 2. For only in this case will the spikes from neurons 1 and 2 arrive at neuron 3 at the same time (or at least close together in time) and fire neuron 3. If neuron 3 fires, this will encode the fact that the low-level feature (vertical bar) represented by neuron 1 is part of the higher level feature (letter T) represented by neuron 2. Thus, to reiterate, if the delay in propagation of a spike from neuron 1 to neuron 3 (6ms) is (at least approximately) equal to the sum of the delays in propagation from neurons 1 to 2 and 2 to 3 (3ms and 3ms respectively), neuron 3 will represent the binding relationship between the low and high-level features encoded by neurons 1 and 2. That is, the firing of neuron 3 will indicate that the vertical bar represented by neuron 1 is part of the letter T represented by neuron 2. The three neurons, 1, 2 and 3, form a PG, in which the binding neuron 3 is embedded. It is hypothesised that large numbers of these PGs with embedded binding neurons will emerge throughout the network during visually-guided learning with synaptic modification driven by STDP. **(b) polychronous group (PG) Representation of Binding.** Here I illustrate how a low-level feature such as a vertical bar may in fact be represented by its own temporal pattern of spikes distributed across a PG of neurons (shown bottom left), the high-level feature or object such as a letter T may also be represented by its own PG (shown top right), and these two PGs may drive a third PG representing the binding relationship between the vertical bar and the letter T (shown bottom right). This more complex scenario, in which the visual features and the binding relations between these features are represented by patterns of spiking activity across their own PGs, is likely to be what actually happens in the brain. The simple three neuron circuit shown in part **(a)** would then be a small part of the three corresponding PGs (representing a vertical bar, letter T, and binding relation between these two features) shown in part **(b)**.

neuron 3 is a hidden neuron within a local layer, say the same layer as either neuron 1 or 2, which may learn to become a binding neuron. Assume that there are the following three synaptic connections between these three neurons: (i) a connection from neuron 1 to neuron 2, (ii) a connection from neuron 1 to neuron 3. (This could be either a lateral or bottom-up connection depending on which layer neuron 3 is in), and (iii) a connection from neuron 2 to neuron 3. (This could be either a lateral or top-down connection depending on which layer neuron 3 is in.)

Let us denote the axonal delay from neuron j to neuron i as $\Delta_{(i,j)}$. Then neuron 1 is participating in driving neuron 2 if and only if a spike emitted by neuron 2 occurs approximately $\Delta_{(2,1)}$ after a spike emitted by neuron 1.

If there is a set of three axonal delays such that

$$\Delta_{(3,1)} = \Delta_{(2,1)} + \Delta_{(3,2)} \tag{6.1}$$

then the spikes from neurons 1 and 2 will converge on neuron 3 (near) simultaneously if and only if neuron 1 is participating in driving neuron 2.

It is assumed that the hidden neuron 3 operates as a 'coincidence detector', and fires only when the volley of spikes from neurons 1 and 2 arrive (near) simultaneously. In this case, neuron 3 will behave as a binding neuron. That is, neuron 3 will fire if and only if neuron 1 is participating in driving neuron 2. In this case, STDP will further strengthen the connections from neurons 1 and 2 onto the binding neuron 3.

It is important that an ideal binding neuron responds if and only if the neurons representing the low-level feature are actually participating in driving the neurons representing the high-level feature. The binding neuron should not respond if the neurons representing the low-level feature and the neurons representing the high-level feature just happen to be co-active, where the former are not actually driving the latter. Such unrelated co-activation of low and high-level features might occur, for example, because of the presence of multiple similar objects within a complex natural scene as explained earlier with Rosenblatt (1961)'s example (Figure 6.2). Suppose a T and L are presented together, then the neurons representing the horizontal bar of the T are co-active with the neurons representing the letter L, but the former are not driving the latter. Thus, the corresponding binding neuron, which would represent that the given horizontal bar was part of the L, should not fire. This kind of temporally specific response is characteristic of a PG, which the three neurons 1, 2 and 3 described above comprise.

I hypothesise that with the inclusion of bottom-up, top-down and lateral connections, there are a variety of possible local network architectures that could self-organise through competitive learning to implement this, with the binding neurons being in any of the nearby lower or higher layers. Wherever the binding neuron is, its activation would still represent that a particular low-level feature is driving the representation of a specific high-level feature or object, and is therefore part of the object. A population of such binding neurons would specify which low-level features within a scene were part of which high-level features or objects, and this information could be read out directly by higher level neurons in the network. One of the examples of such 'binding neuron' is the border ownership neuron described in Chapter 5 and in the next chapter in Chapter 7.

This process could operate across the entire visual field and at every spatial scale within the visual field. Indeed, binding neurons would be expected to emerge throughout successive levels of the feature hierarchy within the network. A rich tapestry of binding neurons through the layers could help to provide a hierarchical structural description of a scene. This proposal may explain why the visual system needs extensive top-down connections between layers and lateral connections within layers, in addition to bottom-up connections.

However, in the brain it is in fact likely that a low-level feature such as a vertical bar and a high-level feature such as the letter T, as well as the binding relationship between these features, would each be represented by their own temporal pattern of spikes distributed across PGs of

neurons. This is illustrated in Figure 6.3b. The binding relations are then represented by PGs (rather than individual neurons), which are replayed if and only if the low-level feature is part of the high-level feature. In this scenario, the simple three neuron circuit shown in part (a) would be a small part of the three corresponding PGs (representing a vertical bar, letter T, and binding relation between these two features) shown in part (b). In the simulations reported below we only focus on identifying individual binding neurons that are part of three neuron circuits of the general form shown in Figure 6.3a.

In this investigation, I specifically look at the emergence of such binding neurons among the learned neuronal representations of three simple visual shapes shown in Figure 6.6, which are presented to the network during visually-guided training. I expected to find evidence for the kind of 3 neuron binding relationships described above and illustrated in Figure 6.3a. These three neuron PGs provide the simplest examples of how the network may learn to represent binding relationships where specific low-level features are part of particular high-level features or objects.

6.2.0.3 Feedforward projection of information about low-level visual features to higher neuronal layers

The above discussion of binding neurons leads directly to a new hypothesis concerning how information about visual features may be projected in a bottom-up (feedforward) manner through successive layers of the network. This might be a useful operation if the behavioural systems of the brain are limited to reading out visual information from the highest layers of the visual system. For example, it is generally conceived that simple visual features such as oriented edges and bars are represented in early cortical visual areas such as V1 and V2, while whole objects and faces are represented in higher visual areas. However, when we look at a visual scene we are perceptually aware of visual features of varying levels of complexity and scale. Does this imply that information about low-level visual features is being projected directly upwards through the visual system in some way that preserves the identity of these features, and at the same time also represents the image context of these features (i.e. binding relationships with higher level features)?

Figure 6.4a shows one simple way in which our network architecture might achieve this. The illustration is very similar to that shown in Figure 6.3a, except that the binding neuron 3 is now in the upper layer, i.e. the same layer as neuron 2 which represents the high-level feature T. Neuron 3 represents the fact that there is a vertical bar in some local region of the retina, which is part of the letter T. In this way, information about the low-level feature (vertical bar at a particular retinal position) along with its image context (the vertical bar is part of the letter T) has been projected up to the same layer as the representation of the high-level feature (the T). This is essentially the same binding mechanism discussed above, but where the binding neuron is situated in the same higher layer as the neuron representing the high-level feature. This mechanism for the bottom-up projection of information about low-level features to higher layers, where this information is modulated by local image context (i.e. the low-level feature is part of a particular high-level feature), may again operate up through successive neuronal layers, hence across the entire visual field and at every spatial scale.

It is possible that the mechanism shown in Figure 6.4a could be repeated iteratively up through the layers. For example, Figure 6.4b shows an example in which information about the vertical bar is first projected up from the first layer to the second layer, where it is represented by binding neuron 3. Neuron 3 represents the fact that there is a vertical bar in a local region of the retina, which is part of the letter T. Then, a similar binding mechanism combines the output from binding neuron 3 with the output of neuron 5 representing a cat, where these combined outputs drive binding neuron 6. Binding neuron 6 then represents the fact that there is a vertical bar in a local region of the retina, which is part of the letter T, which in turn is part of the word CAT. In this case, the information about the lowest level feature is preserved

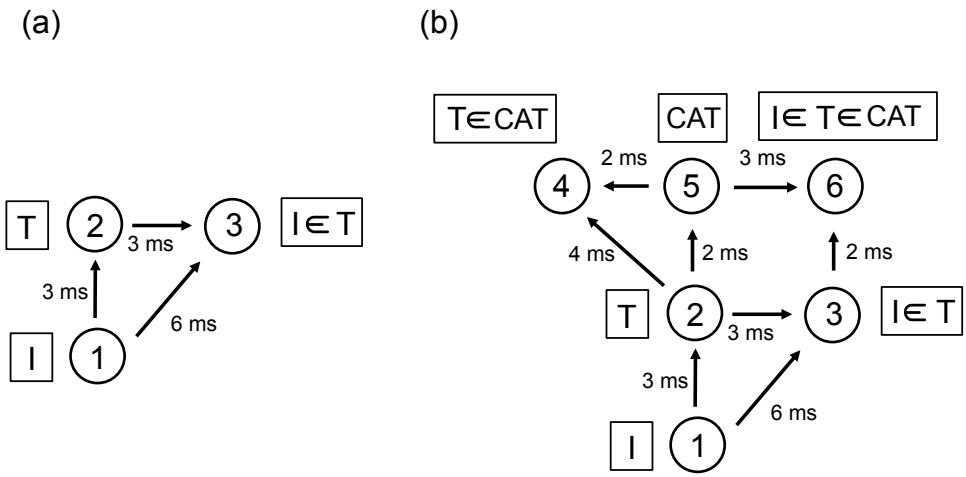

Figure 6.4: Illustrations of how the proposed binding mechanism may project information about low-level visual features such as a vertical bar up through successive layers of the network. The illustration shown in **(a)** is very similar to that shown in Figure 6.3a, except that the binding neuron 3 is now located in the upper layer, i.e. the same layer as neuron 2 that represents the high-level feature T. Neuron 3 represents the fact that there is a vertical bar in some local region of the retina, which is part of the letter T. In this way, information about the low-level feature (vertical bar at a particular retinal position) along with its image context (the vertical bar is part of the letter T) has been projected up to the same layer as the representation of the high-level feature (the T). **(b)** shows how the mechanism illustrated in **(a)** could be repeated iteratively up through the layers. Now a similar binding mechanism combines the output from binding neuron 3 with the output of neuron 5 representing a cat, where these combined outputs drive binding neuron 6. Binding neuron 6 then represents the fact that there is a vertical bar in a local region of the retina, which is part of the letter T, which in turn is part of the word CAT. In this case, the information about the lowest level feature (a vertical bar) is preserved in the highest layer of the network.

in the highest layer of the network. Indeed it is theoretically possible that a very large amount of information could be projected upwards in this manner and preserved in the highest layers for readout by subsequent behavioural systems. We refer to this as a *holographic principle* for spiking network models of biological vision, because information about visual features at every level of complexity and scale may be preserved in the highest layers.

It is important to note that the binding neurons 3 and 6 in the highest layers of the two network architectures shown in Figures 6.4a and 6.4b represent the presence of a vertical bar in some local region of the retina which is explicitly part of a higher level feature (e.g. the letter T) or hierarchy of features (e.g. the letter T, which is part of the word CAT). Thus, these binding neurons do not simply respond to the presence of a vertical bar at some retinal location regardless of local image context (i.e. the higher level features / objects which the vertical bar is part of). So the high-level feature / object still needs to be presented to the network in order to elicit a response from these kinds of binding neuron in the upper layers. Thus, the holographic principle described here is consistent with neurophysiological observations that neurons in the later stages of the ventral visual pathway tend to respond to more complex visual forms than the simple oriented bars represented in early cortical stages such as V1 and V2.

6.2.0.4 Effect of Varying Key Model Parameters

The study also investigates the effects of the following architectural, neuronal, and synaptic parameters on the number of PGs and binding neurons that develop in the network during visually-guided training:

- **Synaptic connectivity.** The investigation will explore the performance of the network with the following synaptic connectivities: (i) purely bottom-up, (ii) bottom-up and top-down, (iii) bottom-up and lateral, and (iv) bottom-up, top-down, and lateral. I test which of these architectures best promotes the emergence of polychronization including the representation of visual stimuli by stimulus-specific PGs.

- **STDP time constant.** I hypothesise that a longer STDP time constant will lead to less temporal sensitivity to spike times in the network, which will lead to the model operating in a more rate-coded manner. This in turn may reduce the emergence of polychronization including the number of stimulus-specific PGs that develop.
- **Multiple synaptic connections between each pair of pre- and postsynaptic neurons.** Within a network with multiple synaptic contacts (each with a different axonal delay) between each pair of pre- and postsynaptic neurons, I hypothesise that STDP will effectively select which delays to strengthen. If STDP is able to selectively strengthen just one (or a subset) of the connections, this should help to promote the emergence of polychronisation. For example, if each pair of pre- and postsynaptic neurons has two synaptic contacts with quite different axonal delays, then it is expected that STDP will increase one connection but weaken the other. I hypothesise that this will in turn increase the number of PGs in the network with maximum information for a particular stimulus.

6.3 Materials & Methods

6.3.1 Model

6.3.1.1 Network Architecture

The neural network model investigated is shown in Figure 6.5 and simulates successive neuronal stages of processing along the primate ventral visual pathway. Specifically, the model is comprised of four layers of excitatory pyramidal neurons, which may be thought of as representing cortical visual areas V2, V4, posterior inferior temporal cortex (TEO) and anterior inferior temporal cortex (TE). There are modifiable bottom-up (feedforward) and top-down (feedback) synaptic connections between excitatory pyramidal neurons in successive layers, as well as modifiable lateral synapses between excitatory pyramidal neurons within each layer. Some simulations reported below explore the importance of the top-down and lateral connections for polychronisation and feature binding by comparing model performance with and without them. Within each layer, there are also inhibitory interneurons with non-plastic lateral synaptic connections to and from the excitatory neurons to produce competition between the excitatory neurons. For all presented simulations, I used $64 \times 64 = 4096$ excitatory neurons and $32 \times 32 = 1024$ inhibitory neurons in each layer, with a fixed number of sparsely distributed topologically organised connections. Table 6.1a shows the different numbers of afferent connections onto each postsynaptic neuron, as well as the fan-in radius of these connections, for the different types of excitatory-excitatory, excitatory-inhibitory and inhibitory-excitatory connections between and within the four neuronal layers.

6.3.1.2 Differential Equations

As originally described in Evans and Stringer (2012), each neuron is based upon the standard conductance-based leaky integrate and fire (LIF) model, whilst the equations for STDP at the Excitatory-Excitatory ($E \rightarrow E$) synapses are adapted from Perrinet et al. (2001). Neuron and synapse constants were chosen to be as biologically realistic as possible based upon the available neurophysiological literature (see Table 6.1 for a full list).

Cell equations The neuron's membrane potential is updated according to Equation (6.2).

$$\tau_m^\gamma \frac{dV_i(t)}{dt} = V_0^\gamma - V_i(t) + R^\gamma I_i(t) \tag{6.2}$$

The cell membrane potential for a given neuron $V_i(t)$ (indexed by i) is driven up by current from excitatory conductance-based synapses, and down towards the inhibitory reversal potential

Table 6.1: Model parameters. Most integrate and fire parameters were taken from Troyer et al. (1998) (derived originally from McCormick et al. (1985) as indicated by §. Plasticity parameters (denoted by †) are taken from Perrinet et al. (2001). Parameters marked with * were tuned for the reported simulations.

(a) Network parameters					
Layer	**1**	**2**	**3**	**4**	
Number of excit. neurons within each layer	64×64	64×64	64×64	64×64	
Number of inhib. neurons within each layer	32×32	32×32	32×32	32×32	
Number of feedforward (FF) afferent excit. connections per excit. neuron (EfE)	30	100	100	100	
Fan-in radius for FF afferent excit. connections to each excit. neuron (EfE)	1.0	8.0	12.0	16.0	
Number of feedback (FB) afferent excit. connections per excit. neuron (EbE)	{0,10}	{0,10}	{0,10}	-	
Fan-in radius for FB afferent excit. connections to each excit. neuron (EbE)	8.0	8.0	8.0	-	
Number of lateral (LAT) afferent excit. connections per excit. neuron (ElE)	{0,10}	{0,10}	{0,10}	{0,10}	
Fan-in radius for LAT afferent excit. connections to each excit. neuron (ElE)	4.0	4.0	4.0	4.0	
Number of LAT afferent excit. connections per inhib. neuron (ElI)	30	30	30	30	
Fan-in radius for LAT afferent excit. connections to each inhib. neuron (ElI)	1.0	1.0	1.0	1.0	
Number of LAT afferent inhib. connections per excit. neuron (IlE)	30	30	30	30	
Fan-in radius for LAT afferent inhib. connections to each excit. neuron (IlE)	8.0	8.0	8.0	8.0	
(b) Parameters for Gabor Filtering of visual images					
Phase shift (ψ)	$0, \pi$				
Wavelength (λ)	2				
Orientation (θ)	$0, \pi/4, \pi/2, 3\pi/4$				
Spatial bandwidth (b)	1.5 octaves				
Aspect ratio (γ)	0.5				
(c) Cellular Parameters					
Excit. cell somatic capacitance (C_m^E) and Inhib. cell somatic capacitance (C_m^I)	500 pF, 214 pF				§
Excit. cell somatic leakage conductance (g_0^E) and Inhib. cell somatic leakage conductance (g_0^I)	25 nS, 18 nS				§
Excit. cell membrane time constant (τ_m^E) and Inhib. cell membrane time constant (τ_m^I)	20 ms, 12 ms				§
Excit. cell resting potential (V_0^E) and Inhib. cell resting potential (V_0^I)	-74 mV, -82 mV				§
Excit. firing threshold potential (Θ^E) and Inhib. firing threshold potential (Θ^I)	-53 mv, -53 mV				§
Excit. after-spike hyperpolarization potential (V_H^E) and Inhib. after-spike hyperpolarization potential (V_H^I)	-57 mV, -58 mV				§
Absolute refractory period (τ_R)	2 ms				§
(d) Synaptic Parameters					
Synaptic neurotransmitter concentration (α_C) and Proportion of unblocked NMDA receptors (α_D)	0.5				†
Presynaptic STDP time constant (τ_C) and Postsynaptic STDP time constant (τ_D)	{5, 25, 125} ms				†
Synaptic learning rate (ρ)	0.1				†
Range of Synaptic Conductance Delay	[0.1, 10.0] ms				†
Synaptic conductance scaling factor for FF excitatory connections from Gabor filters to Layer 1 excit. cells ($\lambda^{GfE} \cdot \Delta g^{GfE}$)	[0, 0.4] nS				*
Synaptic conductance scaling factor for FF excit. connections to excit. cells in layers 2, 3 or 4 ($\lambda^{EfE} \cdot \Delta g^{EfE}$)	[0, 1.6] nS				*
Synaptic conductance scaling factor for FB excit. connections to excit. cells in layers 1, 2 or 3 ($\lambda^{EbE} \cdot \Delta g^{EbE}$)	[0, 1.6] nS				*
Synaptic conductance scaling factor for LAT excit. connections to excit. cells in layers 1, 2, 3 or 4 ($\lambda^{ElE} \cdot \Delta g^{ElE}$)	[0, 1.6] nS				*
Synaptic conductance scaling factor for LAT connections from excit. cells to inhib. cells in layers 1, 2, 3 or 4 ($\lambda^{ElI} \cdot \Delta g^{ElI}$)	40 nS				*
Synaptic conductance scaling factor for LAT connections from inhib. cells to excit. cells in layers 1, 2, 3 or 4 ($\lambda^{IlE} \cdot \Delta g^{IlE}$)	80 nS				*
Excitatory reversal potential ($\hat{V}^E$)	0 mV				§
Inhibitory reversal potential ($\hat{V}^I$)	-70 mV				§
Synaptic time constant for all FF, FB, and LAT connections from Gabor filters and excit. cells to excit. cells ($\tau_{GfE}, \tau_{EfE}, \tau_{EbE}, \tau_{ElE}$)	150 ms				*
Synaptic time constant for LAT connections from excit. cells to inhib. cells (τ_{ElI})	2 ms				§
Synaptic time constant for LAT connections from inhib. cells to excit. cells (τ_{IlE})	5 ms				§
(e) Parameters for numerical simulation by Forward Euler timestepping scheme					
Numerical step size (Δt)	0.02 ms				

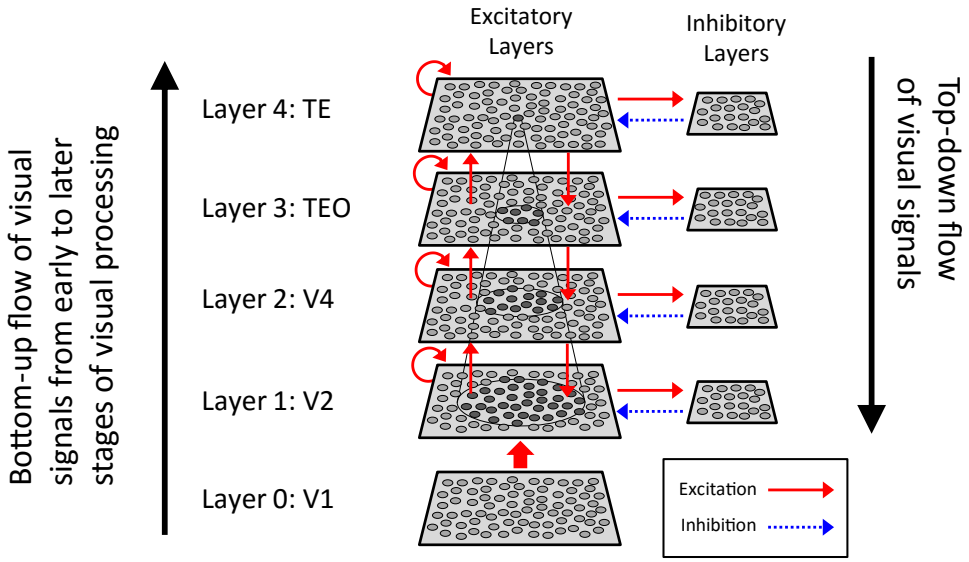

Figure 6.5: Schematic of the four layer neural network architecture investigated. The model represents successive neuronal stages of processing along the primate ventral visual pathway. The model is comprised of five layers of excitatory pyramidal neurons, which may be thought of as representing cortical visual areas V1, V2, V4, posterior inferior temporal cortex (TEO) and anterior inferior temporal cortex (TE). The layer 0 reflects the output of the Gabor filters of the visual input presented to the network, with an imposed Poisson spike rate of neurons in the layer. These neurons establish only feed-forward connection to the layer 1. Each of the following layers of the model (layer 1 - 4) consists of $64 \times 64 = 4096$ excitatory neurons and $32 \times 32 = 1024$ inhibitory neurons. Excitatory modifiable connections (red) include bottom-up (feedforward) and top-down (feedback) connections between excitatory pyramidal neurons in successive layers, and lateral connections between excitatory pyramidal neurons within the same layer (shown by the curved red arrows). Each layer of excitatory pyramidal neurons is connected to a population of inhibitory neurons which implement competition between the excitatory neurons in that layer.

by current from inhibitory conductance-based synapses. Neurons decay back to their resting state over a time course determined by the properties of its membrane. Here τ_m represents the membrane time constant, defined as $\tau_m = C_m/g_0$, where C_m is the membrane capacitance, g_0 is the membrane leakage conductance and R is the membrane resistance ($R = 1/g_0$). V_0 denotes the resting potential of the cell. Class-specific values (excitatory and inhibitory) are indexed by γ for the above neuron parameters. $I_i(t)$ represents the total current input from the afferent synapses (described in Equation (6.3)).

The total synaptic current injected into a neuron is given by the sum of the conductances of all afferent synapses (excitatory and inhibitory), multiplied by the difference between the synapse class reversal potential ($\hat{V}^\gamma$) and neuron membrane potential ($V_i(t)$). Excitatory and inhibitory synapses have positive and negative conductances respectively. The conductance of a given synapse is given by g_{ij} where j and i are the indices of the pre- and postsynaptic neurons respectively.

$$I_i(t) = \sum_\gamma \sum_j g_{ij}(t)(\hat{V}^\gamma - V_i(t)) \tag{6.3}$$

Synaptic conductance equations The synaptic conductance of a particular synapse, $g_{ij}(t)$, is governed by a decay term τ_g and a Dirac delta function (Equation (6.5)) when spikes arrive from the presynaptic neuron j as follows:

$$\frac{dg_{ij}(t)}{dt} = -\frac{g_{ij}(t)}{\tau_g} + \lambda \Delta g_{ij}(t) \sum_l \delta(t - \Delta t_{ij} - t_j^l) \tag{6.4}$$

The conduction delay for a particular synapse, which is ranged from 0.1 to 10.0 ms, is denoted by Δt_{ij} and each presynaptic neuron spike is indexed by l. A biological scaling constant λ has

been introduced to scale the synaptic efficacy Δg_{ij} which lies between unity and zero. The Dirac delta function is defined as follows:

$$\delta(x) = \begin{cases} \infty & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{where, } \int_{-\infty}^{\infty} \delta(x)dx = 1 \tag{6.5}$$

Synaptic learning equations The following differential equations describe the Spike-Timing-Dependent Plasticity (STDP) occurring at each modifiable *Excitatory - Excitatory* ($E \rightarrow E$) synapse. That is, these kinds of modifiable synapses occur at all of the bottom-up, top-down and lateral connections from excitatory cells to excitatory cells throughout layers 1 to 4.

Here i labels the postsynaptic neuron. The recent presynaptic activity, $C_{ij}(t)$, is modelled by Equation (6.6) which may be interpreted as the concentration of neurotransmitter (glutamate) released into the synaptic cleft (Perrinet et al., 2001) and is bounded by [0, 1] for $0 \leq \alpha_C \leq 1$.

$$\frac{dC_{ij}(t)}{dt} = -\frac{C_{ij}(t)}{\tau_C} + \alpha_C(1 - C_{ij}(t)) \sum_l \delta(t - \Delta t_{ij} - t_j^l) \tag{6.6}$$

$C_{ij}(t)$ is governed by a decay term τ_C and is driven up by presynaptic spikes according to the model parameter α_C. The inclusion of the axonal transmission delay Δt_{ij} from presynaptic neuron j to postsynaptic neuron i in equation (6.6) means that the variable $C_{ij}(t)$ is driven up at the time the spike from presynaptic neuron j arrives at the postsynaptic neuron i, rather than the time of emission of the spike from the presynaptic cell.

The recent postsynaptic activity, $D_i(t)$, is modelled by Equation (6.7) and may be interpreted as the proportion of NMDA receptors unblocked by recent depolarization from back-propagated action potentials (Perrinet et al., 2001).

$$\frac{dD_i(t)}{dt} = -\frac{D_i(t)}{\tau_D} + \alpha_D(1 - D_i(t)) \sum_k \delta(t - t_i^k) \tag{6.7}$$

$D_i(t)$ is governed by decay term τ_D and is driven up by postsynaptic spikes according to the model parameter α_D. Postsynaptic neuron spikes are indexed by k. Unlike with the conduction of action potentials towards the synapse, there is no conduction delay associated with D_i, since the cell body is assumed to be arbitrarily close to the receiving synapses and the effects of a postsynaptic spike are assumed to have an equal impact on the neuron's own afferent synapses.

The strength of the synaptic weight, $\Delta g_{ij}(t)$, is modified according to Equation (6.8), which is governed by the time course variable τ_{Δ_g}.

$$\begin{aligned} \tau_{\Delta_g} \tfrac{d\Delta g_{ij}(t)}{dt} = \quad & \rho[(1 - \Delta g_{ij}(t))C_{ij}(t) \textstyle\sum_k \delta(t - t_i^k) \\ & -\Delta g_{ij}(t) D_i(t) \textstyle\sum_l \delta(t - \Delta t_{ij} - t_j^l)] \end{aligned} \tag{6.8}$$

Note that the postsynaptic spike train (indexed by k) is now associated with the presynaptic state variable (C) and vice versa. If C is high (due to recent presynaptic spikes having arrived at the postsynaptic neuron) at the time of a postsynaptic spike, then the synaptic weight is increased (LTP) whereas if D is high (from recent postsynaptic spikes) at the time of a presynaptic spike arriving at the postsynaptic neuron then the weight is decreased (LTD). As noted above, the inclusion of the axonal transmission delay Δt_{ij} in equation (6.6) means that the variable $C_{ij}(t)$ is driven up at the time the spike from presynaptic neuron j actually arrives at the postsynaptic neuron i. Consequently, this form of STDP learning depends directly on the times that spikes from a presynaptic neuron arrive at a postsynaptic neuron rather than the times that the spikes were originally emitted by the presynaptic neuron.

The weight updates are also multiplicative, meaning that the amount of potentiation decreases as the synapse strengthens, as has been found experimentally (Bi and Poo, 1998). Theoretically, this weight-dependent potentiation yields a normal distribution of synaptic efficacies

rather than pushing each weight to one extreme or the other (van Rossum et al., 2000) as would be the case with an additive form of STDP.

6.3.1.3 Numerical Scheme

The differential equations described above are converted to finite difference equations and simulated using the Forward-Euler numerical scheme with a time step $\Delta t = 0.02ms$. In the finite difference equations, the Dirac delta function has been replaced by the discrete approximation $S(x)$ as defined in (Amit and Brunel, 1997). Finally, in the original description, the change in synaptic weight (Equation (6.8)) was instantaneous and so $\Delta t/\tau_{\Delta_g}$ is defined to be a learning rate constant, ρ, in the corresponding finite difference equation.

6.3.1.4 Training and Stimuli

Before the visual images are presented to the first excitatory layer (layer 1), they are preprocessed by a set of Gabor filters, which accord with the general tuning profiles of simple cells in V1 (Jones and Palmer, 1987; Cumming and Parker, 1999; Lades et al., 1993). The filters provide a unique pattern of filter outputs for each transform of each visual object, which is passed through to the first layer of the network. These filters are known to provide a good fit to the firing properties of V1 simple cells, which respond to local oriented bars and edges within the visual field (Jones and Palmer, 1987; Cumming and Parker, 1999). The input filters used are computed by the equations described in Section 1.3.1. In the experiments reported in this chapter, an array of Gabor filters is generated at each of 128 $\times$ 128 retinal locations with the parameters given in Table 6.1. The outputs of the Gabor filters are used as the basis to generate Poisson spike trains as follows.

$$P\{\text{input } cell(x, y, f) \text{ spikes at } t\} = g(x, y, f) \cdot max\_rate\_scaling\_factor \cdot \Delta t \tag{6.9}$$

where f is the index of a filter used for the simulation and *max_rate_scaling_factor* is the maximum input neuron firing rate (set to 100 in the current simulation studies). The outputs of the Gabor filters coded in Poisson spike trains are enacted by the layer 0 (Gabor Filter) cells which propagate activity to the layer 1 excitatory neurons of the network according to the synaptic connectivity given in Table 6.1. That is, each layer 1 neuron receives connections from 30 randomly chosen Gabor filters localised within a topologically corresponding region of the retina. These distributions are defined by a radius shown in Table 6.1.

6.3.2 Performance Measures

6.3.2.1 Information Analysis of Average Firing Rate Responses of Single Cells

Information theory is used to quantify how selective the average firing rate responses of individual neurons are for members of a particular stimulus category. If a neuron responds invariantly to the members of a particular stimulus category but not to members of other stimulus categories, then the neuron carries a maximum amount of information about the presence of its preferred stimulus category.

Information theory is applied to the average firing rate responses of individual neurons in the network in order to be able to compare the information conveyed by the firing rates of neurons with the information conveyed by the temporal spike patterns emitted by PGs (described later in Section 6.3.2.2). In this way, it is possible to demonstrate in the simulations reported below the very large increase in representational capacity that is possible using the temporal spike time coding available with the emergence of polychronisation.

Information theory has been previously used to quantify the performance of single neurons tasked with learning a translation invariant response (across multiple retinal locations) to specific visual stimuli (Eguchi et al., 2015). If the responses r of a neuron carry a high-level of

information about the presence of a particular stimulus s across different retinal locations, then this implies that the neuron will respond selectively to the presence of that stimulus regardless of where the stimulus is presented on the retina.

In this study, transforms of the visual inputs such as translation or rotation are not explicitly introduced. However, since the input neural spike trains are generated based on Poisson distributions, there is a significant degree of stochasticity involved. This means that the exact timings of the input neuron spikes are different at each run. Therefore, in the current simulation study, different presentations of the same visual input to the network are considered as the "transforms" of the same stimulus category.

The amount of stimulus specific information that a specific cell carries is calculated using the formula described in Section 1.5.2.

6.3.2.2 Information Analysis of Temporal Spike Patterns Emitted by Polychronous Groups

Information theory is applied to quantify the amount of information conveyed by the temporal patterns of spikes emitted by PGs. Spike train data consists of time-ordered sequences of spikes. It has been proposed that, in the brain, the temporal spike patterns emitted by PGs may be utilised to encode larger amounts of information than codes relying solely on the average firing rates of neurons.

However, to simplify the analysis, in the simulations below information theory was applied to the analysis of PGs containing only two spikes emitted by a pair of neurons. In the simplest scenario involving only two neurons, A and B, with inter-spike delay k, the PG episode can be represented using the notation $A[k]B$ (Diekman et al., 2014). By applying the analytical technique described in this section to the simulations reported below, it is possible to demonstrate the emergence of frequently repeating PG episodes of the form $A[k]B$ that are specific to a specific stimulus category. A number of these two-neuron interactions could in principle chain together to form longer, more complex multi-neuron PGs.

The same information analysis technique described above in Section 1.5.2.1 is applied to frequently occurring spike-pair PGs of the form $A[k]B$, to investigate whether the network is able to represent different visual stimulus categories using this form of temporal coding. Based on the spike trains recorded during many stimulus presentations to the network, the probabilities that a given spike pair $A[k]B$ will occur in response to the presentation of each of the stimulus categories s are computed. These probabilities are based on the frequency of occurrence of the spike pair $A[k]B$ across multiple transforms of each stimulus s, i.e. across multiple presentations (transforms) of the each stimulus with different stochastic (Poisson) input representations. From these frequency distributions, the following probability table for each stimulus category s is construct:

$$\begin{aligned} ProbTable(i,j,d) = P\{&(\text{Presynaptic cell } j \text{ spikes at time } t-d) \mid \\ &(\text{Postsynaptic cell } i \text{ spikes at time } t)\} \end{aligned} \tag{6.10}$$

where i and j are the indices of two neurons under consideration, t is the time at which the cell i emits a spike, and d is the time interval that neuron i emits a spike after neuron j. Values of d is considered within the range of $[0, 10ms]$, where this time interval is divided into 10 equal bins of 1ms. It is important to note that the probability table is constructed purely based on the actual spike trains emitted by neurons, and does not take into account the actual synaptic connectivity between the neurons. This means that this technique highlights the correlations in spike times emitted by the cells involved but does not necessarily reflect actual synaptic connections. Given this method of analysis, there are potentially 167,772,160 ($nCells * nCells * maxDelay$) distinct spike-pair PG representations the output neuronal layer

can hold. This is 40,960 ($nCells * maxDelay$) times more than the case of a localist rate-coded neuronal representation.

In applying the information analysis methodology of Section 1.5.2.1 to analysing the information carried by spike-pair PGs, the probability table given by Equation (6.10) is regarded as $\vec{R}$, the set of responses to the set of stimuli, used in Equation (1.12). Thus, Equation (1.12) may now be used to compute the information carried by spike-pair PGs about the presence of particular stimulus categories s. With this technique, it can be quantified how selective such temporal spike-pair PGs are for members of a particular stimulus category. In other words, if a particular spike-pair PG responds invariantly to the members of a particular stimulus category s but not to the members of other stimulus categories, then the spike-pair PG would carry maximum information about the presence of its preferred stimulus category.

6.3.2.3 Polychronous Group Counting

A key diagnostic in the simulations reported below is to identify and count the PGs that have emerged in the network after visually-guided training.

As introduced in the Introduction, a PG is defined as the set of neurons that support the associated time-locked spike pattern. More formally, Martinez and Paugam-Moisy (2009) defined that "An s-triggered polychronous group refers to the set of neurons that can be activated by a chain reaction whenever the triggers N_k $(1 \leq k \leq s)$ fire according to the timing pattern t_k $(1 \leq k \leq s)$. The PG is denoted by: $N_1 - N_2 - ... - N_s$ $(t_1, ..., t_s)$ where the firing times t_k are listed in the same order as the corresponding triggers N_k (Martinez and Paugam-Moisy, 2009)".

The algorithm used by Izhikevich (2006) was adopted and modified to be applicable for our conductance based LIF neural network model. The basic algorithm is as follow: (1) identify a set of potential triggers consisting of s neurons with specific spike timings (e.g., $N_1 - N_2 - ... - N_s$ $(t_1, ..., t_s)$), and (2) find PGs by simulating the propagation of activity from activation of this set of triggers.

More specifically, the algorithm first finds all combinations of a given number of s neurons (in our case, $s = 3$) that have at least one postsynaptic cell in common. For each such tuple of neurons, it then looks for the relative spike timings, based on synaptic delay, that can excite the common postsynaptic neuron maximally and enough for it fire. If such neurons exist, then the tuple becomes a trigger set. The algorithm then simulates the firing of the triggers with the identified spike timings and records the propagation of neural activity through the network until it decays to zero. In order to truncate the possible cyclic PGs, an upper limit is set for the time span of a PG and the number of neurons recorded.

6.4 Simulation Studies

In the current simulation study, the network was trained and tested on the abstract visual objects shown in Figure 6.6. The shapes are a circle, a heart, and a star, which are coloured black and presented against a 128×128 light grey background. Each simulation begins with an initial period of visual training. During each training epoch, each of the three object shapes was presented for two seconds to the network. As explained in the model description, the images are convolved with Gabor filters (Equation (1.1) and (1.2)) that mimic the responses of edge detecting V1 simple cells. The stochastically generated Poisson spikes (Equation (6.9)) are then imposed upon layer 0 and are then propagated to the first layer (layer 1) of the network, and thence up through successive layers 2 to 4. During this, the synaptic connections from the Gabor filters to Layer 1 excitatory neurons, as well as the bottom-up, top-down and lateral connections between excitatory neurons across all four layers of the network, were modified using the STDP rule described in Equation (6.8). In order to test the behaviour of the network

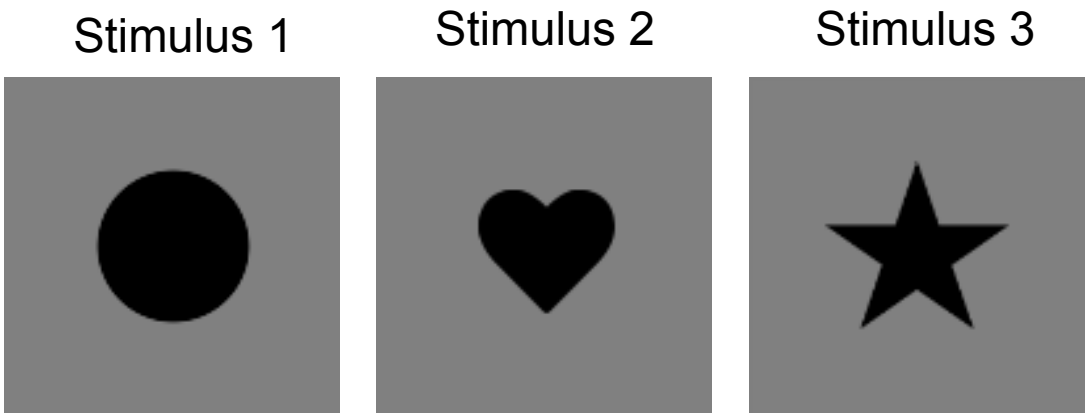

Figure 6.6: A set of three visual stimuli presented to the network: a circle, a heart, and a star.

before and after training, the same set of visual stimuli were also presented to the input layer with STDP turned off before and after training, and the resulting spike trains of neurons in the output layer were recorded for analysis.

6.4.1 Effect of Varying Synaptic Connectivity within Network

In this section, the performance of the model with different kinds of synaptic connectivity present within the network architecture is explored. Specifically, the model with the following four different network connectivities is simulated: 1. Feedforward (FF) connections only, 2. FF + Feedback (FB) connections, 3. FF + Lateral (LAT) connections, and 4. FF + FB + LAT connections. Our aim is to assess the contributions that each of these different types of synaptic connection make towards the operation of the model, including especially the relative amounts of stimulus information carried either in the firing rates of individual neurons or by the spike-pair PGs that emerge after training.

Single cell information analysis, as described in Section 1.5.2.1, was first conducted on the average firing rate responses of individual neurons in the output layer to the three visual stimuli shown in Figure 6.6 before and after training. The aim was to measure how much stimulus information was carried by the output neurons under the assumption of traditional rate coding. In this analysis, there are three different stimulus categories ($n = 3$). The maximum amount of information for a single neuron is $log_2(n)$ where n is the number of stimulus categories = 3. Therefore, the maximum amount of information that a neuron can carry about a particular stimulus is $log_2(3) \approx 1.58$ bits. Each visual stimulus was presented twice during testing, each time for a duration of two seconds. Since the precise spike timings of the input vary for the same visual stimulus between trials due to the stochastic nature of Poisson spike generation, this is conceptually equivalent to presenting two transforms of each stimulus category. Individual layer 4 neurons would have to respond invariantly over the two transforms of a single stimulus category, and not to transforms of the other stimulus categories, in order to carry maximum information about a single stimulus category.

Figure 6.7a shows the information analysis results for the Layer 4 neuron responses, based on the average firing rates over two seconds of presentation of each visual stimulus. Results before training are shown for the full network architecture with feedforward (FF) + feedback (FB) + lateral (LAT) synaptic connections (gray line). Very few output neurons carry the maximal information before training. Results after training are presented for the following four different network connectivities: 1. Feed-forward (FF) connections only (black dotted line), 2. FF + Feedback (FB) connections (black dash-dot line), 3. FF + Lateral (LAT) connections (black dashed line), and 4. FF + FB + LAT connections (black solid line). For all four different types of network connectivity, around 50-100 cells learned to carry the maximum single cell

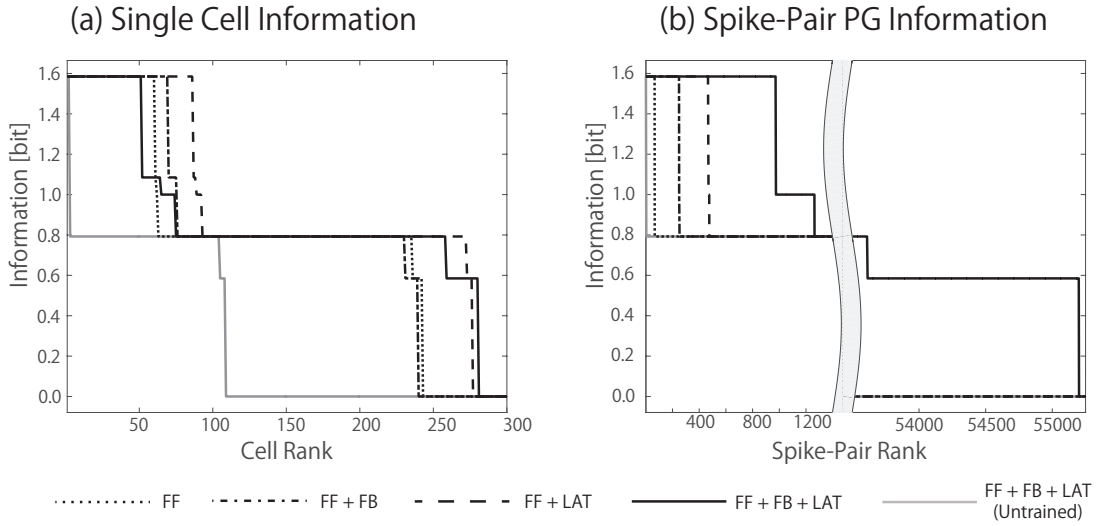

Figure 6.7: **(a) Single cell average firing rate-based information analysis**: The information carried by the output (4th layer) neurons about a specific object shape according to the procedure described in Section 1.5.2.1. The plot shows the maximum single cell information carried by 300 cells in Layer 4, where the cells are plotted along the abscissa in rank order. The results before training for the full network architecture with feedforward (FF) + feedback (FB) + lateral (LAT) synaptic connections are plotted in gray. It can be seen that before training very few output neurons carry the maximal information of 1.58 bits. The four different black lines show the results after training for four different network connectivities: 1. Feedforward (FF) connections only, 2. FF + Feedback (FB) connections, 3. FF + Lateral (LAT) connections, and 4. FF + FB + LAT connections. It is evident that all four types of network architecture have produced around 50 to 100 output neurons with maximal single cell information. **(b) Spike-pair PG information analysis**: The information carried by frequently occurring temporal spike-pair PGs in the output (4th layer) neurons about visual object shape according to the procedure described in Section 6.3.2.2. The plot shows the maximum information carried by spike-pair PGs in Layer 4, where the spike-pair PGs are plotted along the abscissa in rank order. The results before training for the full network architecture with FF + FB + LAT synaptic connections are plotted in gray. It can be seen that before training very few spike-pair PGs carry the maximal information of 1.58 bits. The four different black lines show the results after training for four different network connectivities: 1. FF connections, 2. FF + FB connections, 3. FF + LAT connections, and 4. FF + FB + LAT connections. It is evident that the full network architecture with FF + FB + LAT connections has produced the most spike-pair PGs with maximal information. Indeed, with the full network architecture almost 1,000 spike-pair PGs have reached the maximum information of 1.58 bits.

information (FF: 59, FF+FB: 69, FF+LAT: 85, and FF+FB+LAT: 51). Given that the output layer contains a total of 4,096 neurons, in each simulation only a relatively small fraction of these neurons learned to carry maximal information about stimulus identity in their average firing rates. Moreover, it is noticeable that the network incorporating all three kinds of connections gave the lowest performance.

Next, the new technique introduced in this chapter, spike-pair PG information analysis, which is instead based on frequently occurring temporal spike-pairs as described in Section 6.3.2.2, is applied . Figure 6.7b shows the information analysis results for spike-pair PG responses. Results before training are shown for the full network architecture with FF + FB + LAT synaptic connections (gray line). Before training very few spike-pair PGs carry the maximal information of 1.58 bits. The four different black lines show the results after training for the four different network connectivities: 1. FF connections, 2. FF + FB connections, 3. FF + LAT connections, and 4. FF + FB + LAT connections. All four network architectures produced large numbers of spike-pair PGs that carried the maximal amount of information about stimulus identity (FF: 66, FF+FB: 244, FF+LAT: 469, and FF+FB+LAT: 973). Importantly, it can be seen that the full network architecture with FF + FB + LAT connections produced the most spike-pair PGs with maximal information. Indeed, with the full network architecture almost 1,000 spike-pair PGs reached the maximum information of 1.58 bits. In particular, the number of spike-pair PGs with maximal information in the full network architecture is far greater (about 10 times) than the number of single output neurons achieving maximal information using a firing rate coding shown in Figure 6.7a.

Thus, the full network developed many spike pair PGs during visually-guided learning that

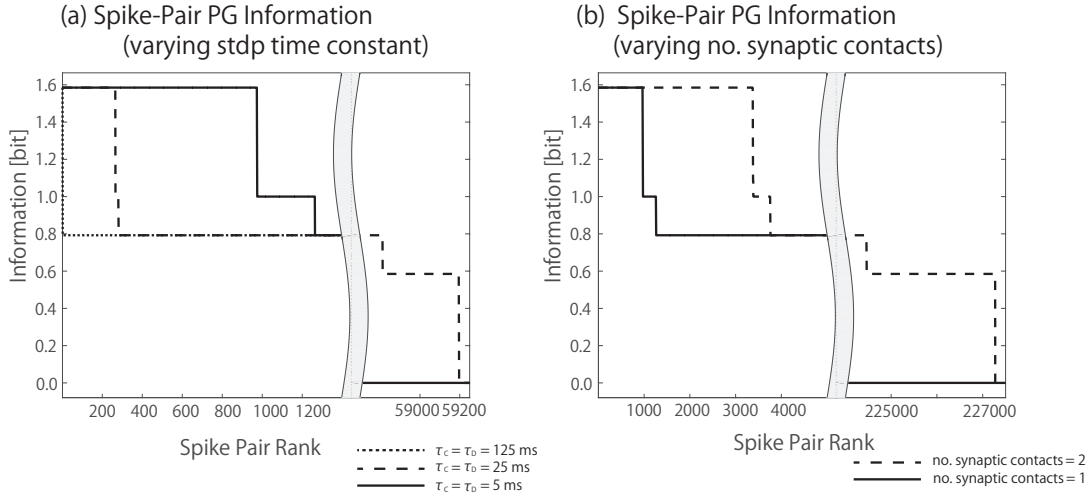

Figure 6.8: **Spike-pair PG information analysis**: The information carried by frequently occurring temporal spike-pair PGs in the output layer of a trained network about visual object shape according to the procedure described in Section 6.3.2.2. The two subplots show the maximum information carried by spike-pair PGs in Layer 4, where the spike-pair PGs are plotted along the abscissa in rank order. **(a)** Spike-pair PG information scores for 3 values of the STDP time constants $\tau_C = \tau_D = 125$ms, 25ms or 5ms. It is evident that shortening the STDP time constants promotes the emergence of spike-pair PGs with maximal information about which stimulus is presented to the network. In particular, the network develops the largest number of such stimulus specific spike-pair PGs when the STDP time constants are shortest (i.e. 5ms). As the STDP time constants are lengthened this reduces the temporal precision of the STDP and so degrades the emergence of PGs. (b) Spike-pair PG information scores for the cases where the number of plastic synaptic contacts between each pair of pre- and postsynaptic excitatory neurons is either one or two. It can be seen that there is a large increase in the number of spike-pair PGs with maximal stimulus information when there are two synaptic contacts with different transmission delays, rather than just one contact, between each pair of pre- and postsynaptic excitatory neurons. The presence of two synaptic contacts between each pair of pre- and postsynaptic excitatory neurons enables the STDP to select which of the transmission delays to strengthen in order to promote the development of PGs.

were tuned to specific stimuli. In particular, a major novel result of the current work is that this self-organisation of stimulus-specific spike-pair PGs occurred even when the stimulus input representations were *randomised* Poisson spike trains, in which the temporal ordering of spikes varied stochastically across different presentations of the same visual stimulus. The development of (spike-pair) PGs using STDP during visual training in such a spiking network is thus a highly robust process that operates perfectly well with randomised stimulus spike patterns in the lower stages of processing. Furthermore, the information results shown in Figure 6.7 clearly illustrate the greater potential of temporal coding over traditional rate coding in terms of representational capacity within a biologically realistic spiking neural network with bottom-up, top-down and lateral connections.

6.4.2 Effects of Varying Key Model Parameters

Next, the effects of varying key model parameters are investigated in order to identify which factors are important to the emergence of temporal coding by PGs. In particular, the effect of varying the STDP time constant and the number of synaptic contacts between each pair of pre- and postsynaptic neurons on the information carried by spike-pair PGs in the output layer is explored. This part of the investigation uses a full model with all three kinds of synaptic connectivity (FF + FB + LAT).

Figure 6.8a shows the spike-pair PG information carried by frequently occurring temporal spike-pairs in the output layer with the STDP time constants τ_C and τ_D both set to either 5 ms (solid line), 25 ms (dashed line) or 125 ms (dotted line). The results show that shortening the STDP time constants promotes the emergence of spike-pair PGs with maximal information about which stimulus is presented to the network. In particular, the network develops the largest number of such stimulus specific spike-pair PGs when the STDP time constants are shortest (i.e. 5ms). However, as the STDP time constant increases, the number of object specific spike-

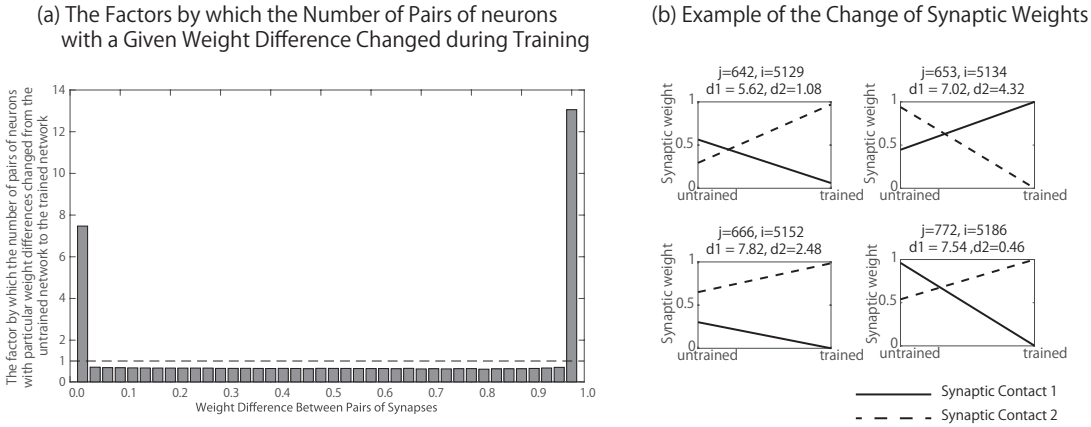

Figure 6.9: Simulation results demonstrating selection of effective synaptic delays by STDP during visual training. In this simulation, the model has two synaptic contacts with different fixed delays between each pair of pre- and postsynaptic excitatory neurons. STDP can then select one of the delays to be strengthened whilst weakening the other during visual training. For each pair of connected pre- and postsynaptic excitatory neurons calculated, both before training and after training, the absolute difference between the synaptic weights of their two synaptic contacts. Then, two frequency histograms, corresponding to before and after training, in which all such pairs of pre- and postsynaptic excitatory neurons were binned according to the absolute difference between the weights of their two synaptic contacts, were computed. These two frequency histograms were then used to compute the plot shown in **(a)**, which shows the result of dividing the frequency histogram after training by the frequency histogram before training on a bin by bin basis. Thus, subplot **(a)** shows the factor by which the number of pairs of neurons with a particular absolute difference between the weights of their two synaptic contacts changes after training. It can be seen that after training there was a large increase in the number of pairs of pre- and postsynaptic excitatory neurons with the maximum possible synaptic weight difference. This implies that during visual training, one of the synaptic weights went to its maximum value of 1.0, while the other synaptic weight went to the minimum value of 0.0. This represents successful synaptic delay selection by STDP. **(b)** shows four examples of the changes in the weights of the two synaptic contacts between a pair of pre- and postsynaptic excitatory neurons that occurred during training. The solid and dashed-lines in each subplot represent the strengths of the two synaptic contacts between a pre- and postsynaptic neuron, each with a different synaptic delay. For each pair of neurons, these plots show selective strengthening of one synaptic contact with a particular delay, but weakening of the other synaptic contact with a different delay. Again, it is clearly evident that STDP is able to selectively strengthen and weaken synaptic connections during visual learning according to their respective transmission delays.

pair PGs decreases. Increasing the STDP time constant makes the precise timing of the spikes less important for learning, making the effect of learning more similar to that expected from traditional Hebbian learning in a rate coded model. This result implies an important role of temporally precise STDP for the development of temporal coding.

Figure 6.8b compares the information carried by frequently occurring temporal spike-pair PGs in the output layer when the number of plastic synaptic contacts between each pair of pre- and postsynaptic excitatory neurons is either one or two. In the latter case, the two synaptic contacts between each pair of pre- and postsynaptic excitatory neurons had different durations that were randomly assigned at the beginning of the simulation and remained fixed throughout. Only the strengths of these connections could be modified by STDP during visually guided learning. The results show that having multiple synaptic contacts, and hence multiple axonal transmission delays, between each pair of excitatory neurons, increases the number of spike-pair PGs with maximum information. The presence of two synaptic contacts between each pair of pre- and postsynaptic excitatory neurons enables the STDP to select which of the transmission delays to strengthen in order to promote the development of (stimulus specific) PGs as hypothesised.

The results shown in Figure 6.8b demonstrate how such a spiking model may exploit the ability to effectively select (self-organise) the durations of the axonal delays in the plastic connections between excitatory neurons. If there is no self-organisation during visual learning over synaptic delay lengths, then it is effectively pre-specified by the initial random distribution of axonal transmission delays within the network whether a given neuron can be a part of a particular PG. By allowing multiple plastic synaptic contacts, with different delays, between each pair of pre- and postsynaptic excitatory neurons, I expected that STDP would effectively select which of these axonal delays to strengthen.

Table 6.2: Statistics of the PGs that emerged in network models with different kinds of synaptic connectivity (FF only, FF + FB, FF + LAT, and FF + FB + LAT) after training.

	FF		**FF+FB**		**FF+LAT**		**FF+FB+LAT**	
Number of PGs	562		1689		827		32317	
	mean	**s.d.**	**mean**	**s.d.**	**mean**	**s.d.**	**mean**	**s.d.**
Total Number of Spikes in PG	8.91	6.16	8.94	7.30	8.70	5.98	11.41	6.56
Longest Path of Spikes in PG	1.00	0.00	1.12	0.44	1.31	0.53	2.55	0.97

I next took a deeper look into the selective strengthening and weakening of synaptic contacts with different axonal transmission delays between each pair of pre- and postsynaptic excitatory neurons. This analysis was carried out on the same simulation with two synaptic contacts with different fixed delays between each pair of neurons. In this case, STDP could select one of the delays to be strengthened while weakening the other during visual training. For each pair of connected pre- and postsynaptic neurons, the absolute difference between the synaptic weights of the two synaptic contacts both before and after training was calculated. The absolute difference in the values of the two synaptic weights after training should reflect how effectively the STDP has selectively strengthened one connection with a particular delay but weakened the other connection with a different delay, which is necessary to promote the emergence of many stimulus specific spike-pair PGs. Specifically, before and after training, frequency histograms in which pairs of pre- and postsynaptic excitatory neurons were binned according to the absolute difference between the weights of their two synaptic contacts was computed.

Figure 6.9a shows the result of dividing the frequency histogram after training by the frequency histogram before training on a bin by bin basis. Thus, subplot **(a)** shows the factor by which the number of pairs of neurons with a particular absolute difference between the weights of their two synaptic contacts changes after training. It can be seen that after training there was a large increase in the number of pairs of pre- and postsynaptic excitatory neurons with the maximum possible synaptic weight difference. This implies that during visual training, one of the synaptic weights went to its maximum value of 1.0, while the other synaptic weight went to the minimum value of 0.0. This represents successful synaptic delay selection by STDP.

Figure 6.9b shows examples of synaptic modifications for four pairs of pre- and postsynaptic excitatory neurons, where each such pair of neurons has two synaptic contacts with different transmission delays. For each pair of neurons, these plots show selective strengthening of one synaptic contact with a particular delay, but weakening of the other synaptic contact with a different delay. These results clearly demonstrate that STDP is able to selectively strengthen and weaken synaptic connections during visual learning according to their respective transmission delays. The model thus selects and self-organises its effective synaptic delays, which can greatly facilitate the emergence of stimulus specific (spike-pair) PGs.

6.4.3 The Emergence of Larger Scale Polychronous Groups

In this section, the development of larger scale PGs (i.e. containing more than just two neurons) is explored. In particular, it is investigated how the development of these PGs is influenced by changing the kind of synaptic connectivity implemented within the network. For each simulation with a different connectivity structure (FF only, FF + FB, FF + LAT, and FF + FB + LAT), all the potential PGs triggered from cells in the third layer of the network were identified based on the synaptic connectivity, conduction delays, and synaptic weights as explained in Section 6.3.2.3. Furthermore, based on the actual spike trains recorded during testing, it was investigated whether any of the activated PGs had learned to be stimulus specific.

Table 6.2 shows the statistics of the PGs that emerged in network models with different kinds of synaptic connectivity after training. The top row shows the total number of PGs that were identified. The general trend is that as the synaptic connectivity becomes more complex, i.e. with more types of connection, the number of PGs increases. In particular, by far the largest number of PGs were found in the full network architecture with FF + FB + LAT connections.

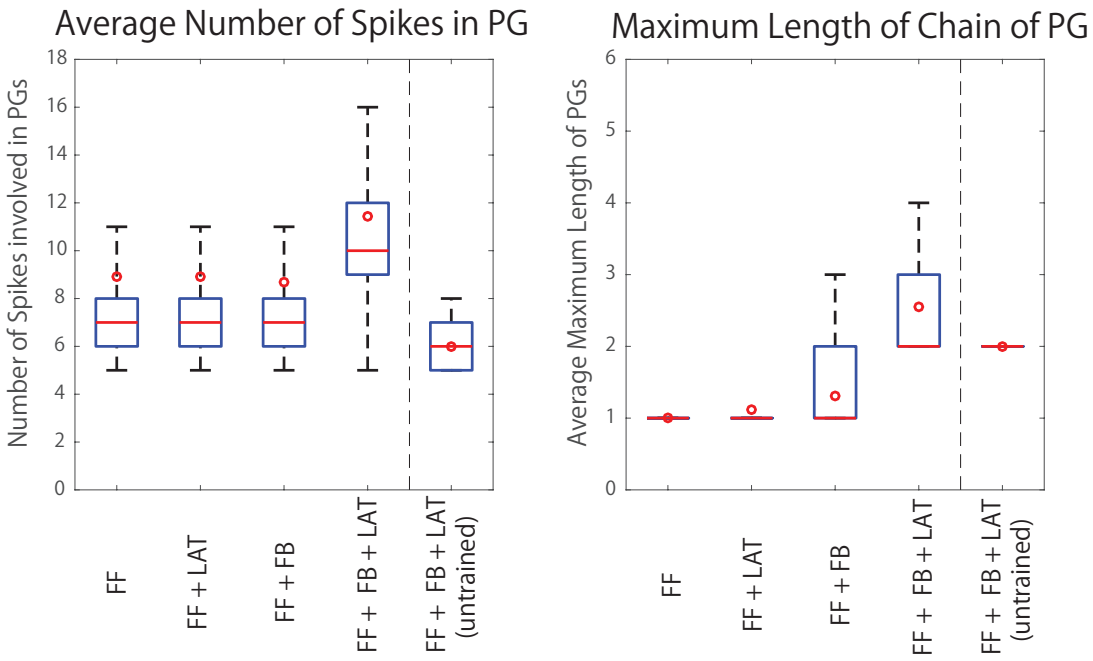

Figure 6.10: Box plots showing key performance statistics of the PGs that emerged in network models with different kinds of synaptic connectivity (FF only, FF + FB, FF + LAT, and FF + FB + LAT) after training. Subplot **(a)** shows the distribution of the average number of spikes in a PG, while subplot **(b)** presents the distribution of the longest path of spikes in a PG. For both subplots the red horizontal lines indicate the median and the red circles indicate the means. It is evident that the full trained network architecture (FF + FB + LAT) gives rise to the largest mean number of spikes in each PG and mean longest path length of each PG compared with the three other reduced network connectivities. The results from the four trained networks are compared with those from the untrained full network architecture (FF + FB + LAT) shown on the right of each subplot. By comparing the results for the full network architecture before and after training, it can be seen that training has led to a significant increase in the mean number of spikes in each PG and the mean longest path length of each PG.

The middle row of Table 6.2 shows the mean number of spikes in each PG. Lastly, the bottom row presents the mean longest path length of each PG, where the longest path is defined as the number of neurons involved in the longest chain of spikes emitted by the PG due to the activation of the trigger neurons (Izhikevich, 2006). It can be seen that both of these statistics get an increase as the network architecture includes more types of synaptic connection. Indeed, the full network architecture (FF + FB + LAT) also gives rise to the largest mean number of spikes in each PG and the mean longest path length of each PG. The full network architecture is clearly the most efficacious for promoting the emergence of polychronisation.

The detailed statistical distributions underlying the mean values shown in the second and third rows of Table 6.2 are shown as box plots in Figure 6.10. Figure 6.10a shows the distribution of the average number of spikes in a PG, while Figure 6.10b presents the distribution of the longest path of spikes in a PG. For both subplots the red horizontal lines indicate the median and the red circles indicate the means. As already shown in Table 6.2, the full trained network architecture (FF + FB + LAT) gives rise to the largest mean number of spikes in each PG and mean longest path length of each PG compared with the three other reduced network connectivities. The results from the four trained networks are compared with those from the untrained full network architecture (FF + FB + LAT) shown on the right of each subplot. By comparing the results for the full network architecture before and after training, it can be seen that training has led to a significant increase in the mean number of spikes in each PG and the mean longest path length of each PG.

Next, it was investigated whether the PGs that developed after training in the full network architecture (with all three connectivity types FF + FB + LAT) had learned to respond to a particular stimulus category. This was done by analysing the responses of the actual trigger events for these PGs in Layer 3 to the three visual stimuli: the circle, heart, and star. The

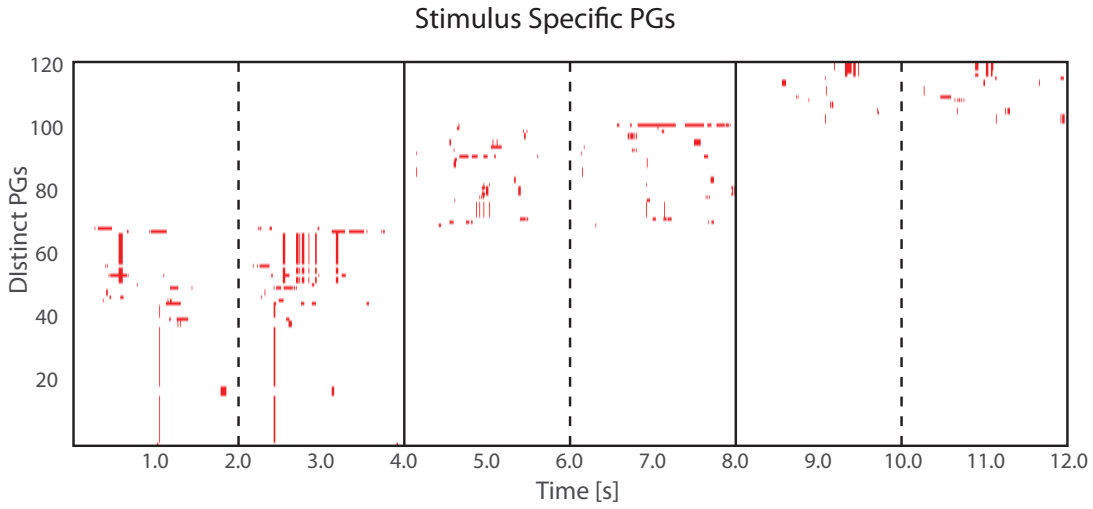

Figure 6.11: Graphical representation of the occurrences of stimulus-selective PG trigger events in Layer 3 of the trained full network architecture (FF + FB + LAT) when tested on the three visual stimuli: the circle, heart, and star. The stimulus-selective PG trigger events were first identified as being selective for one of the stimuli by using information analysis. In the figure, the occurrences of these PG trigger events when each of the stimuli is presented twice to the network, each time for two seconds, are shown. Specifically, the circle is presented during 0 to 2 seconds, and then again during 2 to 4 seconds. Next, the heart is presented during 4 to 6 seconds, and then again during 6 to 8 seconds. Finally, the star is presented during 8 to 10 seconds, and then again during 10 to 12 seconds. The distinct stimulus-selective PG trigger events identified by the information analysis are numbered along the ordinate. It can be seen that PG trigger events 1-70 responded selectively to the circle, PG trigger events 71-102 responded to the heart, and PG trigger events 103-123 responded to the star.

stimulus-selective PG trigger events were identified as being selective for one of the stimuli by using information analysis. This was done using a similar information analysis to that used for single cells in Section 1.5.2.1, but instead using the occurrences of the PG trigger events in the spike trains of Layer 3 neurons as described in Section 6.3.2.3.

Figure 6.11 plots the occurrences of the stimulus-selective PG trigger events (identified by the information analysis) when the network was tested on the three visual stimuli: the circle, heart, and star. The figure shows the occurrences of these PG trigger events when each of the stimuli is presented twice to the network, each time for two seconds. Specifically, the circle is presented twice, followed by two presentations of the heart and then two presentations of the star. It can be seen that PG trigger events 1-70 respond selectively to the circle, PG trigger events 71-102 respond to the heart, and PG trigger events 103-123 respond to the star.

These results confirm that in the trained full network architecture (FF + FB + LAT), large numbers (i.e. greater than 100) of PGs respond selectively to just one of the stimuli, and do so across different presentations of that stimulus.

6.4.4 The Emergence of Binding Neurons

Finally, the PGs from the full network model (FF + FB + LAT) was analysed, and they were found to respond to specific stimuli in Section 6.4.3 for the presence of the hypothesised binding neurons as illustrated in Figure 6.3a. Figure 6.12 shows examples of activated PGs, binding neurons, and visual features represented by input Gabor filters that drive the cells in the PGs. Simulation results are presented from the trained full network architecture when tested on the three visual stimuli: the circle, heart, and star. Each row (a-c) represents a PG of neurons that responds selectively to one of the stimuli (subplot in left pane) and the visual features represented by the input Gabor filters with strong connections to particular neurons explicitly identified in the PGs (two subplots in right pane). The PGs shown in rows (a), (b) and (c) respond selectively to the circle, heart, and star respectively.

In the PG plots shown on the left of Figure 6.12 the neurons are identified by small circles and the strengthened connections between the neurons are represented by lines. In particular,

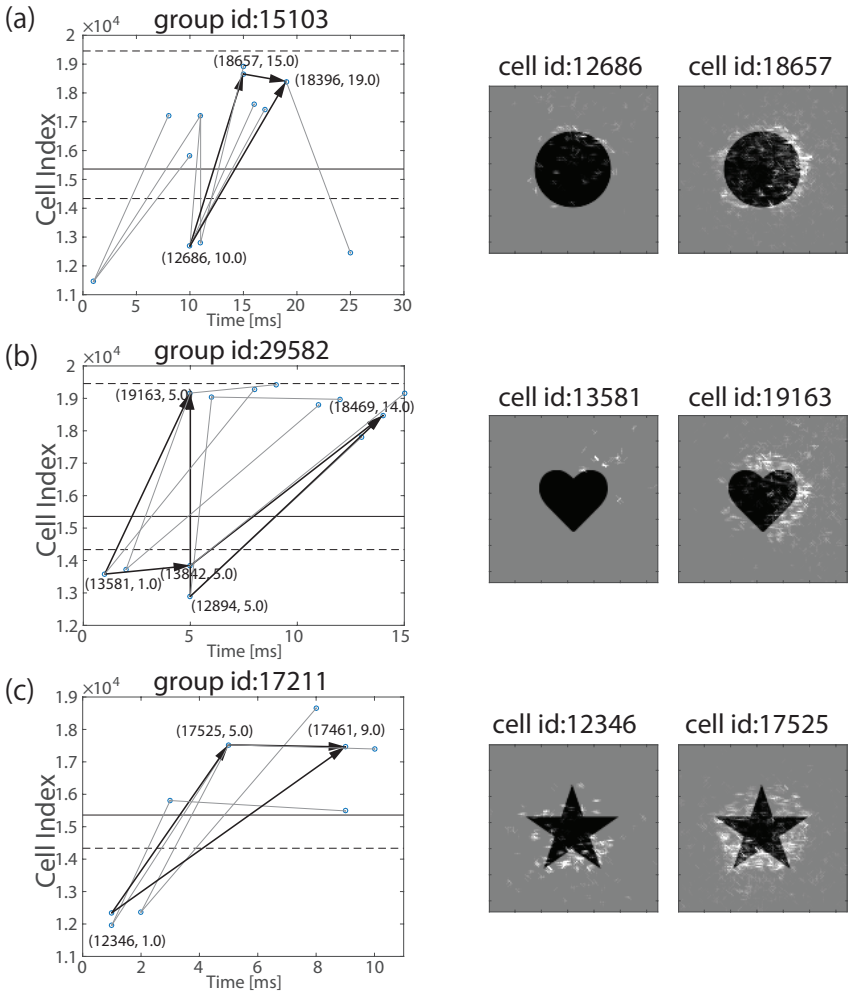

Figure 6.12: **Examples of activated PGs, binding neurons, and visual features represented by input Gabor filters that drive the cells in the PGs.** Simulation results are presented from the trained full network architecture (FF + FB + LAT) when tested on the three visual stimuli: the circle, heart, and star. Each row (a-c) represents a PG of neurons that responds selectively to one of the stimuli (subplot in left pane) and the visual features represented by the input Gabor filters with strong connections to particular neurons explicitly identified in the PGs (two subplots in right pane). Specifically, the two subplots presented in the right panes of rows (a-c) show the visual input features represented by a layer 3 trigger neuron for the PG (left), and a layer 4 neuron from within the same PG (right). The PGs shown in rows (a), (b) and (c) respond selectively to the circle, heart, and star respectively. In the PG plots (shown on the left) the neurons are identified by small circles and the strengthened connections between the neurons are represented by lines. The neurons are plotted along the abscissa according to the relative timings of their spikes within the PGs, which was determined by the axonal transmission delays of the strengthened connections between the neurons. In particular, rows (a) and (c) present clear examples of the hypothesised binding neurons. In these PG plots, the three neurons that make up the three-neuron binding circuit (as illustrated in Figure 6.3a) have bold connections between them. The two subplots in the right pane of row (a) show the visual features represented by the input Gabor filters that have strong feedforward connections to the low-level feature neuron 12686 and the high-level feature neuron 18657. It can be seen that the low-level feature neuron 12686 receives strong connections from a simpler set of input Gabor filters than the high-level feature neuron 18657, which is consistent with our underlying theoretical framework about binding taking place between low-level and high-level features. A similar binding relationship between three neurons is shown in row (c). Row (b) shows that mixtures of polychronous representation types emerge in the same neuronal layers.

rows (a) and (c) present clear examples of the hypothesised binding neurons. In these PG plots, the three neurons that make up the three-neuron binding circuit (as illustrated in Figure 6.3a) have bold connections between them. For example, in row (a), there are three neurons in the binding circuit as follows: neuron 12686 (a PG trigger neuron in layer 3) represents the low-level feature, neuron 18657 (a layer 4 output neuron) represents the high-level feature, and neuron 18396 is the related binding neuron between these two features. It can be seen that the low-level feature neuron 12686 sends a connection to the high-level feature neuron 18657, and both feature neurons 12686 and 18657 send connections to the binding neuron 18396. In particular, it can be seen from the axonal transmission delays shown in the plot that if the low-level feature neuron 12686 is driving the high-level feature neuron 18657, then the spikes emitted by both of these feature neurons will arrive at the binding neuron 18396 at about the same time and so reinforce each other. Thus, the binding neuron 18396 will fire if the low-level feature neuron 12686 is actually driving the high-level feature neuron 18657. A similar binding relationship between three neurons is shown in row (c). Row (b) shows that mixtures of polychronous representation types emerge in the same neuronal layers.

The two subplots presented in the right panes of rows (a-c) in Figure 6.12 show the visual features represented by the input Gabor filters that have strong feedforward connections to two neurons from the PG shown in the left pane. Specifically, the right pane shows a layer 3 trigger neuron for the PG (left), and a layer 4 neuron from within the same PG (right). To produce these subplots, the feedforward synaptic connections between successive layers are traced back to the input Gabor filters in order to determine the specific visual features that drive the responses of the higher layer neurons. Starting from a particular neuron in layer 3 or 4, the connections from the previous layer that have the highest weights is selected, repeating this process through successive layers until the connections reach the Gabor filters in the input layer. This then allows us to plot the pattern of Gabor input filters that the neuron in layer 3 or 4 has become tuned to. Looking at the right panes of Figure 6.12(a,c), it is clear that the layer 3 neurons (left) represent simpler low-level features, whilst the layer 4 neurons (right) represent more global features of the entire object. For example, in row (a) of Figure 6.12 it can be seen that the low-level feature neuron 12686 in layer 3 (left) receives strong connections from a simpler set of input Gabor filters than the high-level feature neuron 18657 in layer 4 (right). Moreover, comparison to the corresponding PG plots in the left panes of Figure 6.12(a,c) shows that the layer four neurons are being driven by the simpler layer 3 neurons, with the outputs of both layer 3 and 4 neurons driving an associated binding neuron. These results are consistent with our underlying theoretical framework about binding taking place between low-level and high-level visual features.

6.4.5 Feedforward projection of information about low-level visual features to higher neuronal layers

Simulations of the full spiking network architecture (FF + FB + LAT) provided examples of the kind of feedforward propagation of visual information hypothesised in Section 6.2.0.3 and illustrated in Figure 6.4a. The binding neurons presented in rows (a) and (c) of Figure 6.12 were in fact in layer 4. Thus, in each of these examples, the low-level feature neuron was in layer 3, the high-level feature neuron was in layer 4, and the binding neuron was also in layer 4. In these cases, information about the low-level feature represented in layer 3, including its local image context (i.e. that the low-level feature represented in layer 3 is part of the high-level feature represented in layer 4), is projected onto the binding neuron in layer 4. These simulation results confirm the feasibility of the hypothesis that low-level visual information is propagated forwards (i.e. bottom-up) to higher layers in the manner proposed in Section 6.2.0.3. This could allow information about low-level features to be represented in the highest layers of the network, where in principle this information could be read out by subsequent behavioural systems.

6.5 Discussion

In this chapter, the operation of a biologically detailed neural network model of the primate ventral visual system was explored. The model incorporates the following key aspects of cortical dynamics and architecture: (i) the model implements spiking neural dynamics in which the timings of action potentials or 'spikes' are simulated explicitly, (ii) STDP is used to modify the synaptic connections during visually-guided learning, (iii) the network architecture incorporates bottom-up, top-down and lateral synaptic connections, (iv) the synaptic connectivity between neurons incorporates distributions of axonal conduction delays of varying durations, (v) in some simulations multiple synaptic connections with different axonal transmission delays are incorporated between each pair of pre- and postsynaptic neurons. These are basic known aspects of the architecture and function of the visual cortex. Using this model architecture, a number of major computational hypotheses was explored as follows.

6.5.1 Emergence of Polychronization

Some previous authors have proposed that a visual scene could be partitioned into separate object regions by synchronisation of neuronal firing (Kreiter and Singer, 1996; Evans and Stringer, 2012, 2013). In this scenario, the spikes emitted by the neurons representing each individual object become synchronised in time, while the spikes emitted by neurons encoding different objects become desynchronised. This mechanism of synchronisation allows a spiking network model to, say, segment and individually bind several different object regions of an image. However, Evans and Stringer (2013) have found that such neuronal synchronisation may be destroyed if natural distributions of axonal transmission delays are included.

I instead hypothesised that even if the visual stimuli (circle, heart, and star) presented to the network were encoded in the input layer by randomised Poisson spike trains, the synaptic connectivity in the later layers of the network would self-organise using STDP during visually-guided learning such that polychronous groups (PGs) would emerge naturally. Moreover, I anticipated that individual PGs would learn to respond to particular stimuli that the network was trained on. This was confirmed in our simulations reported above.

The output (4th) layer was found to carry more stimulus information if a temporal coding based on patterns of spike times within PGs instead of assuming traditional rate coding by individual neurons is assumed. Our results found that the inclusion of feedback and lateral connections in the network structure led to an increase in the number and length of PGs (especially spike-pairs). In particular, the full network architecture with feedforward (FF), feedback (FB) and lateral (LAT) synaptic connections produced the most spike-pair PGs with maximal stimulus information. These spike pair PGs were tuned to specific stimuli.

A major novel result of the current work is that this self-organisation of stimulus-specific spike-pair PGs occurred even when the stimulus input representations were *randomised* Poisson spike trains, in which the temporal ordering of spikes varied stochastically across different presentations of the same visual stimulus. The development of (spike-pair) PGs using STDP during visual training in such a spiking network is thus a highly robust process that operates perfectly well with randomised stimulus spike patterns in the lower stages of processing.

The development of temporal PG codes was shown to be dependent on the temporal specificity of the STDP learning rule used to modify the synaptic connections. It was found that the network develops the largest number of spike-pair PGs with maximal information about which stimulus is presented to the network when the STDP time constants are shortest (i.e. 5ms). However, increasing the STDP time constants in the simulations had the effect of decreasing the number of object specific spike-pair PGs that emerged. The explanation for these observations is that increasing the STDP time constants makes the precise timing of the spikes less important for learning, which in turn makes the synaptic weight modification more similar to traditional Hebbian learning in a rate coded model. Consequently, these simulation results

suggest an important role for temporally precise STDP in the development of temporal coding.

Another novel feature of some of the simulations reported in this chapter was the incorporation of multiple synaptic contacts with different axonal transmission delays between each pair of pre- and postsynaptic neurons. This corresponds to a presynaptic neuron making multiple synaptic connections on different parts of the dendritic branching of a postsynaptic neuron as is seen among real neurons in the brain. In such a network architecture, STDP was able to select which synapses to strengthen and which synapses to weaken, which promoted the visually-guided development of PGs of spiking neurons. Thus, during self-organisation the network is able to effectively select for synaptic transmission delays between pre- and postsynaptic neurons, which results in a greater representational capacity.

6.5.2 Emergence of Binding Neurons

Over the last twenty years, our laboratory has developed a hierarchical, *rate-coded*, neural network model, VisNet, of the primate ventral visual pathway, which has also been used in many studies in this thesis (Wallis and Rolls, 1997; Galeazzi et al., 2015; Eguchi et al., 2016). This network model represents low-level visual features in the lower layers and higher level features or objects in the higher layers, but there is no way to identify which features are part of which objects from the activity of these neurons. How visual features are bound together must underpin how we segment a visual scene into objects and parts of objects, and thus how we make sense of the visual world. Duncan and Humphreys (1989) provide a good description of this hierarchical process as follows:

> "A fully hierarchical representation is created by repeating segmentation at different levels of scale. Each structural unit, contained by its own boundary, is further subdivided into parts by the major boundaries within it. Thus, a human body may be subdivided into head, torso, and limbs, and a hand into palm and fingers. Such subdivision serves two purposes. The description of a structural unit at one level of scale (animal, letter, etc.) must depend heavily on the relations between the parts defined within it (as well as on properties such as color or movement that may be common to the parts). Then, at the next level down, each part becomes a new structural unit to be further described with its own properties, defined among other things by the relations between its own sub-parts. At the top of the hierarchy may be a structural unit corresponding to the whole input scene, described with a rough set of properties (e.g., division into light sky above and dark ground below)."

The new generation of spiking neural network simulations reported in this chapter, in which the timings of action potentials or spikes are explicitly simulated, aim to solve this feature binding problem. Our basic conception is that within the PGs that emerge during visually-guided learning are embedded *binding neurons* that represent the binding relationships between low-level and high-level visual features. It is assumed that neurons in the network behave as 'coincidence detectors' in that they require a volley of spikes from presynaptic cells to arrive simultaneously at the postsynaptic cell in order for the postsynaptic cell to fire itself. The basic three neuron binding circuit is illustrated in Figure 6.3a.

Importantly, our new approach to solving the feature binding problem in biological spiking neural networks relies on polychrony instead of synchrony. In other words, I am interested in how simply segmenting a visual scene into several distinct object regions can accord with the semantically rich, hierarchical visual experience of primate vision as described by Duncan and Humphreys (1989). As discussed earlier, in the brain, the low-level and high-level visual features may in fact be represented by their own temporal patterns of spikes distributed across PG of neurons, and these two PGs may then drive a third PG representing the binding relationship between these visual features. This more complex scenario, in which the visual features and the

binding relations between these features are represented by their own PGs, is likely to be what actually happens in the brain. The simple three neuron binding circuit shown in 6.3a would then be a small part of the three corresponding PGs shown in 6.3b. The use of polychronisation with binding neurons seems to offer far greater richness in terms of the structural and semantic representation of visual scenes.

Simulations of the full spiking network architecture (FF + FB + LAT) presented above demonstrated the emergence of binding neurons, which were part of the same kind of three neuron binding circuit as shown in Figure 6.3a. These simulation results were shown in Figure 6.12. In these simulations, the binding neurons represented the binding relationships between lower level feature neurons in layer 3 and higher level feature neurons in layer 4. Moreover, the individual PGs, in which these binding neurons were embedded, responded to specific visual stimuli (the circle, heart or star). Such binding neurons were originally proposed by von der Malsburg (1999), but without an explanation of how they might emerge naturally during visual development. The simulations reported above demonstrated that such binding neurons may develop automatically within the PGs that emerge during visually-guided learning with STDP. In particular, these binding representations were shown to emerge even when the visual stimuli are encoded by randomised (Poisson) spike trains in the input layer. The binding neurons that develop carry measurable information about which low-level features are driving (and hence part of) which high-level features. Our theory predicts that such binding neurons should develop across the visual field, at every layer of the feature hierarchy, and at every spatial scale within a natural visual image.

Our model of the primate ventral visual pathway contains bottom-up, top-down and lateral synaptic connections in order to reflect the known architecture of this part of the brain. Given this kind of connectivity, there are a variety of ways of realising the three neuron binding circuit shown in Figure 6.3a. For example, the binding neuron might be in the same layer as the low-level feature neuron, or in the same layer as the high-level feature neuron, or in a different area completely. Below the feedforwarded projection of information about low-level features that may occur if the binding neuron is in the same layer as the high-level feature neuron (holographic principle) is discussed. However, wherever the binding neurons are located, they will carry measurable information about the binding relations within a visual scene. Moreover, the theory presented in this chapter implies that binding neurons will develop throughout all visual processing areas of the visual cortex, thus representing the binding relations across the visual field and at every spatial scale. A rich tapestry of binding neurons through the layers could help to provide a rich hierarchical structural description of a scene, rather analogous to that described above by Duncan and Humphreys (1989).

The example given in Figure 6.3a shows how binding neurons may learn to represent the fact that a particular low-level visual feature such as a horizontal or vertical bar is driving, and therefore part of, a given high-level feature such as the letter T. However, binding neurons may learn to represent many other kinds of relationship between features within an image. For example, a binding neuron might learn to respond when a low-level feature (such as a vertical bar) is part of an intermediate-level feature (such as the letter T), which is in turn part of a high-level feature (such as the word CAT). In this case, the binding neuron receives simultaneous inputs from the low-level, intermediate and high-level feature neurons, as shown in Figure 6.13a. Alternatively, a binding neuron could represent that a low-level feature (such as a vertical bar) is simultaneously part of two different higher level features (such as the letter T and the word CAT), as shown in Figure6.13b. Or a binding neuron could represent that two low-level features (such as a vertical bar and a horizontal bar) are both part of a higher level feature (such as the letter T), as shown in Figure6.13c. There are a vast number of such relationships that could be represented by binding neurons. What kinds of relationship actually get represented will depend on the visual images used to train the network model. In future research, I would explore what kinds of binding relationship become represented in the model as it is trained on

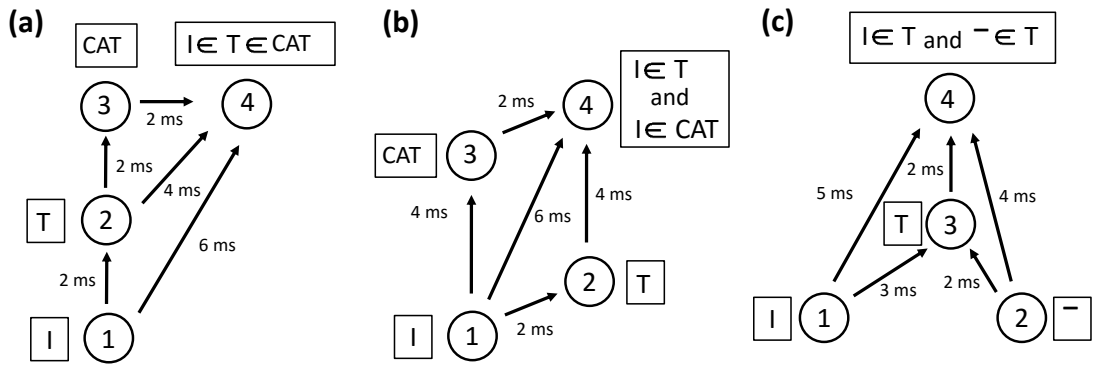

Figure 6.13: Examples of three different kinds of more complex binding relationships. **(a)** a binding neuron might learn to respond when a low-level feature (such as a vertical bar) is part of an intermediate-level feature (such as the letter T), which is in turn part of a high-level feature (such as the word CAT). In this case, the binding neuron receives simultaneous inputs from the low-level, intermediate and high-level feature neurons. **(b)** Alternatively, a binding neuron could represent that a low-level feature (such as a vertical bar) is simultaneously part of two different higher level features (such as the letter T and the word CAT). **(c)** a binding neuron could represent that two low-level features (such as a vertical bar and a horizontal bar) are both part of a higher level feature (such as the letter T).

lots of natural images. Such binding information is essential to the rich semantic analysis and interpretation of visual images performed by the visual brain.

This proposal sharply contrasts with the feature integration theory (FIT) of Treisman and Gelade (1980), which posits that that there is only a single locus of attention within the visual field where visual features are bound together. Some researchers have tried to relate feature binding to the speed of visual search for target objects among nontarget distractors. Given that feature integration theory assumes there is only a single locus of attention where feature binding takes place, this implies a serial search for a visual search task that requires feature binding, but allows faster parallel search for other search tasks that do not require feature binding. In contrast, I proposed that feature binding is carried out by binding neurons that operate simultaneously across the whole visual field including at every spatial scale. In this case, there is no need for binding to be limited to a single spatial locus of attention, and the time taken for visual search would not be governed by the need to perform a serial search with a single locus of attention. Instead, binding may operate in parallel across the visual field, and the search time would be related to other factors determining the intrinsic difficulty of the task. This is supported by the study of Duncan and Humphreys (1989). These authors found no clear dichotomy between serial and parallel modes of search. Instead, they reported that search efficiency was found to decrease as the targets became more similar to nontargets, or if the nontargets became more dissimilar to each other. This finding contradicts the assumption of feature integration theory that there is a single locus of feature binding, which leads to serial search for those tasks that require feature binding and parallel search for tasks that do not.

However, although our theory permits feature binding to operate in parallel across the entire visual field, it would still be expected that visual processing, which includes feature binding, would be somewhat degraded away from the spatial locus of attention. This could occur because the neural representation of the part of the visual scene at the site of spatial attention, which might be highlighted due to high acuity foveal fixation or top-down attentional facilitation, would compete strongly with visual processing of the rest of the scene through inhibitory interneurons. This strong inhibition from the attended part of the scene would likely degrade visual processing elsewhere, including feature binding operations. This would explain various psychophysical findings about binding in human vision Wolfe and Cave (1999). However, this is quite different from the underlying assumption of feature integration theory, which actually requires only a single spatial locus of attention to perform feature binding, and so cannot permit any binding elsewhere.

6.5.3 Feedforward projection of information about low-level visual features to higher neuronal layers

The simulations presented in this chapter showed that visual information about low-level features was, in fact, being propagated up through the neuronal layers of the network in a similar fashion to that illustrated in Figure 6.4a. This kind of feedforward propagation of low-level visual information may be important if the behaviour-related areas of the brain are restricted to reading out visual information from only the higher processing stages of the visual system. As discussed above, low-level visual features such as oriented bars and edges are represented in the earliest cortical stages (e.g. V1 and V2) of visual processing. However, when we perceive an object we are aware of visual features at every spatial scale and complexity of visual form. The simulations reported in this chapter show how all of this visual information could in principle be projected upwards through successive stages of visual processing. In particular, the neural representation of a low-level feature in the higher layers of the network encodes both the identity of the low-level feature as well as its local image context in terms of hierarchical binding relationships to higher level features. For example, in 6.4a, binding neuron 3 represents the fact that there is a vertical bar in some localised region of the retina, and that this vertical bar is part of the alphabetic letter T.

The bottom-up projection of low-level visual information through successive layers of visual processing is an automatic consequence of our hypothesised solution to the feature binding problem using polychronisation and the emergence of binding neurons. The most simple mechanism for achieving the bottom-up projection of low-level visual information is illustrated in Figure 6.4a. The mechanism is essentially the same as the three neuron binding circuit shown in Figure 6.3a, but with the binding neuron 3 situated in the same higher layer as neuron 2 that represents the high-level feature T. As described above, binding neuron 3 represents that there is a vertical bar in some local region of the retina, and that this vertical bar is part of the letter T. Thus, figure 6.4a shows how information about the presence of a low-level feature (i.e. there is a vertical bar in some localised region of the retina) in the lower layer has been projected up to the higher layer along with its local image context (i.e. the vertical bar is part of the letter T). This proposed mechanism for the bottom-up projection of information about low-level features could operate through successive cortical stages of visual processing, including across the visual field and at every spatial scale.

Simulations of the full network architecture (FF + FB + LAT) provided examples of this kind of feedforward propagation of visual information. The binding neurons presented in rows (a) and (c) of Figure 6.12 were in fact in layer 4. Thus, in each of these examples, the low-level feature neuron was in layer 3, the high-level feature neuron was in layer 4, and the binding neuron was also in layer 4. In these cases, information about the low-level feature represented in layer 3, including its local image context (i.e. that the low-level feature represented in layer 3 is part of the high-level feature represented in layer 4), is projected onto the binding neuron in layer 4. These simulation results confirm the feasibility of the hypothesis that low-level visual information is propagated forwards (i.e. bottom-up) to higher layers in the manner proposed in Section 6.2.0.3.

Figure 6.4b shows how the basic mechanism illustrated in Figure 6.4a could be repeated iteratively up through successive layers in order to project information about low-level features into the highest (output) layer of the network. In Figure 6.4b, visual information about the presence of a vertical bar is first projected up from the first neuronal layer to the second layer, where it is represented by binding neuron 3. Neuron 3 represents the fact that there is a vertical bar in some localised region of the retina, and that this vertical bar is part of the alphabetic letter T. Then, a similar binding mechanism combines the output from binding neuron 3 with the output of neuron 5 representing a cat, where these combined outputs drive binding neuron 6. Binding neuron 6 then represents the fact that there is a vertical bar in a local region of the retina, which is part of the letter T, which in turn is part of the word CAT. In this way, the

information about the lowest level feature is projected upwards and preserved in the highest layer of the network. Indeed it is possible that a large amount of information about low-level features could be projected upwards in this manner and preserved in the highest layers for readout by subsequent behavioural systems. We refer to this as a *holographic principle* because information about visual features at every level of complexity and scale may be preserved in the highest (output) layer(s) of the network. In using the term holographic principle, we are conscious of a somewhat similar usage to describe the preservation of information at the event horizon surface of a black hole (Susskind, 1995).

It is important to note that the binding neurons 3 and 6 in the highest layers of the two network architectures shown in Figures 6.4a and 6.4b represent the presence of a vertical bar in some local region of the retina which is explicitly part of a higher level feature / object (e.g. the letter T) or hierarchy of features (e.g. the letter T, which is part of the word CAT). Consequently, such binding neurons do not simply respond to the presence of a vertical bar at some retinal location regardless of local image context (i.e. the higher level features / objects which the vertical bar is part of). The high-level feature / object still needs to be presented to the network in order to elicit a response from these kinds of binding neuron in the upper layers. The holographic principle described here is thus consistent with neurophysiological observations that neurons in the later stages of the ventral visual pathway tend to respond to more complex visual forms than the simple oriented bars represented in early cortical stages such as V1 and V2.

Lastly, the bottom-up projection of information about low-level visual features, as illustrated in Figure 6.4, would seem to negate the need for top-down synaptic connections in the network architecture. This presents something of a conundrum. If the holographic principle holds true in some way, then we will need to develop a theory of how top-down signal transmission fits into this framework. This might lead to much greater complexity in visual processing than so far considered here. However, I believe that the observed architecture and neurodynamics of the visual cortex provide the necessary signposts for eventually understanding and simulating the singular semantic richness of biological vision.

6.5.4 Future Work

The utilisation of polychronisation within a spiking neural network allows the model to develop binding neurons with the crucial property that they respond if a low-level feature neuron is actually participating in driving a high-level feature neuron. Only in this case will the binding neuron be fully informative that the low-level feature is part of the high-level feature. It is important that the binding neuron does not fire whenever the low-level feature neuron and the high-level feature neuron happen to be simultaneously active. For this reason, I propose that binding may not be soluble within a traditional rate-coded network, but will instead require the full spiking neuronal dynamics of the brain.

The kind of spiking network architecture discussed in this chapter seems to be needed to solve the feature binding problem. Generally, the parietal lobe is the main (Friedman-Hill et al., 1995; Shafritz et al., 2002) contributor to the visual feature binding. However, it is important to aware that the border ownership representation found in the primary visual cortex can also be seen as an example of the 'binding neuron' which requires a type of visual feature binding. In rate coded models, such as our own VisNet model, *individual postsynaptic neurons do not record which subset of presynaptic neurons are actually driving them*, and consequently the network as a whole does not maintain an explicit representation of which presynaptic neurons are driving particular postsynaptic neurons throughout the network. Thus, in a sense, the rate-coded network 'leaks' this essential binding information, which is necessary for representing, and making sense of, how the visual features within a scene are related to each other. This is particularly problematic when postsynaptic neurons represent high-level visual features (such as a complex visual form, object or face) with some degree of transform (e.g. location, view

or scale) invariance, as is typical in the higher layers of the primate ventral visual pathway (Wallis and Rolls, 1997; Booth and Rolls, 1998; Perry et al., 2010). It is particularly in this situation that the network needs to maintain a representation of exactly which presynaptic neurons are driving each postsynaptic neuron in order to represent the relationships between the lower level and higher level features throughout the visual field and at every spatial scale. In this current chapter, the network has not been trained to develop transform (e.g. location) invariant responses to the visual stimuli. However, this problem is challenged in the following Chapter 7.

Further theoretical evidence that rate coding may be insufficient to solve feature binding has been provided in Chapter 5, which demonstrated the failure of binding within a rate-coded model of *border ownership cells* (Eguchi and Stringer, 2016). This class of visual cells, which have been found in cortical areas V1 and V2, respond to oriented edges like simple cells, but are also modulated by which side of an object the edge occurs on (Zhou et al., 2000). Such border ownership cells are clearly modulated by top-down visual signals about local object context from outside their classical receptive field. Importantly, border ownership cells are thought to represent the binding relationship between a localised border edge region of an object and the object itself. I provided a detailed argument for why our rate-coded model of border ownership cells failed on binding when more than one object was presented to the network at a time, and also proposed that spiking dynamics would be needed to solve this problem (Eguchi and Stringer, 2016). In the final chapter, I explore how border ownership cells may develop in the kind of spiking network model investigated in this chapter, where the border ownership cells are examples of our hypothesised binding neurons.

The binding hypothesis proposed in this chapter also provides a way in which the visual system might localise (parts of) objects in space. When we look at a visual scene, we are aware of visual features at all such spatial scales. In particular, we are aware of the (e.g. retinal) locations of low-level features such as the edges of objects. This kind of information may be represented by edge detecting (e.g. simple) cells in lower visual areas, which have small receptive fields of about 1 or 2 degrees in size. Neurons with such small receptive fields can effectively localise the edge of an object in space. However, through a process of feature binding, we also see these edges as parts of the boundaries of their respective objects. Thus, the binding of a localised edge represented in an early visual area (e.g., border ownership cells (Zhou et al., 2000)) to an object representation at a higher stage of processing provides a way in which (the parts of) objects may be localised in space. Hence, the development of binding neurons within PGs as proposed in this chapter provides a plausible explanation for how such binding might operate and play a key role in the localisation of (parts of) objects in space.

A particularly interesting feature of the proposed theories in this chapter is that it potentially reveals a sharp contrast between processing in the visual brain and the operation of biologically implausible rate-coded neural network algorithms such as backpropagation of error. The architecture of the visual cortex, which is simulated in the spiking neural network models presented in this chapter, could potentially the development of binding neurons that represent the binding relationships between low-level and high-level features at all spatial scales throughout a visual scene. However, a biologically implausible neural network algorithm such as rate-coded backpropagation of error (Hertz et al., 1991) would not develop binding neurons and so could not represent such binding information. That is, although rate coded networks (trained by backpropagation of error or otherwise) might be efficient at learning arbitrary mappings they would not be able to represent the essential binding information needed to semantically analyse natural visuospatial scenes in the same way as the primate brain.

As a first step towards this, in future work I propose to develop *hybrid* neural networks that combine the kind of biologically-inspired spiking (unsupervised learning) network presented in this chapter with a more traditional engineering (supervised learning) network such as backpropagation of error. In such a hybrid network, the biological network may operate as a

preprocessing stage that extracts not only the visual features but also the binding relationships between those features across the visual field and at every spatial scale. All of this visual information may then be propagated from the biological network to the engineering network for, say, image classification.

Chapter 7

The Neural Basis of Border Ownership Representations in a Spiking Neural Network Model

In this chapter, a solution to the classic feature *binding problem* is demonstrated by modelling the visually-guided development of *border ownership cells* in cortical areas V1 and V2 of the primate visual cortex. These neurons respond to oriented edges, like classic simple cells, but are also sensitive to which side of an object the boundary edge occurs. In this way, the neurons are thought to represent which object a particular edge belongs to. Border ownership cells are thus thought to play a key role in feature binding, in this case binding a relatively low-level feature such as a boundary edge to a higher level object. In Chapter 5, the development of border ownership cells in an established rate-coded neural network model, VisNet, of the primate ventral visual pathway (Eguchi and Stringer, 2016) was modelled. The border ownership cells were top-down modulated by neurons in higher layers that mimicked the responses of V4 neurons encoding the local curvature of object boundary elements. However, the border ownership cells within our earlier rate-coded model failed to respond properly when the network was presented with visual scenes containing more than one visual object. This was shown to be due to the fact that the top-down modulation from higher cortical stages was not specific to particular retinal locations. In the current work presented in this chapter, it is shown how the problem may in fact be solved in a spiking neural network, in which the timings of action potentials or 'spikes' are explicitly simulated. That is, in the spiking network simulations reported, it is shown that the border ownership cells are able to maintain their proper response characteristics even when multiple objects are presented to the network simultaneously. The new spiking network model exploits the emergence of *polychronization* during training using Spike-Timing-Dependent Plasticity (STDP), in which groups of neurons learn to fire in regularly repeating temporal sequences. In particular, the border ownership cells that develop in the simulations are found to be examples of the *binding neurons* hypothesised in Chapter 6 to provide a general solution to the feature binding problem. Importantly, the well-known problem of "superposition catastrophe" that occurs within a rate-coded model when multiple objects are presented simultaneously may be overcome within the current spiking network model. Taken together, the failed rate-coded simulations reported in Chapter 5 (Eguchi and Stringer, 2016) and the successful spiking network results presented here provide strong theoretical support for the binding hypothesis advanced in Chapter 6.

7.1 Introduction

7.1.1 Overview

In this chapter, computer simulation is used to demonstrate a solution to the classic feature *binding problem*, which concerns how the brain represents the hierarchical relationships between lower and higher level features within a visual scene (von der Malsburg, 1999), by modelling the visually-guided development of *border ownership cells*. This class of visual cells, which have been found in cortical areas V1 and V2 of the primate visual cortex, respond to oriented edges like simple cells, but are also modulated by which side of an object the edge occurs on (Zhou et al., 2000). Such border ownership cells are clearly modulated by top-down visual signals about local object context from outside their classical receptive field. This top-down modulation enables border ownership cells to represent which object a particular edge belongs to. Border ownership cells are thus thought to play a key role in solving feature binding in the visual brain, specifically representing the binding relationship between a localised border edge of an object and the object itself.

In Chapter 5, the development of border ownership cells in VisNet (Eguchi and Stringer, 2016) was modelled. The border ownership cells were top-down modulated by neurons in higher layers that mimicked the responses of V4 neurons encoding the local curvature of object boundary elements (Pasupathy and Connor, 2002). However, the border ownership cells within our earlier rate-coded model failed to respond properly when the network was presented with visual scenes containing more than one visual object. It is in situations where multiple objects are seen together, where the visual brain must bind lower level features to their correct higher-level features or objects, that the binding problem especially rears its head (von der Malsburg, 1999). The failure of the rate-coded model was shown in Chapter 5 to be due to the fact that the top-down modulation from higher cortical stages was not specific to particular retinal locations (Eguchi and Stringer, 2016). I proposed that the problem may be solved in a spiking neural network, in which the timings of action potentials or 'spikes' are explicitly simulated.

In the current work presented in this chapter, the visually-guided development of border ownership cells within a spiking neural network model of the primate visual system is modelled. In the new spiking network model, the border ownership cells are able to maintain their proper response characteristics even when multiple objects are presented to the network simultaneously. Training the spiking network model using Spike-Timing-Dependent Plasticity (STDP) (Bi and Poo, 1998; Markram et al., 1997) leads to the emergence of *polychronization*, in which groups of neurons learn to fire in regularly repeating temporal sequences (Izhikevich, 2006). In particular, the border ownership cells that develop in the simulations are found to be examples of the *binding neurons* hypothesised in Chapter 6 to provide a general solution to the feature binding problem. Importantly, the well-known problem of "superposition catastrophe" (von der Malsburg, 1999) that occurs within a rate-coded model when multiple objects are presented simultaneously is overcome within the current spiking network model. The current chapter demonstrates how exactly border ownership cells may develop in a spiking network model, where the border ownership cells are examples of the hypothesised binding neurons.

In Chapter 6, it was proposed that the spiking network architecture utilised in our modelling study seems to be needed to solve such feature binding problems. One important reason given for this is as follows. In traditional rate-coded models, such as the model used in Chapter 5, individual postsynaptic neurons do not record which subset of presynaptic neurons are actually driving them (Eguchi and Stringer, 2016). Consequently, these models do not maintain an explicit representation of which presynaptic neurons are driving particular postsynaptic neurons throughout the network. Thus, in a sense, the rate-coded network 'leaks' this essential binding information, which is necessary for representing, and making sense of, how the visual features within a scene are related to each other. This is particularly problematic when postsynaptic neurons represent high-level visual features (such as a complex visual form, object or face) with

some degree of transform (e.g. location, view, or scale) invariance, as is typical in the higher layers of the primate ventral visual pathway (Wallis and Rolls, 1997; Booth and Rolls, 1998; Perry et al., 2010). It is particularly in this situation that the network needs to maintain a representation of exactly which presynaptic neurons are driving each postsynaptic neuron in order to represent the relationships between the lower and higher level features throughout the visual field and at every spatial scale. If this kind of binding information is leaked away in a rate-coded network, then it will not be possible for a translation invariant neuron representing part of an object boundary in a higher neuronal layer (e.g. corresponding to cortical area V4) to selectively top-down modulate the correct subset of border ownership cells at just one retinal location in the lower layers (e.g. corresponding to cortical areas V1 and V2). This fundamental limitation of rate-coding was key to why the model in Chapter 5 failed when the network was presented with more than one object simultaneously (Eguchi and Stringer, 2016).

The binding mechanism proposed in Chapter 6, which is applied to modelling border ownership cells in this current study, provides a way in which the visual system might localise (parts of) objects in space. When we look at a visual scene, we are aware of visual features at all spatial scales. In particular, we are aware of the precise (retinotopic) locations of low-level features such as the edges of objects. This kind of information may be represented by edge detecting (e.g. simple) cells in lower visual areas such as V1, which have small receptive fields of about 1 or 2 degrees in size. Neurons with such small receptive fields can effectively localise the edge of an object in space. However, through a process of feature binding, we also see these edges as parts of the boundaries of their respective objects. Thus, the binding of a localised edge represented in an early visual area by border ownership cells to an object representation at a higher stage of processing provides a way in which (the parts of) objects may be localised in space. Hence, the simulated development of border ownership cells as examples of the binding neurons hypothesised in Chapter 6, as demonstrated in this chapter, provides a plausible explanation for how such binding might operate and play a key role in the localisation of (parts of) objects in space.

Taken together, the failed rate-coded simulations in Chapter 5 (Eguchi and Stringer, 2016) and the successful spiking network results presented below provide strong theoretical support for the binding hypothesis advanced in Chapter 6.

In the following three subsections, the followings are discussed: the earlier rate-coded neural network model of border ownership cells developed in Chapter 5 (Eguchi and Stringer, 2016), the failure of this rate-coded model of border ownership cells when multiple visual stimuli are presented simultaneously, and our hypothesised operation of a *spiking* neural network model of border ownership cells as demonstrated in this current chapter.

7.1.2 Rate-coded neural network model of border ownership cells developed in Chapter 5

In Chapter 5, the visually-guided development of border ownership cells within a rate-coded neural network model of VisNet (Eguchi and Stringer, 2016) was simulated. The VisNet architecture consisted of a hierarchical series of four competitive layers of neurons. The version of VisNet used in their study incorporated both feedforward (bottom-up) and feedback (top-down) synaptic connections. These connections were modified during visually-guided training by a trace learning rule, which drove the development of neurons in the higher layers that displayed responses that were translation invariant as a visual stimulus was shifted across different retinal locations. The VisNet model was *rate-coded* in the sense that it did not explicitly represent the actual timings of the action potentials or spikes emitted by neurons, but instead represented only the average firing rate of each neuron at any given moment in time. This model simplification was made to reduce the computational cost of the simulations.

When VisNet was trained on images of many different objects in Chapter 5, the neurons in the higher layers of the model developed the same response characteristics as neurons observed

in area V4 of the primate visual system by Pasupathy and Connor (2002). Such neurons represented the boundary contour elements of 2-dimensional object shapes. That is, individual neurons responded selectively to boundary elements with a specific curvature at a particular location with respect to the centre of mass of the object. For example, some higher stage neurons might learn to respond to the presence of a vertical edge on the left boundary of an object, while other neurons might learn to respond to the presence of a vertical edge on the right boundary of an object. A subpopulation of such neurons can provide a distributed encoding of the entire boundary shape of a visual object. Moreover, the neurons responded with translation invariance as an object was shifted across different retinal locations.

The self-organisation of the feedforward connections through visual training, during which the neurons in the higher layers learn to encode the local boundary elements of objects in a similar manner to the V4 neurons reported by Pasupathy and Connor (2002), was in fact originally described and demonstrated in Chapter 2 (Eguchi et al., 2015). Further details are therefore given in this earlier chapter. In particular, it was shown that while individual neurons in the higher layers developed strong feedforward connections from neurons in the lower layers representing the same preferred boundary element, the higher layer neurons also developed slightly weaker connections from other neurons in the lower layers representing other nearby boundary elements that could co-occur depending on the shape of an object presented to the network. These connections from other lower layer neurons were important in that they ensured that the higher layer neurons responded selectively according to where their preferred boundary contour element occurred on the object boundary.

As VisNet was continued to train on the same object images in Chapter 5, strong polysynaptic feedback connections subsequently developed from the higher layer neurons encoding the local boundary elements of objects to neurons in the lower layers such as V1 and V2. These self-organised feedback connections were then able to modulate the responses of edge-detecting neurons in layers V1 and V2 in a manner that depended on where their preferred edge element occurred on the boundary of an object, rather like the response characteristics of border ownership neurons observed by Pasupathy and Connor (2002). For example, some of the V1/V2 edge-detecting neurons that developed in the lower layers of VisNet responded to the presence of a vertical straight edge at a particular retinal location only when that edge was at a particular boundary location (e.g. left or right) with respect to the centre of the object. A detailed mathematical description of how this process works is provided in Chapter 5 (Eguchi and Stringer, 2016).

In summary, in Chapter 5, it has been shown that the firing characteristics of border ownership neurons reported by Zhou et al. (2000), in which the responses of V1 and V2 neurons are modulated by which side of an object the edge occurs on, may be replicated by incorporating both feedforward (bottom-up) and feedback (top-down) associatively modifiable connections within VisNet (Eguchi and Stringer, 2016). This allows neurons in the early layers to develop their firing responses through visually-guided competitive learning driven by a combination of both bottom-up and top-down visual signals. After training, visual information about local image context is conveyed by the top-down connections to modulate the responses of neurons in the lower layers, which may then mimic the observed firing characteristics of border ownership neurons.

7.1.3 Failure of rate-coded model of border ownership cells when multiple visual stimuli are presented simultaneously

The rate-coded model used in Chapter 5 developed border ownership neurons that maintained their responses even when the network was tested on any one of a large number of different object shapes. However, it was found that the rate-coded model failed to display the proper firing properties of border ownership neurons under more general stimulus conditions in which more than one object was presented to the network at the same time after training. It is

in situations where multiple objects are seen together that feature binding, in which lower-level features must be bound to the correct higher-level features or objects, becomes a more challenging problem (von der Malsburg, 1999). The reason for the failure of the rate-coded model developed in Chapter 5 was carefully analysed and may be summarised as follows.

It is started by considering how the feedforward (bottom-up) connections self-organise when the network is trained with a large collection of differently shaped objects. For example, let us assume that two of these objects, Objects 1 and 2, have vertical straight edges on either their left or right boundaries, respectively. In the simulations carried out in Chapter 5 the network was trained on one object at a time shifting across different retinal locations. During this training phase, neurons in a higher layer (corresponding to cortical visual area V4) learn to encode the local boundary elements of objects in a similar manner to the V4 neurons reported by Pasupathy and Connor (2002). In particular, the higher layer neurons have a relatively large fan-in of polysynaptic connections from the retina, and the trace learning rule operating in the bottom-up connections enables these neurons to develop responses that are translation invariant across all trained retinal locations. That is, the same subset of higher layer neurons learns to respond to a straight vertical edge on their preferred side of an object regardless of where that object appears on the retina. In this case, after training, Objects 1 and 2 will excite higher layer neurons representing a straight vertical edge on either the left or right object boundary, respectively, regardless of the retinal location in which each of the objects is seen.

Now consider how the feedback (top-down) connections self-organise during training. The trace learning rule operating in the top-down connections ensures that the higher layer neurons representing a straight vertical edge on either the left or right object boundary will develop strong polysynaptic connections via competitive learning with the corresponding subsets of lower layer neurons representing a straight vertical edge on either the left or right object boundary, respectively. However, most importantly, each of the higher layer neurons will develop strong connections to the corresponding lower layer neurons across *all* of the trained retinal locations. In this case, after training, individual higher layer neurons will simultaneously modulate the responses of corresponding subsets of lower layer neurons over all trained retinal locations. Hence, after the rate-coded model has been trained, the translation invariant higher layer neurons representing a straight vertical edge on the left object boundary stimulate subsets of lower layer neurons representing a straight vertical edge on the left object boundary across all trained retinal locations, *including the locations of both Objects 1 and 2 presented together*. Similarly, translation invariant higher layer neurons representing a straight vertical edge on the right object boundary stimulate subsets of lower layer neurons representing a straight vertical edge on the right object boundary across all trained retinal locations.

To summarise, the overall effect of presenting two objects simultaneously after training, where Object 1 has a straight vertical edge on its left boundary while Object 2 has a straight vertical edge on its right boundary, is as follows. If the first object has a straight vertical edge on its left boundary, then this excites higher layer neurons representing a straight vertical edge on the left object boundary, which in turn modulates lower layer neurons representing a straight vertical edge on the left object boundary at both object locations. Similarly, if the second object has a straight vertical edge on its right boundary, then this excites higher layer neurons representing a straight vertical edge on the right object boundary, which in turn also modulates lower layer neurons representing a straight vertical edge on the right object boundary at both object locations. This results in both subsets of lower layer cells, representing straight vertical edges on the left and right object boundaries, being activated at the locations of both objects simultaneously. Thus, in the rate-coded model, the top-down modulation of lower layer neurons representing object boundary elements by polysynaptic connections from higher layer neurons fails to be specific to the actual locations of the object boundary elements on the retina. The border ownership information in the lower layer is thus lost in this situation of multiple objects, which represents a failure of feature binding. A similar failure will occur even for the simpler

situation of a single object presented to the network with vertical straight edges on both the left and right boundaries.

This result could be counter-intuitive to those who remember the basic principle of trace learning rule that can associate any visual inputs cluttered in time. However, it is important to notice that the rate-coded network in fact has successfully developed the translation invariant representation of a particular shape of a local contour element with trace learning as promised. On the other hand, the border Ownership representation is not a translation invariant representation. That is instead location specific but only shapes (other than the border part) invariant representation. In other words, while the network tries to associate any object that contains a particular contour element on a particular side with trace learning rule, the border ownership cell needs to be selective only to the subset of those when it is presented at a particular location on the retina. This kind of representation has never been promised to be learned with trace learning rule in any preceding study.

The above argument suggests that the incorporation of additional top-down connections, although necessary, is not sufficient by itself to allow the network to develop robust border ownership representations in the early layers and thus solve this kind of feature binding problem. So what further biological details need to be incorporated into the model to allow it to form robust border ownership representations under the more general stimulus conditions of multiple objects seen together?

7.2 Hypothesis

In Chapter 5, I hypothesised that the failure of the border ownership representations in the VisNet model when multiple objects were presented was primarily due to the implementation of rate-coding in the model. That is, VisNet only represents the average firing rate of each neuron, and not the actual timings of the action potentials emitted by neurons as occur in the brain. The architecture and operation of the visual cortex of the primate brain differs from the VisNet model used in Chapter 5 in the following important ways. Firstly, real neurons in the brain communicate by emitting electrical pulses called action potentials or spikes. Secondly, the modification of synaptic strengths during learning has been found to depend on the relative timings of the spikes emitted by the pre- and post-synaptic neurons. For example, neurophysiology studies have shown that a synapse may be strengthened when the spike from the pre-synaptic neuron occurs about 20ms before the spike from the post-synaptic neuron, but weakened when the spike from the pre-synaptic neuron occurs about 20ms after the spike from the post-synaptic neuron (Bi and Poo, 1998; Markram et al., 1997). This kind of learning is called Spike-Timing-Dependent Plasticity (STDP). Thirdly, the passage of an action potential from one neuron to the next may be subject to an axonal transmission delay of several milliseconds, where different axonal connections between neurons may have different time delays. Though it should be noted that the transmission delay associated with an individual axonal connection between any two neurons tends to be fairly constant through time. Modelling studies have shown that when such randomised distributions of axonal delays are incorporated into a spiking neural network, then this produces memory patterns in the form of repeating temporal loop of neuronal firings. This phenomenon has been termed *polychronization* (Izhikevich, 2006). Our laboratory has shown that the emergence of these temporal memory loops is further enhanced within a recurrently connected spiking network with randomised distributions of axonal conduction delays when the strengths of synaptic connections are modified by STDP (Chapter 6). Recognising the potential importance of all these biological features of cortical operation, in Chapter 5, I hypothesised that such a spiking model, which also incorporates bottom-up, top-down and lateral associatively modifiable excitatory connections, may develop border ownership neurons in the lower visual layers (corresponding to cortical areas V1 or V2) that respond selectively to a vertical straight edge on either the left or right boundary of an object presented at the neuron's preferred

retinal location, *in a way that is unaffected by the presence of another object seen at a different nearby location on the retina.* More generally, in Chapter 5, I hypothesised that the biological features discussed above all play important roles in how the primate visual system solves the feature binding problem (von der Malsburg, 1999). Consequently, in this current chapter, I explore below how border ownership representations may develop in a new spiking neural network version of the VisNet model, which incorporates bottom-up, top-down and lateral excitatory connections, distributions of axonal transmission delays, and STDP.

7.2.1 The Proposed Role of Polychronization and Feature Binding in the Development of Border Ownership Cells

Our hypothesised mechanism for the development and operation of border ownership cells exploits the general solution to the feature binding problem demonstrated in the modelling study conducted in Chapter 6. The behaviour of a biologically realistic hierarchical neural network model of the primate ventral visual system was investigated. The architecture of this model is shown in Fig. 6.5. The model represents successive neuronal stages of processing along the primate ventral visual pathway. It is comprised of five layers of excitatory pyramidal neurons, which may be thought of as loosely representing cortical visual areas V1, V2, V4, TEO, and TE. However, since the architecture of the model is still quite a crude simplification of actual cortical structure, this putative correspondence between the layers of the model and specific areas of the visual cortex should not to be regarded as precise. Layer 0 represents the input layer, which corresponds to cortical area V1. The responses of neurons in Layer 0 are set in accordance with the outputs of Gabor filters that mimic the responses of edge-detecting simple cells. Importantly, the model incorporates the following important general features of cortical operation:

- (i) The model implements spiking neural dynamics in which the timings of action potentials or 'spikes' are simulated explicitly.
- (ii) STDP is used to modify the synaptic connections during visually-guided learning. If a spike from a presynaptic neuron arrives at a postsynaptic neuron just before the postsynaptic neuron emits a spike, then the synapse is strengthened (LTP). Otherwise, if the spike from the presynaptic neuron arrives at the postsynaptic neuron just after the postsynaptic neuron emits a spike, then the synapse is weakened (LTD).
- (iii) The network architecture incorporates bottom-up, top-down, and lateral excitatory synaptic connections, which are associatively modifiable during visual training. This connectivity reflects the known architecture of the visual cortex.
- (iv) The synaptic connectivity between neurons incorporates distributions of axonal conduction delays of varying durations, from a few milliseconds to tens of milliseconds.
- (v) There are multiple (2) synaptic connections between each pair of pre- and postsynaptic neurons, where these connections have different axonal transmission delays. This permits STDP to strengthen just one (or a subset) of these connections in order to effectively select the functional transmission delay between the two neurons.

The simulations carried out in Chapter 6 using the above spiking network architecture developed many groups of neurons, which we refer to as *polychronous groups*, that emitted regularly repeating temporal chains of spikes in response to the presentation of visual objects. This phenomenon is known as polychronization (Izhikevich et al., 2004). In particular, embedded within these polychronous groups were examples of what we called *binding neurons*, which had the crucial property that they responded if and only if a low-level feature neuron was actually participating in driving a high-level feature neuron. In this way, these neurons encoded the binding

relationships between lower and higher level visual features represented in different neuronal layers. Indeed, large numbers of binding neurons developed through successive layers of the hierarchical network simulated in Chapter 6, representing binding relationships across the visual field and at every spatial scale. In the current work, the same kind of spiking neural network architecture is used to simulate the development of border ownership cells that are able to maintain their proper firing properties when the network is presented with visual scenes containing multiple objects. The border ownership cells that develop in the simulations reported below are examples of the hypothesised binding neurons that developed in the simulations carried out in Chapter 6.

It is a key property of the model developed in Chapter 6 and implemented in this current chapter that visual objects are encoded in the input Layer 0 by spiking neurons with *randomised* Poisson distributions of spikes. That is, the visual stimuli are represented in the input layer by spike patterns with no regularly repeating temporal structure. Although the average firing rates of individual input neurons are set according to the outputs of Gabor filters that mimic the responses of simple cells in cortical visual area V1. Nevertheless, in Chapter 6, it was shown that, after an initial period of visually-guided learning with STDP, the network developed polychronous groups of neurons that fired their spikes in regularly repeating temporal patterns. In particular, in the study reported in Chapter 6, some of these polychronous groups were found to contain the hypothesised binding neurons. In our simulations reported below, these emergent polychronous groups are found to contain border ownership cells that are able to maintain their proper firing characteristics when more than one object is presented to the network at a time.

7.2.2 Mechanisms Underpinning the Development of V4-like Object Boundary Contour Element Cells in the Higher Network Layers

In Chapter 2 and Chapter 5, it has been shown how training the rate-coded VisNet model on many differently shaped objects leads to the development of neurons in higher layers that encode the boundary contour elements of objects in a similar manner to the V4 neurons reported by Pasupathy and Connor (2002). In particular, the responses of such neurons depend on where the preferred contour element occurs on an object boundary. For example, some higher layer neurons might learn to respond to the presence of a straight vertical edge on the left boundary of an object, while other neurons might learn to respond to the presence of a straight vertical edge on the right boundary of an object. In Chapter 2, I demonstrated how, even though the network is always trained on whole objects, neurons in the higher layers learn to represent localised object boundary contour elements (Eguchi et al., 2015). This is due to the statistical decoupling between differently shaped boundary contour elements over a sufficiently large set of visual objects used to train the network. In Chapter 2, a detailed analysis of the feedforward connectivity that had developed during visual training, which endowed the object boundary contour element neurons with their firing properties, was carried out. It was found, for example, that an object boundary contour element neuron in layer 3 that had learned to respond to a straight vertical edge on the right boundary of an object, in fact, had developed strengthened feedforward polysynaptic connections from a rich pattern of input Gabor filters that represented a main straight vertical edge as well as many surrounding edges of different orientations to the left of the straight vertical edge (see Fig. 2.9 in Chapter 2). The connections from the input Gabor filter representing the straight vertical edge were strongest, while the connections from the surrounding Gabor filters on the left were weaker and provided the local image context needed to ensure that the object boundary contour element cell only responded when the straight vertical edge was on the right of an object. So, to summarise, an individual neuron representing a particular kind of object boundary contour element, with firing properties similar to those observed by Pasupathy and Connor (2002), is driven in the models by a spatial constellation of input Gabor filters representing many local edges in the image, where the neuron receives its strongest connection from the input Gabor filter representing the edge corresponding to the

object boundary contour element and weaker connections from other surrounding Gabor filters representing local image context to ensure that the neuron fires only when the edge occurs at the correct position with respect to the centre of mass of the object.

In the modelling studies reported in Chapter 2 and Chapter 5, the neurons representing object boundary contour elements in the higher layers developed translation invariant responses by the use of a trace learning rule in the feedforward connections. Such a learning rule is able to drive the development of translation invariant neuronal responses by encouraging post-synaptic neurons to learn to respond to subsets of input patterns that tend to occur close together in time. In Chapter 2 and Chapter 5, it was demonstrated that if each of the objects is seen shifting across different retinal locations during visual training, perhaps due to a series of saccades, with different retinal views of the same object clustered together in time, then the trace learning rule will produce neurons that respond selectively to a particular boundary contour element with translation invariance, that is, no matter where the object is seen on the retina. However, the form of the trace learning used earlier utilised rate-coding. How might trace learning arise naturally in a more biologically realistic spiking neural network with STDP? This question was addressed by Evans and Stringer (2012), who showed that a trace learning effect can be achieved in a spiking network when the time constant governing the exponential decay of each synaptic conductance after an incoming pre-synaptic spike is sufficiently long. In this case, the conductance channels remain open for longer after each incoming spike, which in turn keeps the post-synaptic neuron firing longer. In this case, subsequent object views can become associated through STDP learning with the same active post-synaptic neuron. In this way, the post-synaptic neuron may learn to respond to a particular visual object over a number of different retinal locations. In the simulations described below, the same kind of trace learning mechanism as originally demonstrated by Evans and Stringer (2012) is implemented in order to drive the development of V4-like neurons in the higher layers that encode the conformation of boundary contour elements at particular positions with respect to the centre of mass of an object regardless of the location of the object on the retina as reported by Pasupathy and Connor (2002).

7.2.3 Mechanisms Underpinning the Development of V1/V2-like Border Ownership Cells in the Lower Network Layers

I hypothesised that the emergence of polychronous groups within the network, whereby groups of neurons learn to emit their action potentials in regularly repeating temporal sequences, during visually-guided training with STDP could also produce border ownership cells that maintain their proper firing characteristics when multiple visual objects are presented together to the network. Specifically, I hypothesised that border ownership cells would develop automatically within particular kinds of polychronous group during visual training, where the border ownership cells would become tuned through STDP learning to respond if and only if an edge-detecting V1-like simple cell in a lower layer is participating in driving a V4-like object boundary contour element cell in a higher layer, where the edge represented by the V1-like simple cell directly corresponds to the object boundary element represented by the object boundary contour element cell. In this case, the border ownership cell will carry measurable information that the edge represented by the edge-detecting V1-like simple cell is part of the object boundary element represented by the V4-like object boundary contour element cell. It should be noted that, in this scenario, border ownership cells are not examples of top-down modulated V1-like simple cells as modelled in Chapter 5 in a rate-coded network, but are instead a new category of neurons that is driven by converging inputs from V1-like simple cells in lower layers and object boundary contour element cells in higher layers (Eguchi and Stringer, 2016). In fact, as stated above, the border ownership cells that develop in the simulations reported below are examples of the hypothesised binding neurons that developed in the simulations carried out in Chapter 6. It is demonstrated in the simulations presented below that such border ownership cells develop

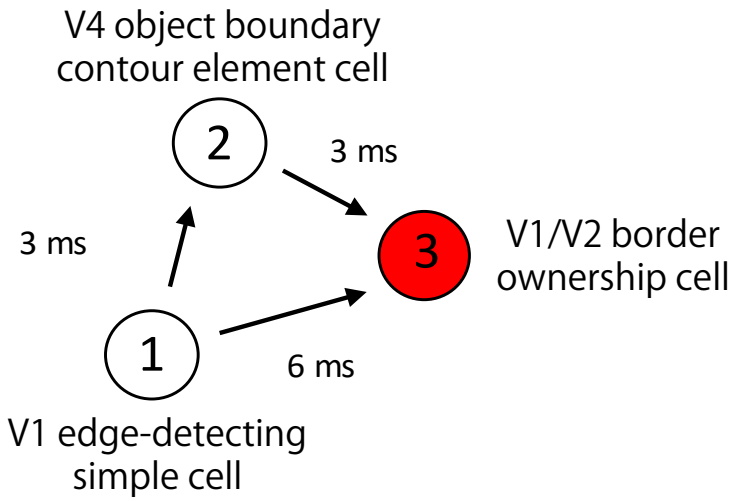

Figure 7.1: **Local network connectivity resulting in the emergence of a border ownership cell.** A local circuit of three linked neurons situated across different nearby stages of the primate ventral visual pathway is considered: neuron 1 is an edge-detecting simple cell in a lower visual layer such as V1, neuron 2 is an object boundary contour element cell in a higher layer such as V4, and neuron 3 is a border ownership neuron within a local layer such as V1 or V2. Assume that an image of an object is presented to the network, and that the edge represented by neuron 1 corresponds directly to the localised object boundary contour element represented by neuron 2. In the visual cortex, there are non-zero axonal transmission delays of around several milliseconds in the time taken for an action potential to pass from one neuron to another. In the circuit shown, the delay from neuron 1 to neuron 2 is 3ms, the delay from neuron 2 to neuron 3 is 3ms, and the delay from neuron 1 to neuron 3 is 6ms. Each neuron is assumed to operate as 'coincidence detector' in that it requires a volley of spikes from a number of sending neurons to arrive at about the same time in order for the neuron to reach its firing threshold. In our simulations, this is achieved by implementing fast neuronal and synaptic time constants, which ensures that the cell potential decays quickly between incoming spikes. So, it is assumed that neuron 3 will fire if and only if the incoming spikes from neurons 1 and 2 arrive at approximately the same time, that is, within a couple of milliseconds of each other. In this case, given the set of axonal delays shown in the figure, neuron 3 will fire if and only if neuron 1 is actually participating in driving neuron 2, because only then will the spikes from neurons 1 and 2 arrive together at neuron 3. In such a neuronal circuit, the firing of neuron 3 will encode the fact that the edge represented by neuron 1 is part of the object boundary contour element represented by neuron 2. The three neurons, 1, 2 and 3, form a polychronous group, in which the border ownership neuron 3 is embedded. It is hypothesised that large numbers of these polychronous groups with embedded border ownership neurons will emerge within the lower layers of the network, representing many different border ownership relationships across the visual field, during visually-guided learning with synaptic modification driven by STDP. Most importantly, these kinds of 3-neuron circuits, which each encode the binding relationship between a particular edge and the object boundary contour element of which the edge is a part, should not be susceptible to interference from the presence of other nearby objects within the visual field. In this way, the spiking dynamics described here are able to solve the central problem of how border ownership neurons are able to maintain their proper firing properties when the visual system is presented with scenes containing multiple different objects. This is essential to solving feature binding across complex visual scenes with multiple objects present.

automatically within the polychronous groups that emerge during visually-guided learning with STDP.

Figure 7.1 shows how border ownership cells can operate as members of polychronous groups that may emerge within a spiking neural network after visual training. A linked polychronous group of three neurons at different stages of the ventral visual pathway is considered: (i) neuron 1 is an edge-detecting simple cell in a lower visual layer such as V1, (ii) neuron 2 is an object boundary contour element cell in a higher layer such as V4, and (iii) neuron 3 is a hidden neuron within a local layer such as V1 or V2, which has the potential to operate as a border ownership neuron. Assume that there are the following three synaptic connections between these three neurons: (i) a connection from neuron 1 to neuron 2, (ii) a connection from neuron 1 to neuron 3. (This could be either a lateral or bottom-up connection depending on which layer neuron 3 is in), and (iii) a connection from neuron 2 to neuron 3. (This could be either a lateral or top-down connection depending on which layer neuron 3 is in.)

Let $\Delta_{(i,j)}$ denote the axonal transmission delay from a pre-synaptic neuron j to post-synaptic neuron i. Given this notation, it is evident that neuron 1 is participating in driving neuron 2 if and only if a spike emitted by neuron 2 occurs approximately $\Delta_{(2,1)}$ after a spike emitted by neuron 1.

If the axonal transmission delays between the three neurons 1, 2 and 3 shown in Fig. 7.1 have the following approximate relationship

$$\Delta_{(3,1)} \approx \Delta_{(2,1)} + \Delta_{(3,2)} \tag{7.1}$$

then the spikes from neurons 1 and 2 will converge on neuron 3 at approximately the same time if and only if neuron 1 is actually participating in driving neuron 2.

It is assumed that neurons in the network operate as 'coincidence detectors', in that they only fire when a volley of incoming spikes from pre-synaptic neurons arrive close together in time. This can be effected in the spiking network simulations by implementing fast neuronal and synaptic time constants, which allows a cell potential to decay rapidly between incoming spikes. Thus, neuron 3 is assumed to only fire when the volley of spikes from neurons 1 and 2 arrive near simultaneously. In this case, neuron 3 will behave as a border ownership neuron. That is, neuron 3 will fire if and only if neuron 1 is participating in driving neuron 2, which indicates that the edge represented by neuron 1 is actually part of the object boundary contour element represented by neuron 2. Moreover, in this case, STDP will further strengthen the connections from neurons 1 and 2 onto the border ownership neuron 3, thus further enhancing the effect.

7.2.4 Responses of Spiking Border Ownership Cells to Visual Scenes with Multiple Objects

The key property of the border ownership cells illustrated in Fig. 7.1 is that an individual cell responds if and only if the lower layer neuron representing the edge is actually participating in driving the higher layer neuron representing the corresponding object boundary contour element. Only in this case will the border ownership cell be fully informative that the edge is part of the object boundary contour element. The border ownership neuron should not respond if the lower layer neuron representing the edge and the higher layer neuron representing the object boundary contour element simply happen to be co-active, where the former is not actually driving the latter. Such unrelated co-activation of a lower layer edge-detecting cell and higher layer object boundary contour element cell might occur, for example, because of the presence of multiple objects within a natural scene. However, our proposed mechanism for generating border ownership cells, as illustrated in Fig. 7.1, ensures that these cells only become activated if the lower layer edge-detecting cell is actually participating in driving the higher layer object boundary contour element cell. This kind of temporally specific response is characteristic of a polychronous group, which the three neurons 1, 2 and 3 described above comprise. This mechanism potentially solves the limitations of rate-coding described in section 7.1.3. Indeed, for these reasons, I propose that modelling the development of border ownership cells, and more generally feature binding, may not be soluble within a traditional rate-coded network, but will instead require the full spiking neuronal dynamics of the brain.

In the simulations presented below, the emergence of such border ownership neurons during visually-guided training is looked at. I expected to find evidence for the kind of 3 neuron polychronous groups described above and illustrated in Figure 7.1.

7.3 Materials & Methods

7.3.1 Spiking Neural Network Model

7.3.1.1 Network Architecture

The neural network model investigated is shown in Figure 6.5 and simulates successive neuronal stages of processing along the primate ventral visual pathway. Specifically, the model is comprised of four layers of excitatory pyramidal neurons, which may be loosely thought of as

representing cortical visual areas V2, V4, posterior inferior temporal cortex (TEO) and anterior inferior temporal cortex (TE). There are modifiable bottom-up (feedforward) and top-down (feedback) synaptic connections between excitatory pyramidal neurons in successive layers, as well as modifiable lateral synapses between excitatory pyramidal neurons within each layer. Within each layer, there are also inhibitory interneurons with non-plastic lateral synaptic connections to and from the excitatory neurons to produce competition between the excitatory neurons. For all presented simulations, $64 \times 64 = 4096$ excitatory neurons and $32 \times 32 = 1024$ inhibitory neurons in each layer are used, with a fixed number of sparsely distributed topologically organised connections. Table 7.1a shows the different numbers of afferent connections onto each postsynaptic neuron, as well as the fan-in radius of these connections, for the different types of excitatory-excitatory, excitatory-inhibitory and inhibitory-excitatory connections between and within the four neuronal layers. The detailed descriptions of the model is provided in 6.3.1.

7.3.2 Network Performance Measures

7.3.2.1 Information Analysis of Average Firing Rate Responses of Single Cells

Information theory is used to quantify how selective the average firing rate responses of individual neurons are for members of a particular stimulus category. If a neuron responds invariantly to the members of a particular stimulus category but not to members of other stimulus categories, then the neuron carries a maximum amount of information about the presence of its preferred stimulus category.

Information theory was used to quantify the performance of single neurons tasked with learning a translation invariant response (across multiple retinal locations) to specific visual stimuli in Chapter 2 (Eguchi et al., 2015). If the responses r of a neuron carry a high-level of information about the presence of a particular stimulus s across different transforms such as changes in retinal location or orientation, then this implies that the neuron will respond selectively to the presence of that stimulus regardless of where the stimulus is presented on the retina or its orientation with respect to the observer. A detailed description of the analysis is provided in Sec 1.5.2.

7.3.2.2 Finding Polychronous Groups and Binding Neurons

A key diagnostic in the simulations reported below is to identify the polychronous groups that have emerged in the network after visually-guided training so that any tuple of cells that shows the binding relationship can be looked for. The technique used is described in Section 6.3.2.3.

7.4 Simulation Studies

7.4.1 Development of V4-like object boundary contour element cells and V1/V2-like border ownership cells

In this first simulation study, the visually-guided development of V4-like object boundary contour element cells in Layer 4 and V1/V2-like border ownership cells in Layer 1 as the network is trained on a number of different visual objects is investigated.

In this simulation study, the network was trained and tested on the two abstract visual objects shown in Figure 7.2. The two objects were a semicircle with a straight vertical edge on its right boundary (Fig. 7.2(a)) and a semicircle with a straight vertical edge on its left boundary (Fig. 7.2(b)). However, each of these two objects was seen in the four different transforms shown in each row of Figure 7.2. Firstly, the objects were black when presented on a light grey background or light grey when presented on a black background. Secondly, each object was presented in two locations on the left and right of the retina. Whenever an object

Table 7.1: Model parameters. Most integrate and fire parameters were taken from Troyer et al. (1998) (derived originally from McCormick et al. (1985) as indicated by §. Plasticity parameters (denoted by †) are taken from Perrinet et al. (2001). Parameters marked with * were tuned for the reported simulations.

(a) Network parameters					
Layer	**1**	**2**	**3**	**4**	
Number of excit. neurons within each layer	64×64	64×64	64×64	64×64	
Number of inhib. neurons within each layer	32×32	32×32	32×32	32×32	
Number of feedforward (FF) afferent excit. connections per excit. neuron (EfE)	30	100	100	100	
Fan-in radius for FF afferent excit. connections to each excit. neuron (EfE)	12.0	12.0	18.0	24.0	
Number of feedback (FB) afferent excit. connections per excit. neuron (EbE)	5	5	5	-	
Fan-in radius for FB afferent excit. connections to each excit. neuron (EbE)	12.0	12.0	12.0	-	
Number of lateral (LAT) afferent excit. connections per excit. neuron (ElE)	30	30	30	30	
Fan-in radius for LAT afferent excit. connections to each excit. neuron (ElE)	8.0	8.0	8.0	8.0	
Number of LAT afferent excit. connections per inhib. neuron (ElI)	30	30	30	30	
Fan-in radius for LAT afferent excit. connections to each inhib. neuron (ElI)	1.0	1.0	1.0	1.0	
Number of LAT afferent inhib. connections per excit. neuron (IlE)	30	30	30	30	
Fan-in radius for LAT afferent inhib. connections to each excit. neuron (IlE)	8.0	8.0	8.0	8.0	
(b) Parameters for Gabor Filtering of visual images					
Phase shift (ψ)	$0, \pi$				
Wavelength (λ)	2				
Orientation (θ)	$0, \pi/4, \pi/2, 3\pi/4$				
Spatial bandwidth (b)	1.5 octaves				
Aspect ratio (γ)	0.5				
(c) Cellular Parameters					
Excit. cell somatic capacitance (C_m^E) and Inhib. cell somatic capacitance (C_m^I)	500 pF, 214 pF				§
Excit. cell somatic leakage conductance (g_0^E) and Inhib. cell somatic leakage conductance (g_0^I)	25 nS, 18 nS				§
Excit. cell membrane time constant (τ_m^E) and Inhib. cell membrane time constant (τ_m^I)	20 ms, 12 ms				§
Excit. cell resting potential (V_0^E) and Inhib. cell resting potential (V_0^I)	-74 mV, -82 mV				§
Excit. firing threshold potential (Θ^E) and Inhib. firing threshold potential (Θ^I)	-53 mv, -53 mV				§
Excit. after-spike hyperpolarization potential (V_H^E) and Inhib. after-spike hyperpolarization potential (V_H^I)	-57 mV, -58 mV				§
Absolute refractory period (τ_R)	2 ms				§
(d) Synaptic Parameters					
Synaptic neurotransmitter concentration (α_C) and Proportion of unblocked NMDA receptors (α_D)	0.5				†
Presynaptic STDP time constant (τ_C) and Postsynaptic STDP time constant (τ_D)	5 ms				†
Synaptic learning rate (ρ)	0.1				†
Range of synaptic conductance delays	[0.1, 10.0] ms				†
Synaptic conductance scaling factor for FF excitatory connections from Gabor filters to Layer 1 excit. cells ($\lambda^{GfE} \cdot \Delta g^{GfE}$)	[0, 2] nS				*
Synaptic conductance scaling factor for FF excit. connections to excit. cells in layers 2, 3 or 4 ($\lambda^{EfE} \cdot \Delta g^{EfE}$)	[0, 1] nS				*
Synaptic conductance scaling factor for FB excit. connections to excit. cells in layers 1, 2 or 3 ($\lambda^{EbE} \cdot \Delta g^{EbE}$)	[0, 2] nS				*
Synaptic conductance scaling factor for LAT excit. connections to excit. cells in layers 1, 2, 3 or 4 ($\lambda^{ElE} \cdot \Delta g^{ElE}$)	[0, 2] nS				*
Synaptic conductance scaling factor for LAT connections from excit. cells to inhib. cells in layers 1, 2, 3 or 4 ($\lambda^{ElI} \cdot \Delta g^{ElI}$)	1 mS				*
Synaptic conductance scaling factor for LAT connections from inhib. cells to excit. cells in layers 1, 2, 3 or 4 ($\lambda^{IlE} \cdot \Delta g^{IlE}$)	25 mS				*
Excitatory reversal potential ($\hat{V}^E$)	0 mV				§
Inhibitory reversal potential ($\hat{V}^I$)	-70 mV				§
Synaptic time constant for all FF, FB, and LAT connections from Gabor filters and excit. cells to excit. cells ($\tau_{GfE}, \tau_{EfE}, \tau_{EbE}, \tau_{ElE}$)	150 ms				*
Synaptic time constant for LAT connections from excit. cells to inhib. cells (τ_{ElI})	2 ms				§
Synaptic time constant for LAT connections from inhib. cells to excit. cells (τ_{IlE})	5 ms				§
(e) Parameters for numerical simulation by Forward Euler timestepping scheme					
Numerical step size (Δt)	0.02 ms				

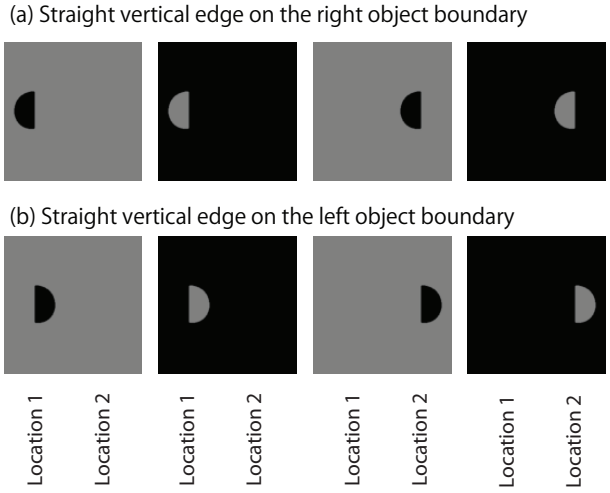

Figure 7.2: The visual object stimuli used for the simulation study of section 7.4.1, in which the network is trained and tested on scenes containing only a single object at a time. There are two object shapes used to both train and test the network: a semicircle with a straight vertical edge on its right boundary (a) and a semicircle with a straight vertical edge on its left boundary (b). However, each of these two objects was seen in the four different transforms shown in each row. Firstly, the objects were black when presented on a light grey background or light grey when presented on a black background. Secondly, each object was presented in two locations on the left and right of the retina. Whenever an object was presented on the left of the retina, the vertical straight edge on its (left or right) boundary was precisely aligned with retinal Location 1. Similarly, whenever the object was presented on the right of the retina, the vertical straight edge on its (left or right) boundary was aligned with retinal Location 2. In the simulation study reported in section 7.4.1, the network was presented with input images containing only a single object stimulus at a time during both training and testing.

was presented on the left of the retina, the vertical straight edge on its (left or right) boundary was precisely aligned with retinal Location 1. Similarly, whenever the object was presented on the right of the retina, the vertical straight edge on its (left or right) boundary was aligned with retinal Location 2.

In the simulation reported in this section, the network was trained and tested on visual scenes containing only a single object at a time.

Each simulation begins with an initial period of visual training. During each training epoch, each of the two objects shown in Figure 7.2 is presented in turn to the network in a randomised series of different transforms (i.e. black or light grey shading, and left or right retinal locations) a total of 16 times. The duration of each stimulus presentation is 200 ms. It is important that first one object is presented 16 times and then the other object is presented 16 times in order to ensure that the different transforms of each object are seen clustered together in time. This is necessary for trace learning to be able to drive the development of translation invariant object boundary contour element neurons in the highest layer of the network, as shown in the previous rate-coded simulations reported in Chapter 5 and spiking network simulations of Evans and Stringer (2012). Also remember that each different presentation of the same training image, corresponding to a particular object with a particular shading and retinal position, may also be regarded as a different transform because such identical images will in fact be represented by different randomised spatiotemporal spike patterns in the input Layer 0, as discussed above in section 7.2. Given this training regime, the total duration of one training epoch is 2 (vertical straight edge on the left or right object boundary) $\times$ 16 $\times$ 200 ms. In this manner, the network is trained for a total of 20 training epochs.

After training is completed, the *steady state* firing responses of neurons in Layer 1 and Layer 4 to the same stimuli that were used to train the network as shown in Figure 7.2 are analysed. Specifically, during testing, each stimulus is initially presented for one second in order to allow enough time for visual signals to propagate up and back down the layers, and the neurons throughout the network to settle into steady state firing rates. After this initial stimulus presentation period of one second had passed, the stimulus is then continued to be

presented for a further one second during which we recorded the number of spikes emitted by each of the neurons throughout the network. Then the average steady state firing rate response of each neuron to each stimulus is computed over this further one second period of stimulus presentation. In other words, although the network has spiking dynamics, which I hypothesize are critical to the development of border ownership cells within the network, the neuronal responses are still analysed in a rate-coded manner. This is because such a rate-coded analysis is sufficient to demonstrate the firing characteristics of border ownership cells, even if the underlying dynamics required for the operation of the network as described above in section 7.2 are spiking. The network is tested in this manner both before and after visual training in order to assess the effect that training has on the response properties of neurons in Layer 1 and Layer 2, and in particular, the development of object boundary contour element cells in Layer 4 and border ownership cells in Layer 1.

The firing properties of the output (Layer 4) neurons are firs tested to investigate whether these neurons had learned after training to respond like the object boundary contour element neurons reported by Pasupathy and Connor (2002). That is, did these neurons learn to respond selectively to the presence of a vertical straight edge on either the left boundary or right boundary of an object, and do so regardless of the position of the object on the retina (i.e. retinal Location 1 or Location 2) or the shading of the object (i.e. black or light grey)? The same set of stimuli used to train the network shown in Figure 7.2 was presented to the network during testing, and the firing rate of each neuron in the output layer of the network was recorded.

In order to quantify the performance of the output layer neurons, information analysis was conducted as described in Section 7.3.2.1. In the analysis of Layer 4 neurons, there are two different stimulus categories ($n = 2$) corresponding to stimuli with a vertical straight edge on either the left object boundary or right object boundary, respectively. In Figure 7.2, stimuli from the first category with a vertical straight edge on the left object boundary are shown in row (b), while stimuli from the second category with a vertical straight edge on the right are shown in row (a). Since each category member was defined by its shading (black or light grey) and retinal location (Location 1 or Location 2), there were $2 \times 2 = 4$ members (transforms) of each of the two stimulus categories. In order for a Layer 4 neuron to successfully mimic the firing characteristics of an object boundary contour element neuron, it had to respond invariantly over the four members (transforms) of its preferred stimulus category, and not respond to any members of the other stimulus category. In this case the neuron would carry maximum information about its preferred stimulus category.

Figure 7.3(a) shows the information analysis of the steady state response properties of Layer 4 neurons at the end of each stimulus presentation. Results are presented before and after training. The maximum amount of information possible for the simulation is $log_2(n)$ where n is the number of stimulus categories = 2, that is 1 bit. Before training, only one neuron reached 1 bit of information and in fact most neurons carried much less than 1 bit. However, after training, around 40 neurons carried the maximum 1 bit of information. This result confirms that some neurons in Layer 4 had successfully learned to respond selectively to a vertical straight edge either on the left or on the right of an object boundary, regardless of the retinal location of the object or shading of the object. These neurons successfully replicated the observed firing characteristics of the object boundary contour element cells described by Pasupathy and Connor (2002).

Figure 7.3(b) shows the steady state firing rate responses of two typical Layer 4 neurons (15552) and (15665) after each stimulus had been presented for one second to allow the neuronal firing rates to settle to stable values as described above. The firing rate responses are plotted before and after training. Specifically, the plot shows the responses of the two neurons to all transforms of the object with a vertical straight edge on its right boundary (1-4) and the object with a vertical straight edge on its left boundary (5-8). These results show that after training neuron (15552) learned to respond selectively to all transforms of the object with a

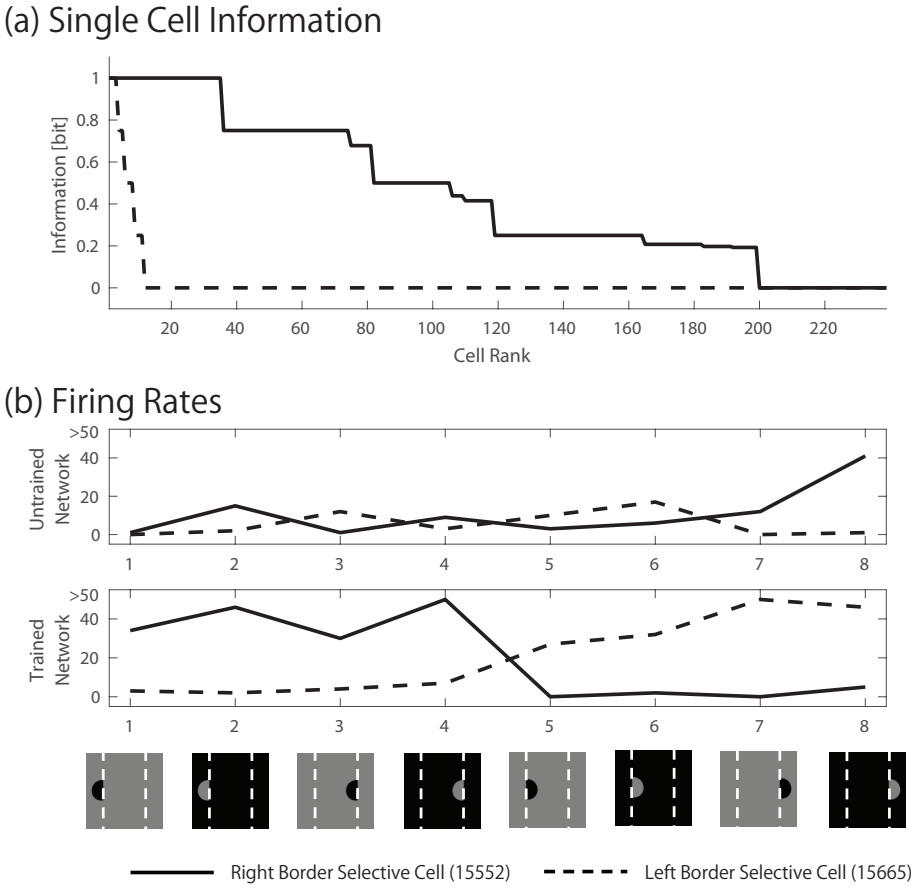

Figure 7.3: The development of V4-like object boundary contour cells in output Layer 4 of the network model after visually-guided training on the set of objects shown in Figure 7.2). The steady state firing rate responses of Layer 4 neurons when the network is tested on the same set of visual objects are analysed. That is, the average firing rate response of each neuron is computed by summing the total number of spikes emitted by that neuron in response to each visual object, and dividing this by the duration of the stimulus presentation. In other words, although the network has spiking dynamics, the neuronal responses are still analysed in a rate-coded manner. **(a) Information analysis:** The information carried by the output (4th layer) neurons about whether the vertical straight edge was on the left or right boundary of each object transform presented to the network before and after training, is computed. Since there are $n = 2$ stimulus categories, perfectly discriminating Layer 4 neurons carry a maximum of 1 bit of information. The plot shows the maximum single cell information carried by each of the top 240 neurons in Layer 4 about which one of these two stimulus categories was presented, where the neurons in Layer 4 are plotted along the abscissa in rank order. The results show that training led to a large increase in the number of Layer 4 neurons that responded selectively to a vertical straight edge either on the left or on the right of an object boundary, regardless of the shading or retinal location of the object. Indeed, after training, about 40 Layer 4 neurons reached the maximum level of single cell information of 1 bit. **(b) Firing rate responses of two example Layer 4 neurons that had maximum single cell information:** the plot shows the responses of two Layer 4 neurons to all transforms of the object with a vertical straight edge on its right boundary (1-4) and the object with a vertical straight edge on its left boundary (5-8). These results show that after training neuron (15552) learned to respond selectively to all transforms of the object with a vertical straight edge on the right, while neuron (15665) learned to respond to all transforms of the object with a vertical straight edge on the left. These two neurons thus display the characteristic firing properties of the V4-like object boundary contour element cells reported by Pasupathy and Connor (2002), in that they represent a particular boundary contour element conformation (i.e. vertical straight edge) at a particular position on the object boundary (i.e. left or right of object boundary) regardless of the retinal location of the object (i.e. on the left or right of the retina).

vertical straight edge on the right, while neuron (15665) learned to respond to all transforms of the object with a vertical straight edge on the left. These two neurons thus display the characteristic firing properties of the V4-like object boundary contour element cells reported by Pasupathy and Connor (2002), in that they represent a particular boundary contour element conformation (i.e. vertical straight edge) at a particular position on the object boundary (i.e. left or right of object boundary) regardless of the retinal location of the object (i.e. on the left or right of the retina). These kinds of firing responses were simulated in the rate-coded

models used in Chapter 2 and Chapter 5. As described in section 7.2, V4-like object boundary contour element cells need to develop in the higher layers of our model in order to provide appropriate top-down signals that drive the development of border ownership neurons in lower layers corresponding to cortical areas V1/V2.

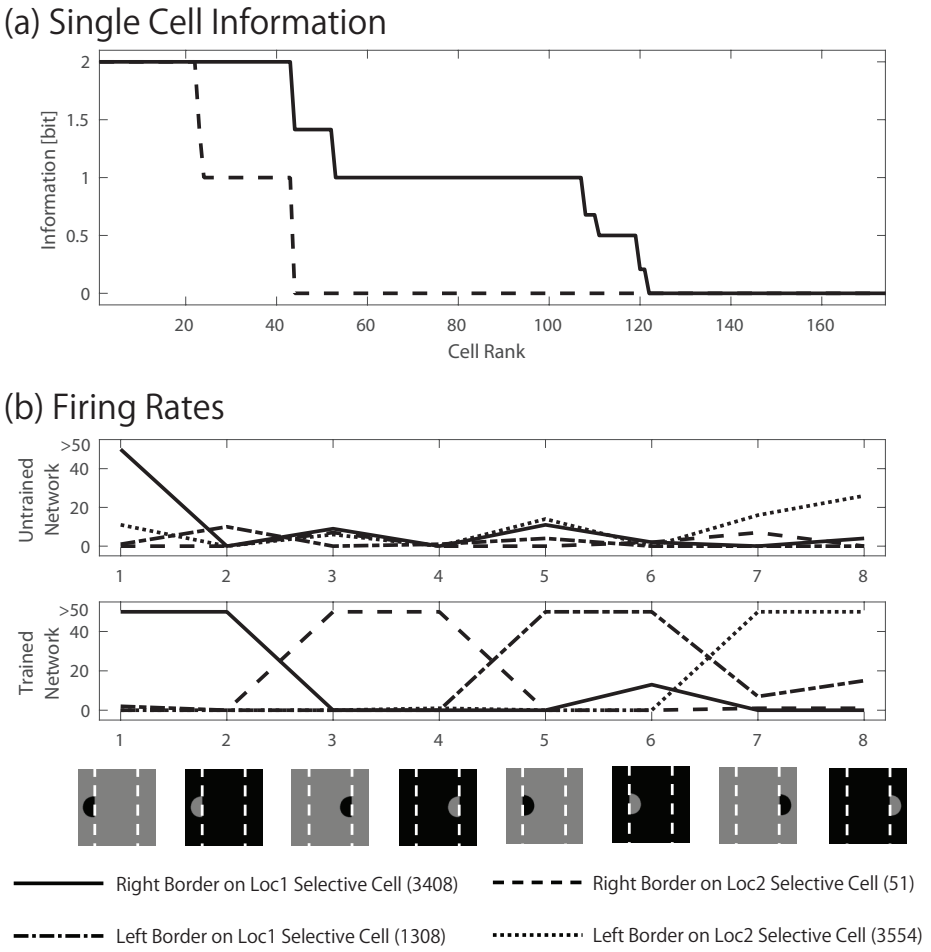

Figure 7.4: The development of V1/V2-like border ownership cells in Layer 1 of the network model after visually-guided training on the set of objects shown in Figure 7.2). The steady state firing rate responses of Layer 1 neurons when the network is tested on the same set of visual objects are analysed. **(a) Information analysis:** I investigated whether Layer 1 neurons learned to respond selectively to one of the following four stimulus categories: (i) the object has a vertical straight edge on its left boundary with the vertical straight edge positioned at retinal Location 1, (ii) the object has a vertical straight edge on its right boundary with the vertical straight edge positioned at retinal Location 1, (iii) the object has a vertical straight edge on its left boundary with the vertical straight edge positioned at retinal Location 2, and (iv) the object has a vertical straight edge on its right boundary with the vertical straight edge positioned at retinal Location 2. The plot shows the maximum single cell information carried by each of the top 170 neurons in Layer 1 about which one of these four stimulus categories was presented, where the neurons in Layer 1 are plotted along the abscissa in rank order. The results show that training led to a large increase in the number of Layer 1 neurons that responded selectively to a vertical straight edge either on the left or on the right of an object boundary, with the object presented in a particular retinal position, regardless of the shading of the object. Indeed, after training, about 40 Layer 1 neurons reached the maximum level of single cell information of 2 bits. These Layer 1 neurons have learned to respond with perfect selectivity to one of the four stimulus categories. **(b) The firing rate responses of four Layer 1 neurons with maximum single cell information:** the plot shows the responses of four Layer 1 neurons to all transforms of the object with a vertical straight edge on its right boundary (1-4) and the object with a vertical straight edge on its left boundary (5-8). In particular, the first two visual stimuli 1-2 shown along the abscissa have the object with a vertical straight edge on its right boundary presented in retinal Location 1, the next two stimuli 3-4 have the object with a vertical straight edge on its right boundary presented in retinal Location 2, the next two stimuli 5-6 have the object with a vertical straight edge on its left boundary presented in retinal Location 1, and the last two stimuli 7-8 have the object with a vertical straight edge on its left boundary presented in retinal Location 2. The results show that, after training, each of the four neurons had learned to respond selectively to one of the four stimulus categories. These four neurons thus display the characteristic firing properties of the V1/V2-like border ownership neurons reported by Zhou et al. (2000), in that they respond selectively to a vertical straight edge on either the left or right of an object when the object is presented in a particular retinal position.

I next tested whether Layer 1 neurons had developed the kind of border ownership representations reported by Zhou et al. (2000). In other words, I tested whether the feedback (top-down) connections implemented in the network, which carried top-down signals from object boundary contour element neurons in Layer 4 through consecutive lower layers, enabled the development of neurons in Layer 1 (corresponding to visual areas V1/V2) that responded to straight vertical edges in either retinal Location 1 or 2 but which also responded selectively depending on whether the straight vertical edge was on either the left or right boundary of the object.

In order to quantify the performance of Layer 1 neurons, the information carried by the steady state responses of these cells at the end of each 2 second stimulus presentation was computed as described above. The results of this analysis are presented in Figure 7.4(a), where the information carried by Layer 1 neurons before and after training is shown. The border ownership neurons reported by (Pasupathy and Connor, 2002) occur in early cortical stages of visual processing, and so they do not display much translation invariance across different retinal positions. Similarly, Layer 1 neurons in our model are not expected to develop translation invariance across different retinal locations due to the small fan-in of connections from the retina. Therefore, information that was specific to either retinal Location 1 or Location 2 was computed. Specifically, the analysis calculated the information carried by the Layer 1 neurons about whether the vertical straight edge in the object stimulus presented to the network was an example from one of four stimulus categories: (i) the object has a vertical straight edge on its left boundary with the vertical straight edge positioned at retinal Location 1, (ii) the object has a vertical straight edge on its right boundary with the vertical straight edge positioned at retinal Location 1, (iii) the object has a vertical straight edge on its left boundary with the vertical straight edge positioned at retinal Location 2, and (iv) the object has a vertical straight edge on its right boundary with the vertical straight edge positioned at retinal Location 2. Since there are $n = 4$ stimulus categories, perfectly discriminating neurons carry a maximum of $log_2(n) = 2$ bits of information.

Figure 7.4(a) shows the single cell information analysis. The plot shows the maximum information carried by each of the top 170 neurons in Layer 1 about which one of the four stimulus categories was presented. It can be seen that training the network has led to double the number of neurons carrying the maximum 2 bits of information. After training, around 40 cells learned to carry the maximum single cell information, which implies that these cells have learned to respond with perfect selectivity to one of the four stimulus categories. These Layer 1 neurons thus provide the kind of border ownership representations experimentally observed in cortical visual areas V1 and V2 by Zhou et al. (2000).

Figure 7.4(b) shows the steady state firing rate responses of four typical Layer 1 neurons at the end of each stimulus presentation for 2 seconds. Results are compared before and after training. The plot shows the responses of four Layer 1 neurons to all transforms of the object with a vertical straight edge on its right boundary (1-4) and the object with a vertical straight edge on its left boundary (5-8). The results show that, after training, each of the four neurons had learned to respond selectively to one of the four stimulus categories. Specifically, neuron (3408) responded when the object had a vertical straight edge on its right boundary with the vertical straight edge positioned at retinal Location 1, neuron (51) responded when the object had a vertical straight edge on its right boundary with the vertical straight edge positioned at retinal Location 2, neuron (1308) responded when the object had a vertical straight edge on its left boundary with the vertical straight edge positioned at retinal Location 1, and neuron (3554) responded when the object had a vertical straight edge on its left boundary with the vertical straight edge positioned at retinal Location 2. These four neurons thus display the characteristic firing properties of the V1/V2-like border ownership neurons reported by Zhou et al. (2000), in that they respond selectively to a vertical straight edge on either the left or right of an object when the object is presented in a particular retinal position.

7.4.2 Maintenance of border ownership representations when the network is tested with multiple visual objects simultaneously

In the above simulations, the model was tested by presenting a single object to the network at a time. However, the primate visual system is usually presented with multiple objects simultaneously in real world scenes. In section 7.1.3, it was discussed how this more realistic situation exposed a weakness in the previous rate-coded model of border ownership neurons in Chapter 5 (Eguchi and Stringer, 2016), which failed to maintain border ownership representations when multiple objects were seen together simultaneously. It is in this situation that binding between lower-level and higher-level features becomes a more difficult problem (von der Malsburg, 1999). In particular, in a rate-coded network the top-down modulation of V1/V2 neuronal firing rates is not specific to retinal location. This effectively destroys the local border ownership (binding) information carried by the V1/V2 neurons. However, in section 7.2, I proposed how this problem may be solved in a spiking neural network, in which the border ownership cells are examples of the binding neurons hypothesised in Chapter 6. Next, this proposal is explored by testing the spiking network model on multiple objects simultaneously.

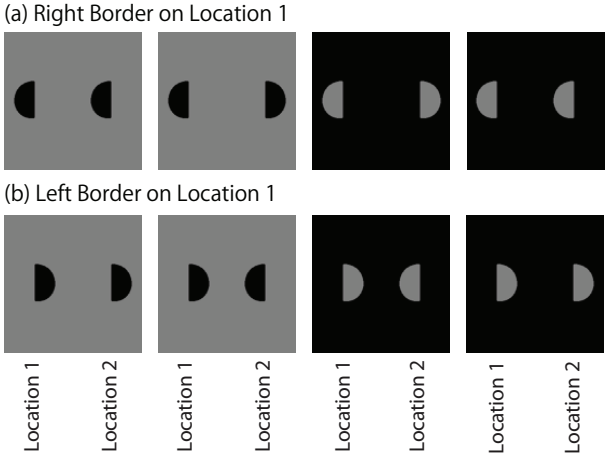

Figure 7.5: The set of visual stimuli used in section 7.4.2 to test the performance of the network when two objects are presented simultaneously during testing. There are two new categories of visual stimuli. The first stimulus category (a) consists of all possible combinations two objects where one of the objects has a vertical straight edge on its *right* boundary which is positioned at retinal Location 1. The second stimulus category (b) consists of all possible combinations two objects where one of the objects has a vertical straight edge on its *left* boundary which is positioned at retinal Location 1. These two stimulus categories were chosen in order to test whether the border ownership cells in Layer 1, which had been found in section 7.4.1 to respond selectively to a vertical straight edge on either the left or right object boundary in retinal Location 1, were able to maintain their response selectivity when an additional object was simultaneously presented with its vertical straight edge in retinal Location 2.

In this section, the model was trained with the set of objects shown in Figure 7.2, where these objects were presented to the network one at a time during training as described in the simulations above. However, the network was then tested with *two* objects shown together during each visual presentation, where the set of test images shown in Figure 7.5 is used. The steady state firing responses of neurons in each layer of the network at the end of each such visual presentation are analysed. These results were compared with those in which only a single object was presented to the network at a time during testing. In order to facilitate comparison of the results for the two test situations, in each case I analysed how much information neurons carried about the two border ownership stimulus categories, i.e. whether a straight vertical edge was present on either the left or right object boundary, that were associated with retinal Location 1.

The set of images used for testing the network with two objects at a time are shown in Figure 7.5. There are two different stimulus categories. The first stimulus category, shown in Figure 7.5(a), consists of all possible combinations two objects where one of the objects has a vertical

straight edge on its *right* boundary which is positioned at retinal Location 1. The second stimulus category, shown in Figure 7.5(b), consists of all possible combinations two objects where one of the objects has a vertical straight edge on its *left* boundary which is positioned at retinal Location 1. Each of the two stimulus categories undergoes 4 transforms due to variations in the following two stimulus features: 2 sides of an object (left or right) on which a vertical straight edge may occur at retinal Location 2 $\times$ 2 kinds of shading contrast between objects and background. These two stimulus categories were chosen in order to test whether the border ownership cells in Layer 1, which had been found in section 7.4.1 to respond selectively to a vertical straight edge on either the left or right object boundary in retinal Location 1, were able to maintain their response selectivity when an additional object was simultaneously presented with its vertical straight edge in retinal Location 2.

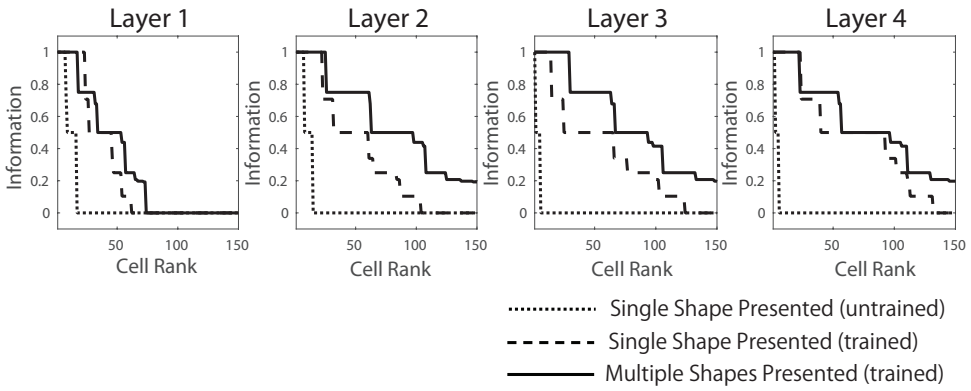

Figure 7.6: A quantitative comparison of the border ownership information carried by neurons in each layer when the network is tested with objects shown individually (untrained - dotted line; trained - dashed line) or when tested on two objects presented together (solid line). The network performance is assessed using single-cell information analysis. The information analysis is applied to the steady state firing responses of neurons in each layer at the end of each stimulus presentation for 2 seconds. In both test situations, the model was initially trained with the individual objects shown in Figure 7.2, as described in section 7.4.1. **Solid lines:** the performance of the model when tested with *two* objects shown together during each visual presentation using the test images shown in Figure 7.5. Here the single cell information carried by each neuron about the two stimulus categories described in Figure 7.5 was computed. Since there are two stimulus categories, neurons may carry up to a maximum of 1 bit of information. **Dashed lines:** the performance of the model when tested with the objects shown in Figure 7.2 presented one at a time during testing. The single cell information carried by each neuron about the four stimulus categories described in section 7.4.1 was computed. In this figure the maximum single-cell information ($log_2(4) = 2$) has been rescaled to 1. The information carried by neurons in each layer about stimulus categories (i) and (ii) is shown in each plot. It can be seen that even when multiple object shapes are presented to the network simultaneously, the border ownership selective responses of cells developed in each layer remain comparable to when objects are presented individually. This behaviour in a spiking network is in sharp contrast to the former rate-coded simulation results reported in Chapter 5 (Eguchi and Stringer, 2016), which showed a large drop in border ownership selectivity when multiple objects were presented to the network together during testing.

Figure 7.6 compares the border ownership information carried by neurons in each layer when the network is tested with objects shown individually (dashed line) or when tested on two objects presented together (solid line). The performance of the network was assessed using single-cell information analysis. The information analysis is applied to the steady state firing responses of neurons in each layer at the end of each stimulus presentation.

The dashed lines in Figure 7.6 show the performance of the model when tested with the objects shown in Figure 7.2 presented one at a time during testing. The single cell information carried by each neuron about the four stimulus categories previously described in section 7.4.1 was computed. That is, the information about whether the vertical straight edge in the object stimulus was from one of the following four stimulus categories: (i) the object has a vertical straight edge on its left boundary with the vertical straight edge positioned at retinal Location 1, (ii) the object has a vertical straight edge on its right boundary with the vertical straight edge positioned at retinal Location 1, (iii) the object has a vertical straight edge on its left boundary with the vertical straight edge positioned at retinal Location 2, and (iv) the object

has a vertical straight edge on its right boundary with the vertical straight edge positioned at retinal Location 2. Since there are four stimulus categories, perfectly discriminating neurons carry a maximum of 2 bits of information. However, in this figure the maximum single-cell information has been rescaled to 1. Results for the two stimulus categories (i) and (ii), which are associated with retinal Location 1 are plotted. The information carried by neurons in each layer about stimulus category (i) and (ii) is shown in the plot.

The solid lines in Figure 7.6 show the performance of the model when tested with *two* objects shown together during each visual presentation using the test images shown in Figure 7.5. Here the single cell information carried by the neurons about the two stimulus categories described in Figure 7.5, which are both associated with retinal Location 1, was computed. The first stimulus category includes all combinations of two objects where one of the objects has a vertical straight edge on its *right* boundary which is positioned at retinal Location 1, while the second stimulus category includes all combinations of two objects where one of the objects has a vertical straight edge on its *left* boundary positioned at retinal Location 1. Since there are two stimulus categories, neurons may carry up to a maximum of 1 bit of information. It can be seen that even when the network is tested on two objects at a time, the border ownership selective responses of cells developed in each layer remain comparable to when objects are presented individually. This result supports our prediction that border ownership information carried by Layer 1 (V1/V2) neurons in the spiking model may remain even when the network is presented with multiple visual objects during testing. This behaviour in a spiking network is in sharp contrast to the former rate-coded simulation results reported in Chapter 5, which showed a large drop in border ownership selectivity when multiple objects were presented to the network together during testing. It therefore appears that the spiking dynamics investigated in this chapter may be necessary to model the behaviour of border ownership neurons in the more challenging situation in which multiple objects are seen together simultaneously.

Interestingly, the simulation results shown in Figure 7.6 reveal the strong presence of border ownership representations in higher layers of the network, including the output Layer 4. There were no border ownership neurons with maximum single cell information before visual training as shown in Figure 7.6 with dotted lines. However, from Figure 7.6 it can be seen that there was a large number of perfectly tuned border ownership cells in Layer 4 after training. Moreover, these border ownership neurons maintained their response selectivity even when multiple objects were presented to the network together. I propose that the development of border ownership neurons in our spiking network simulations may be an example of the *holographic principle* originally proposed in Chapter 6, in which information about low level features is projected upwards through the layers.

In Chapter 6, I suggested that the bottom-up projection of visual information about lower-level visual features such as object border edges to the higher layers of visual processing might be important if the visual information used to guide behaviour is only read out from the later stages of the visual system. For example, it is usually thought that information about low-level visual features such as oriented bars and object edges is represented in early cortical layers such as V1 and V2. While more complex visual stimuli such as whole objects is represented in later stages of visual processing. However, when we observe a visual scene we are aware of visual features at every spatial scale and level of complexity, including object edges. Therefore, if information is only accessible from the highest stages of visual processing, then somehow this low-level information must be projected upwards through successive visual layers. Moreover, this feedforward projection of low-level visual information must be dependent on local image context, for example, representing the fact that an edge is part of a particular object.

The hypothesised mechanism underpinning the holographic principle is illustrated in Figure 6.4 in Chapter 6. We considered an example in which the word "CAT" is presented to the network, and considered how information about the elemental parts of the word may be projected upwards to the higher layers.

An experimental prediction of these simulation results is that border ownership information should be represented in higher cortical visual layers, but which may depend on the whole object stimulus presented to the visual system.

7.4.3 The Emergence of Polychronization and *Binding Neurons* within the Spiking Network Model

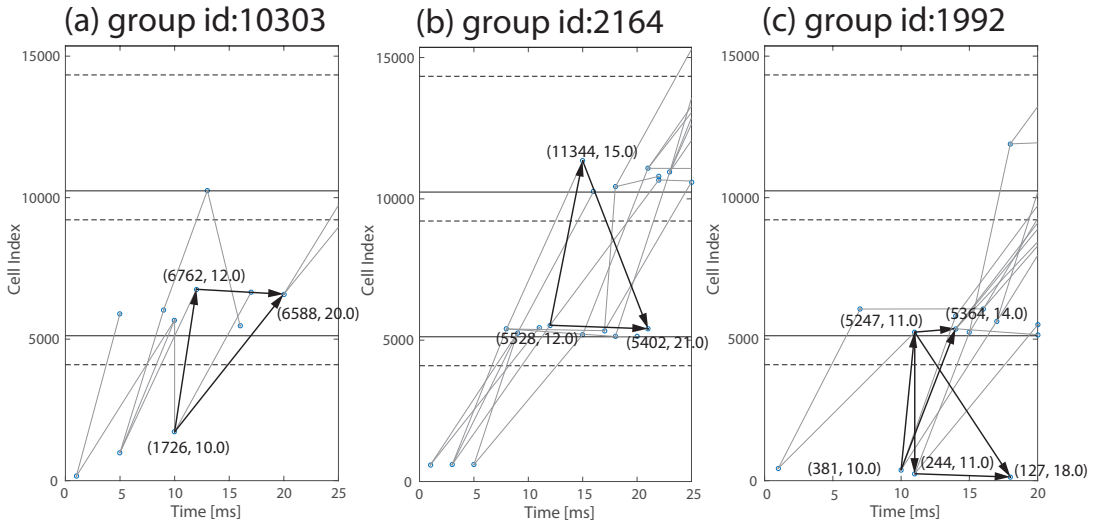

Figure 7.7: **Examples of polychronous groups that encode binding relationships between lower and higher level visual features.** Simulation results are presented for the network after it has been trained on the visual objects shown in Figure 7.2. Each of the three subplots shows one or two examples of 3-neuron polychronous groups similar to those hypothesised and illustrated in Figure 7.1. In these plots, the neurons are identified by small circles and the strengthened connections between the neurons are represented by bold lines. The neurons are plotted along the abscissa according to the relative timings of their spikes within the polychronous groups, which was determined by the axonal transmission delays of the strengthened connections between the neurons. In each of these examples, it can be seen that two pre-synaptic neurons send strengthened connections to a third neuron, where the spikes from the first two neurons arrive simultaneously at the third neuron during a stimulus presentation. The third neuron thus operates like a binding neuron, as hypothesised in section 7.2.3. However, in the polychronous groups shown, it was not in fact possible to clearly identify the firing properties of the two sending neurons, which had rather more obscure firing characteristics than originally hypothesised in Figure 7.1.

Finally, for the same trained network, all the potential polychronous groups triggered from cells in the input layer of the network are identified based on the synaptic connectivity, conduction delays, and synaptic weights as explained in Section 7.3.2.2. Figure 7.7 shows the examples of polychronous groups that form binding relationships. In these plots, the neurons are identified by small circles and the strengthened connections between the neurons are represented by bold lines. The neurons are plotted along the abscissa according to the relative timings of their spikes within the polychronous groups, which was determined by the axonal transmission delays of the strengthened connections between the neurons. Each of the three subplots shows one or two examples of 3-neuron polychronous groups similar to those hypothesised and illustrated in Figure 7.1. Figure 7.7(a) shows an example of a 3-neuron polychronous group in which neuron 1726 in the input layer sends a spike to neuron 6762 in the second layer, and these two neurons send spikes that arrive simultaneously at neuron 6588 in the second layer which becomes a binding neuron. Similarly, Figure 7.7(b) shows another example where a neuron in the second layer and another neuron in the third layer both send spikes which arrive simultaneously at a binding neuron in the second layer. Figure 7.7(c) shows an interesting example where the same neuron in the second layer is a part of different binding relationships. As our theory predicts, such binding neurons should develop across the visual field, at every layer of the feature hierarchy, and at every spatial scale within a natural visual image. However, in the polychronous groups shown, it was not in fact possible to clearly identify the firing properties of the two sending

neurons in each subplot, which had rather more obscure firing characteristics. Our original hypothesis proposed that the two sending neurons might be examples of an edge-detecting simple cell in a lower visual layer and an object boundary contour element cell in a higher layer, while the third neuron is a border ownership neuron within a local layer. However, in practice, visual processing is spread over several (i.e. four) hierarchical layers, with the firing properties of neurons gradually developing over these stages. Consequently, the mechanisms underpinning the development of border ownership cells may include neurons with intermediate varieties of response characteristics, and so may not operate in quite such a simple manner as illustrated in Figure 7.1.

7.5 Discussion

Zhou et al. (2000) reported the existence of border ownership cells in cortical areas V1 and V2 of the primate visual cortex that respond to oriented edges, like classic simple cells, but which are also sensitive to which side of an object the boundary edge occurs. Consequently, these kinds of neuron encode which object a particular edge belongs to. Border ownership cells are thus thought to play a key role in feature binding, in this case binding a relatively low-level feature such as a boundary edge to a higher level object.

In the rate-coded neural network model described in Chapter 5, the responses of the border ownership cells arose due to modulation of V1/V2-like edge-detecting simple cells by top-down signals from object boundary contour element cells that developed in the higher layers. These latter cells represent the boundary contour elements of 2-dimensional object shapes. That is, individual object boundary contour element cells respond selectively to boundary elements with a specific curvature at a particular location with respect to the centre of mass of the object (Pasupathy and Connor, 2002). However, our rate-coded model of border ownership cells failed when it was presented with visual scenes containing multiple objects in Chapter 5 (Eguchi and Stringer, 2016). This is interesting because it is when visual scenes contain multiple objects that feature binding becomes more of a challenging problem, as discussed by von der Malsburg (1999). I argued that the problem with a rate-coded model is the lack of spatial specificity in the top-down modulation by signals from object boundary contour element cells in the higher layers of the network, as discussed in detail in Chapter 5 and summarised in section 7.1.3.

In the current work presented above, it was demonstrated that this problem may be solved within a more biologically realistic spiking neural network, in which the timings of action potentials or spikes emitted by neurons are explicitly modelled. In particular, although the set of stimuli used was simplified due to the limitation of computational resources, it was shown that the border ownership cells that develop in our spiking network simulations actually maintain their proper firing characteristics when the network is presented with multiple objects simultaneously. In this way, our new spiking network model provides a solution to the "superposition catastrophe" that may occur within a rate-coded model when multiple stimuli are presented together von der Malsburg (1999).

I proposed that the border ownership neurons that develop in our spiking network model are operating in a manner similar to the feature binding hypothesis proposed in Chapter 6, where the border ownership cells in our simulations are examples of the *binding neurons*. How such cells might operate was discussed in section 7.2.3. In this scenario, border ownership cells are embedded within the kinds of 3-neuron polychronous groups illustrated in Figure 7.1. The emergence of many similar 3-neuron polychronous groups are observed in our simulations, some of which are shown in Figure 7.7. It is important to note that these polychronous groups emerged in the model despite the fact that the visual stimuli are represented by randomised Poisson spike patterns in the input Layer 0 during both training and testing, as described in section 6.3.1.4.

However, in the polychronous groups shown in Figure 7.7, the two sending neurons in each

subplot had rather obscure firing characteristics. Both in our simulations and the primate brain, visual processing is spread over several hierarchical layers, and the firing properties of neurons will gradually develop over these stages. Consequently, the mechanisms underpinning the development of border ownership cells may include neurons with intermediate varieties of response characteristics, and so may not operate in quite such a simple manner as illustrated in Figure 7.1. In future research, I will investigate exactly how this process operates when spread over several cortical stages, with a series of intermediate neuronal response types.

One interesting observation in our simulations is that many neurons developed the response characteristics of border ownership cells throughout all four layers 1 to 4. This behaviour was consistent with the *holographic principle* proposed in Chapter 6. I posited that information about low-level visual features may propagated upwards to the higher (output) layers of the network, where such information may then be read out by later brain systems. The computational plausibility of this hypothesis was supported by the simulations presented above, where border ownership information was seen to be projected upwards and encoded by neurons in the highest layers of the network.

In conclusion, I have shown how a more biologically realistic spiking neural network may develop border ownership neurons that maintain their response selectivities when presented with visual scenes containing multiple objects. I have proposed and provided evidence that, in such a spiking network model, the border ownership cells are operating in a manner somewhat analogous to the binding neurons hypothesised in Chapter 6. Thus, taken together, the failed rate-coded simulations reported in Chapter 5 (Eguchi and Stringer, 2016) and the successful spiking network results described above provide theoretical support for the binding hypothesis advanced in Chapter 6. In other words, I suggest that the existence of border ownership cells in the brain, which maintain their firing properties in multi-object scenes, may only be explained by the general binding hypothesis in Chapter 6. This begs the question, what other kinds of neuronal response property found in the brain require the same kind of binding mechanism? In future research, I will work with experimental neuroscientists to identify such neuronal responses in the brain in order to provide further support for this potential solution to the binding problem.

Chapter 8

Conclusion

In this doctoral thesis, I conducted a series of simulation studies to advance understanding of how the primate brain learns to process visual input from natural scenes. Throughout the studies, I provided a number of important theoretical insights and predictions based on the modelling results. For example, I show that the transform invariant neuronal representation of the local contour elements of a visual object and the later integrated representations of those features can be developed solely based on the visual statistics in a feed-forward neuronal network. Also, I propose that the development of the monotonic tunings of the spatial relationships of facial features can be developed simply due to the limited size of receptive fields, and the integration of those representations can contribute to the development of more global attributes of faces such as identity and expression. This investigation into the neuronal representation of faces in the visual brain was then extended to model the neurobiological basis of a family of clinical treatments for depression known as cognitive bias modification (CBM). These computer models now provide a potentially powerful tool for optimising the design and efficacy of these clinical procedures. At the same time, I also identified a limitation of the traditional rate-coded models in that they leak away essential information about the binding relations between lower and higher level features within a visual scene, and so cannot effectively solve the feature binding problem. I argued that this led to a failure of our rate-coded model of border ownership neurons when tested on multiple objects presented together. Finally, I provided a general solution to the visual feature binding problem within a new spiking neural network model, in which the timings of the action potentials or spikes are explicitly represented. These spiking networks developed polychronous groups of neurons, within which were embedded 'binding neurons' that represent the relations between lower and higher level visual features. Using the new spiking network model, I was able to successfully simulate the development of border ownership neurons that maintained their proper firing properties when tested on scenes containining multiple objects. I next provide a summary of the simulation studies reported in each chapter of the thesis.

8.1 Object Shape Representations

In Chapter 2, I demonstrated that when a neural network model, VisNet, of the primate ventral visual pathway is trained on many objects with different boundary shapes, the neurons in the higher layers of the network learn to respond to localised boundary contour elements, which are defined by the curvature and location of the boundary element in the frame of reference of the object. Interestingly, neurons learn to respond to these boundary elements rather than learning to respond to the whole objects that were actually presented during training. Moreover, the neurons were able to learn to respond with translation invariance as visual objects are shifted across different retinal locations. This was shown to be successful when VisNet was trained with either the artificially constructed visual stimuli used in Studies 1 and 2, or with images of natural visual objects in Study 3. A population of such neurons, representing many different

boundary elements of different curvature and position within the object, could be used to provide a distributed coding of the entire boundary shape of an object. This has been demonstrated with real neurons in primate visual area V4 by Pasupathy and Connor (2002). As such, these neurons are likely to play an important role in how the primate visual system represents the shapes of objects.

The primary contribution of Chapter 2 is to elucidate and test two key biologically plausible learning mechanisms that can combine to promote the development of these neuronal response characteristics. First, if the network is trained on many objects with different boundary shapes, where each boundary is comprised of a different constellation of contour elements, then this leads to a statistical decoupling between the boundary elements. This is sufficient to allow the competitive layers of VisNet to develop neurons that respond to individual boundary elements defined by curvature and position within the object, which are similar to the neurons reported in the physiological experiments conducted by Pasupathy and Connor (2001). Secondly, neurons may learn to respond with translation invariance across different retinal locations through the use of a trace learning rule. This kind of learning places constraints on the statistics of how the eyes move and visual objects change or transform on the retina. Specifically, it is assumed that the eyes usually shift about a visual scene faster than the objects are changing or rotating on the retina. These two mechanisms together provide a biologically plausible account of how neurons in the primate ventral visual pathway may learn to represent localised boundary contour elements of objects as revealed by Pasupathy and Connor (2001).

Furthermore, neurophysiological experiments carried out by Brincat and Connor (2004) have shown that neurons in the later stages of the ventral visual pathway, TEO and posterior TE, integrate information from multiple boundary contour elements. In our simulations, the number of cells that were tuned to combinations of multiple contours increased in the higher layers. Tracing back the feed-forward synaptic connectivity to these output neurons confirmed that their selectivities were built by combining inputs from neurons representing each local boundary contour in the preceding layer.

The simulations reported in Chapter 2 are the first to show how neuronal responses encoding the local boundary conformation of objects may develop through a biologically plausible process of visually-guided learning. Both the Hebb learning rule (1.7) and trace learning rule (1.8), (1.9) used in the simulation studies are biologically plausible in that they are 'local' learning rules, which only use locally available biological quantitites, such as the activity of the pre- and post-synaptic neurons, to modify the synaptic weights. This is in sharp contrast to other modelling studies that manually set up the synaptic weights in a non-local manner.

Another important factor that underpins the biological plausibility of the simulations carried out in this chapter is that the network model was always trained on whole objects rather than carefully pre-segmented and isolated parts of objects corresponding to local boundary elements. Indeed, in Study 3, VisNet was trained on a random assortment of whole natural visual objects. Nevertheless, the network was still able to develop neurons that were specifically tuned to localised boundary segments of objects. I also found the performance of the model to be extremely robust, which gives additional credence to the learning mechanisms explored in this chapter.

8.2 Face Representations

In Chapter 3, I presented biologically plausible neural network simulations of the visually-guided development of facial representations in the VisNet model (Section 1.3) (Wallis and Rolls, 1997). In particular, I trained VisNet with realistic human face stimuli constructed using FaceGen (FaceGen, 2013). As a result, I found that the network successfully developed various kinds of cells with response properties similar to those reported in neurophysiological studies. To further advance our understanding of the learning mechanisms involved, additional simulations

were performed within simplified one-layer competitive network models.

Our initial simulations with the VisNet model showed the development of neurons that learned to respond to individual facial features such as the eyes and mouth, as well as combinations of these features, as has been reported in single cell recordings in the macaque brain (Freiwald et al., 2009). However, the question was how neurons might learn to respond to individual facial features if the facial features are always seen together within whole faces during training. Particular facial features such as the eyes occur in different shapes across different faces. Thus, across a population of faces the network will be exposed to different combinations of facial feature shapes on different occasions. This will lead to a statistical decoupling (Stringer et al., 2007; Stringer and Rolls, 2008) between the individual facial features, which I hypothesised may force the neurons in higher layers to learn to represent the individual features rather than whole faces. This hypothesis was confirmed in the VisNet simulations, where it was found that the output neurons switched to predominantly representing the individual facial features as the number of possible shapes of any facial feature p used to generate the set of training faces increased from 1 to 2.

I further hypothesised that as the number of shapes of any facial feature p increased further, an invariance learning mechanism known as continuous transformation (CT) learning would begin to drive the development of neurons that responded invariantly to many or all of the shape variations of a particular facial feature. Such neurons would represent a facial feature such as a mouth irrespective of the particular shape of that feature. This hypothesis was also confirmed in VisNet simulations as p was increased to 5, 10, and 30. At $p = 30$ there was a sharp rise in the number of neurons that responded to all 50 of the differently shaped eyes used to test the network.

Furthermore, the VisNet simulations also developed some cells with monotonically increasing or decreasing tuning responses to gradually changing spatial relations between facial features such as inter-eye distance, as has been observed in neurophysiology studies (Freiwald et al., 2009). The question was how such monotonic response properties develop. In complementary simulations of a one-layer competitive network, I found that the finite receptive field of a neuron due to a topologically restricted fan-in of afferent synaptic connections, as well as the nature of the competition within the output layer, both played important roles in the emergence of neurons with monotonic tuning.

I also found that VisNet developed neurons encoding global facial attributes such as face identity and facial expression as reported in neurophysiology studies (Morin et al., 2014). The question was how different sub-populations of higher layer neurons can learn to respond selectively to either face identity or expression if the network is always exposed to both attributes simultaneously, and the same retinal input neurons represent both global attributes simultaneously in a complex distributed manner. In complementary simulations of a one-layer competitive network, I showed that the network can develop separate representations of multiple perceptual input spaces such as facial identity and expression even if the input neurons encoding these spaces are fully overlapping. In particular, this may occur when the input patterns vary independently between the different input spaces. This result provides a possible mechanism for the simultaneous development of multiple global facial representations such as facial identity and expression.

Furthermore, one of the most important arguments I raise is that the neurons that encode global attributes of faces (such as facial identity and expression) and the neurons that encode a spatial relationship between facial features (such as inter-eye distance) are essentially the same. More specifically, I propose that neurons encoding different global attributes such as expression simply represent different spatial relationships between local features with monotonic tuning curves or particular combinations of these spatial relations. In this way, the population response of a set of facial features would be amplified for extreme compared with intermediate feature values along the visual pathway, and thereby explain why faces with more deviant appearances

are recognized better than those which are more typical (Rhodes, 1997; Benson and Perrett, 1991; Bruce and Young, 2011). In particular, this proposal contrasts sharply with the idea of neurons being assigned in an entirely random distributed manner to represent particular facial identities. Instead, neurons encoding facial identity are in fact representing specific structural information about the faces they encode. Our simulation results provide convincing evidence for this argument.

8.3 Cognitive Bias Modification (CBM)

In Chapter 4, I described and modelled two alternative Cognitive Bias Modification (CBM) training mechanisms; continuous transformation (CT) learning (Stringer et al., 2006) and trace learning (Foldiak, 1991; Wallis and Rolls, 1997). These learning mechanisms were previously used to model how the primate ventral visual pathway learns to perform transform invariant visual object recognition. CT learning binds together input stimuli onto the same categorical output representation using spatial continuity, while trace learning binds together stimuli using temporal continuity. Experimental support for these two learning mechanisms has been provided by previous psychophysical studies, which have confirmed that human subjects bind together different images onto a single categorical representation using a mixture of both spatial continuity (CT learning) and temporal continuity (trace learning) (Perry et al., 2006). Our current simulations have shown that these same learning mechanisms may be implemented in neural network computer models to rewire the synaptic connectivity in order to eliminate the kind of negative cognitive biases associated with clinical depression.

The results of these simulations are highly informative for the development of experimental protocols to develop optimal CBM training methodologies with human participants. One such suggestion is to have a clearer focus on the way in which a bias might be altered. This thesis demonstrated a shift in cognitive bias from negative to more positive through the exploitation of two visual learning mechanisms. For a change in cognitive bias to occur, some sort of learning must take place. Therefore, it follows that we should use what we know about learning to inform CBM procedures. Here the stimuli are optimised for use with trace and CT learning, but future work could look at other types of learning. For example, findings from reinforcement learning research could be used to optimise CBM procedures using feedback.

The first of the CBM retraining mechanisms, CT learning, utilizes a Hebbian learning rule (1.7) with weight vector renormalisation (1.10) and (1.11). During training, the face stimuli gradually transform from happy to sad. The initial presentation of the happy face stimulates the happy output representation, which then stays active while the faces morph continuously through more neutral to sad faces. The continual application of Hebbian learning at each face presentation then remaps the more neutral faces onto the active happy output representation. In this way, the positive output neurons that originally fire only to very happy faces are remapped to also fire to the more neutral faces.

The second of the CBM retraining mechanisms, trace learning, utilizes a trace learning rule (1.8) and (1.9) with weight vector renormalisation (1.10) and (1.11). Trace learning encourages output neurons to respond to input patterns that tend to occur close together in time. During training, the sad and neutral faces are presented to the network in an interleaved manner. The application of trace learning at each face presentation then binds the happy and neutral faces together onto the same happy output representation. In this way, positive output neurons are remapped to also respond to neutral faces.

The two CBM training mechanisms, CT learning and trace learning, were first tested in a simplified one-layer neural network model in order to investigate the operation of these learning mechanisms in a highly controlled way. These computer simulations allowed us to explore the neural and synaptic dynamics underpinning the two CBM training mechanisms. It was found that both CT learning and trace learning were able to remap the synaptic connectivity such

that the happy output cell responded to the happy to neutral portion of the stimulus range, while the sad output cell responded only to the sad end of the stimulus range. Thus CBM retraining by either CT learning or trace learning produced successful CBM, where the bias in the network was shifted from negative to positive.

Next I tested the CBM training methodologies in a much more biologically detailed multi-layer model, VisNet, of the primate ventral visual pathway with realistic face images generated using the FaceGen 3D face modelling software package. The network was first pretrained on 100 randomly generated faces with a variety of facial identities and expressions ranging from happy to sad. During this pretraining stage, the network developed output neurons responding preferentially to either happy or sad facial expressions, as shown in Chapter 3 (Eguchi et al., 2016). Then the network underwent CBM retraining using either CT learning or trace learning as described in the sections on Experiment 2a and Experiment 2b, respectively. It was found that both CT learning and trace learning were able to remap the more neutral faces away from the sad output neurons and onto the happy output neurons, thus shifting the cognitive bias in the network connectivity from negative to positive.

To the authors' knowledge, this is the first study that has modelled the application of the CT learning and trace learning mechanisms to CBM-Interpretation. Previous experimental studies have found that CBM-Interpretation can reduce negative cognitive biases in human participants (e.g. Grey and Mathews, 2000; Mathews and Mackintosh, 2000), which in turn can reduce the risk for depression recurrence (Holmes et al., 2009). This Chapter provides potential explanations at the neuronal and synaptic level for how such a shift in interpretational bias might occur through CBM training. Understanding the way in which biases can be shifted is crucial at present, given the mixed results seen in CBM research so far (Fox et al., 2014).

In the future, our laboratory plans to develop computer simulations aimed at shedding light on the operation of antidepressants in the brain as an extension to the work presented in this thesis. The mechanisms by which antidepressants shift information processing from negative to positive may be similar to what has been proposed in the simulation studies for the CBM training methodologies. However, the initial shift in cognitive bias caused by antidepressants must now be achieved by some form of pharmaceutically driven global neuromodulation. Such neuromodulation may affect the processing of every day sensory experiences, which in turn may drive further synaptic modification resulting in reduced innate negative biases. Therefore, I will explore how global changes to neuronal and synaptic model parameters may lead to shifts in cognitive bias similar to those described above for the CBM training methodologies.

8.4 Border Ownership Representations

In Chapter 5, I investigated through computer simulation how top-down connections may play a fundamental role in the development of border ownership representations in the early cortical visual layers V1/V2. In terms of the novelty, this work is different from previous modelling studies that have already proposed hypothetical neural circuits for such coding in that I investigated how such circuits may develop using a biologically plausible, local, trace learning rule to modify the synaptic connectivity during visual experience.

The simulations reported in the chapter have demonstrated how top-down connections may help to guide competitive learning in lower layers, thus driving the formation of lower level (border ownership) visual representations in V1/V2 that are modulated by higher level (object boundary element) representations in V4. More precisely, I showed that simple cells in area V1 representing a vertical straight edge at a particular retinal location can learn to be modulated by top-down connections from higher level representations of object shape in, for example, area V4 (Pasupathy and Connor, 2001, 2002). However, more importantly, I also identified a key limitation of the rate-coded model in that it does not maintain the proper firing characteristics of border ownership cells when more than one object stimulus is presented to the network at the

same time after training. This is in contrast to neurophysiological studies which have shown that border ownership cells in the brain do indeed maintain their firing properties for multi-figure displays (Qiu et al., 2007; Martin and von der Heydt, 2015).

The result suggests that the incorporation of additional top-down connections, although necessary, is not sufficient by itself to allow the network to develop robust border ownership representations in the early layers and thus solve this kind of feature binding problem. Our rate-coded model failed because the network is not able to maintain an explicit representation of which features are part of which objects. To solve this problem, I propose that it is important to have a form of binding neuron (e.g. border ownership neuron in V1/V2) that responds if and only if a neuron representing a low-level feature (such as a simple oriented edge) is actually participating in driving a neuron representing a high-level feature (such as an object boundary element). The binding neuron should not respond if the neuron representing the low-level feature and the neuron representing the high-level feature just happen to be co-active, where the former is not actually driving the latter. Such unrelated co-activation of low and high-level feature neurons might occur, for example, because of the presence of multiple similar objects within a complex natural scene. Then the question is what further biological details need to be incorporated into the model in order to allow it to form such binding (e.g. border ownership) representations that maintain their firing properties under these more general stimulus conditions.

8.5 Polychronization and Feature Binding in a Spiking Network Model

In the Chapter 6, I explored the operation of a biologically detailed spiking neural network model of the primate ventral visual system. The model incorporates the following key aspects of cortical dynamics and architecture: (i) the model implements spiking neural dynamics in which the timings of action potentials or 'spikes' are simulated explicitly, (ii) Spike-Timing-Dependent Plasticity (STDP) is used to modify the synaptic connections during visually-guided learning, (iii) the network architecture incorporates bottom-up, top-down and lateral synaptic connections, (iv) the synaptic connectivity between neurons incorporates distributions of axonal conduction delays of varying durations, (v) in some simulations multiple synaptic connections with different axonal transmission delays are incorporated between each pair of pre- and post-synaptic neurons. These are basic known aspects of the architecture and function of the visual cortex. Using this model architecture, I explored a number of major computational hypotheses.

For example, I hypothesised that even if the visual stimuli (circle, heart, and star) presented to the network were encoded in the input layer by randomised Poisson spike trains, the synaptic connectivity in the later layers of the network would self-organise using STDP during visually-guided learning such that polychronous groups (PGs) would emerge naturally. Moreover, I anticipated that individual PGs would learn to respond to particular stimuli that the network was trained on. This was confirmed in our simulations reported above.

The output (4th) layer was found to carry more stimulus information if I assumed a temporal coding based on patterns of spike times within PGs instead of assuming traditional rate coding by individual neurons. Our results found that the inclusion of feedback and lateral connections in the network structure led to an increase in the number and length of PGs (especially spike-pairs). In particular, the full network architecture with feedforward (FF), feedback (FB), and lateral (LAT) synaptic connections produced the most spike-pair PGs with maximal stimulus information. These spike pair PGs were tuned to specific stimuli.

A major novel result of the work is that this self-organisation of stimulus-specific spike-pair PGs occured even when the stimulus input representations were *randomised* Poisson spike trains, in which the temporal ordering of spikes varied stochastically across different presentations of the same visual stimulus. The development of (spike-pair) PGs using STDP during visual

training in such a spiking network is thus a highly robust process that operates perfectly well with randomised stimulus spike patterns in the lower stages of processing.

The development of temporal PG codes was shown to be dependent on the temporal specificity of the STDP learning rule used to modify the synaptic connections. It was found that the network develops the largest number of spike-pair PGs with maximal information about which stimulus is presented to the network when the STDP time constants are shortest (i.e. 5ms). However, increasing the STDP time constants in the simulations had the effect of decreasing the number of object specific spike-pair PGs that emerged. The explanation for these observations is that increasing the STDP time constants makes the precise timing of the spikes less important for learning, which in turn makes the synaptic weight modification more similar to traditional Hebbian learning in a rate coded model. Consequently, these simulation results suggest an important role for temporally precise STDP in the development of temporal coding.

Another novel feature of some of the simulations reported in this thesis was the incorporation of multiple synaptic contacts with different axonal transmission delays between each pair of pre- and postsynaptic neurons. This corresponds to a presynaptic neuron making multiple synaptic connections on different parts of the dendritic branching of a postsynaptic neuron as is seen among real neurons in the brain. In such a network architecture, STDP was able to select which synapses to strengthen and which synapses to weaken, which promoted the visually-guided development of PGs of spiking neurons. Thus, during self-organisation the network is able to effectively select for synaptic transmission delays between pre- and postsynaptic neurons, which results in a greater representational capacity.

Our new approach to solving the feature binding problem in biological spiking neural networks relies on polychrony, in which polychronous groups of neurons emit their spikes in rich spatiotemporal patterns. This is in contrast to previous proposals based on synchrony, in which groups of neurons emit their spikes simultaneously. Our move away from synchrony is in part driven by the psychological view that simply segmenting a visual scene into several distinct object regions by synchrony cannot accord with the semantically rich, hierarchical visual experience of primate vision as described by Duncan and Humphreys (1989). In particular, in simulations of the full spiking network architecture (FF + FB + LAT), I observed the emergence of binding neurons embedded within 3-neuron polychronous groups of the form illustrated in Figure 6.3a. These binding neurons represent the hierarchical relations between lower and higher level visual features. The simulation results were shown in Figure 6.12. In these simulations, the binding neurons represented the binding relationships between lower level feature neurons in layer 3 and higher level feature neurons in layer 4. Moreover, the individual PGs, in which these binding neurons were embedded, responded to specific visual stimuli (the circle, heart or star). Such binding neurons were originally proposed by von der Malsburg (1999), but without an explanation of how they might emerge naturally during visual development. In the simulations reported above I demonstrated that such binding neurons may develop automatically within the PGs that emerge during visually-guided learning with STDP. In particular, these binding representations were shown to emerge even when the visual stimuli are encoded by randomised (Poisson) spike trains in the input layer. The binding neurons that develop carry measurable information about which low-level features are driving (and hence part of) which high-level features. Our theory predicts that such binding neurons should develop across the visual field, at every layer of the feature hierarchy, and at every spatial scale within a natural visual image.

8.6 Border Ownership Representations in a Spiking Network Model

In Chapter 5, I hypothesised that the failure of the border ownership representations in the VisNet model when multiple objects were presented was primarily due to the implementation of

rate-coding in the model. That is, VisNet only represents the average firing rate of each neuron, and not the actual timings of the action potentials emitted by neurons as occur in the brain. Meanwhile, modelling studies have shown that when randomised distributions of axonal delays are incorporated into a more biologically plausible spiking neural network, then this produces memory patterns in the form of repeating temporal loops of neuronal firings. This phenomenon has been termed *polychronization* (Izhikevich, 2006). In Chapter 6, I showed that the emergence of these temporal memory loops is further enhanced within a recurrently connected spiking network with randomised distributions of axonal conduction delays when the strengths of synaptic connections are modified by STDP (Eguchi et al., 2017a). Recognising the potential importance of all these biological features of cortical operation, I hypothesised that such a spiking model, which also incorporates bottom-up, top-down, and lateral associatively modifiable excitatory connections, may develop border ownership neurons in the lower visual layers (corresponding to cortical areas V1 or V2) that respond selectively to a vertical straight edge on either the left or right boundary of an object presented at the neuron's prefered retinal location, *in a way that is unaffected by the presence of another object seen at a different nearby location on the retina*. More generally, I hypothesised that the biological features discussed in Chapter 6 play important roles in how the primate visual system solves the feature binding problem (von der Malsburg, 1999). Consequently, in the final chapter, I explored how border ownership representations may develop in a new spiking neural network version of the VisNet model, which incorporates bottom-up, top-down, and lateral excitatory connections, distributions of axonal transmission delays, and STDP.

The simulations carried out in Chapter 6 developed many groups of neurons, which we refer to as *polychronous groups*, that emitted regularly repeating temporal chains of spikes in response to the presentation of visual objects. Embedded within these polychronous groups were binding neurons that represented the relations between low and high level features. I hypothesised that similar spiking mechanisms could produce border ownership cells that maintain their proper firing characteristics when multiple visual objects are presented together to the network. It was proposed that in such a spiking network, the border ownership cells would operate in a similar manner to the binding neurons that emerged in Chapter 6. Specifically, I hypothesised that border ownership cells would develop automatically within particular kinds of polychronous group during visual training, where the border ownership cells would become tuned through STDP learning to respond if and only if an edge-detecting V1-like simple cell in a lower layer is participating in driving a V4-like object boundary contour element cell in a higher layer, where the edge represented by the V1-like simple cell directly corresponds to the object boundary element represented by the object boundary contour element cell. In this case, the border ownership cell will carry measurable information that the edge represented by the edge-detecting V1-like simple cell is part of the object boundary element represented by the V4-like object boundary contour element cell. It should be noted that, in this scenario, border ownership cells are not examples of top-down modulated V1-like simple cells as modelled in Chapter 5 in a rate-coded network, but are instead a new category of neurons that is driven by converging inputs from V1-like simple cells in lower layers and object boundary contour element cells in higher layers. I proposed that such border ownership cells would develop automatically within the polychronous groups that emerge during visually-guided learning with STDP.

The key property of the border ownership cells that I hypothesised would develop in Chapter 7 is that an individual cell responds if the lower layer neuron representing the edge is actually participating in driving the higher layer neuron representing the corresponding object boundary contour element. Only in this case will the border ownership cell be fully informative that the edge is part of the object boundary contour element. The border ownership neuron should not respond if the lower layer neuron representing the edge and the higher layer neuron representing the object boundary contour element simply happen to be co-active, where the former is not actually driving the latter. Such unrelated co-activation of a lower layer edge-detecting cell

and higher layer object boundary contour element cell might occur, for example, because of the presence of multiple objects within a natural scene.

Our proposed mechanism for generating border ownership cells ensures that these cells only become activated if the lower layer edge-detecting cell is actually participating in driving the higher layer object boundary contour element cell. This kind of temporally specific response is characteristic of a polychronous group, which the tuple of cells in such binding relationships comprise. I proposed that this mechanism potentially solves the limitations of rate-coding described in Chapter 5. Indeed, for these reasons, I propose that modelling the development of border ownership cells, and more generally feature binding, may not be soluble within a traditional rate-coded network, but will instead require the full spiking neuronal dynamics of the brain.

However, in the polychronous groups shown in Figure 7.7, the two sending neurons in each subplot had rather obscure firing characteristics. Both in our simulations and the primate brain, visual processing is spread over several hierarchical layers, and the firing properties of neurons will gradually develop over these stages. Consequently, the mechanisms underpinning the development of border ownership cells may include neurons with intermediate varieties of response characteristics, and so may not operate in quite such a simple manner as illustrated in Figure 7.1. Future research should investigate exactly how this process operates when spread over several cortical stages, with a series of intermediate neuronal response types.

One interesting observation in our simulations is that many neurons developed the response characteristics of border ownership cells throughout all four layers 1 to 4. This behaviour was consistent with the *holographic principle* proposed in chapter 6. I posited that information about low-level visual features may propagated upwards to the higher (output) layers of the network, where such information may then be read out by later brain systems. The computational plausibility of this hypothesis was supported by the simulations presented in chapter 7, where border ownership information was seen to be projected upwards and encoded by neurons in the highest layers of the network.

8.7 Conclusion

As summarised above, this doctoral thesis provided a number of important contributions towards understanding how the visual system learns to encode the detailed spatial structure of objects and faces within scenes. In particular, I shed light on the learning mechanisms that may give rise to a number of related neuronal firing properties. Nevertheless, it still remains a difficult challenge to understand exactly how many other kinds of neuron that have not been investigated in this thesis may also develop their response properties through visually guided learning. Such examples include the 3-dimensional form of objects (Yamane et al., 2008), a mirror symmetric encoding of the different views of a face (Freiwald and Tsao, 2010; Kietzmann et al., 2012), the material surface properties of objects (Goda et al., 2016), and the perceived colour of object surfaces (Zeki et al., 1999) (A self-organisation of colour opponent receptive fields has been investigated in Eguchi et al. (2014)). A fundamentally important question is how does the brain integrate all of this diverse visual information in order to provide a unified percept of whole scenes, with correct binding between all of the features at every spatial scale and across the visual field? I began to answer this by moving towards a new generation of more biologically realistic spiking neural networks with learning based on STDP, bottom-up, top-down, and lateral synaptic connections, and distributions of axonal transmission delays. Such networks are able to represent the binding relations throughout a visual scene through the emergence of polchronization and binding neurons. I believe that this new generation of biological spiking neural networks may ultimately prove to be of great utility in many engineering applications such as the recognition of objects, facial identity and expression, and indeed making sense of whole scenes. However, we are just at the beginning of exploring the operation

of such networks.

Bibliography

Abeles, M. (1991). *Corticonics: Neural Circuits of the Cerebral Cortex*. Cambridge University Press, Cambridge ; New York.

Akolkar, H., Meyer, C., Clady, X., Marre, O., Bartolozzi, C., Panzeri, S., and Benosman, R. (2015). What Can Neuromorphic Event-Driven Precise Timing Add to Spike-Based Pattern Recognition? *Neural Computation*, 27(3):561–593.

Almond, S. and Healey, A. (2003). Mental health and absence from work: New evidence from the UK Quarterly Labour Force Survey. *Work, Employment and Society*, 17(4):731–742.

Amit, D. J. and Brunel, N. (1997). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral Cortex*, 7(3):237–252.

Anderson, M. C. and Green, C. (2001). Suppressing unwanted memories by executive control. *Nature*, 410(6826):366–369.

Angelucci, A. and Bullier, J. (2003). Reaching beyond the classical receptive field of V1 neurons: horizontal or feedback axons? *Journal of Physiology-Paris*, 97(2-3):141–154.

Baek, K. and Sajda, P. (2005). Inferring figure-ground using a recurrent integrate-and-fire neural circuit. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(2):125–130.

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

Beck, A. T. (2008). The evolution of the cognitive model of depression and its neurobiological correlates. *American Journal of Psychiatry*, 165:969–977.

Becker, S. (1999). Implicit learning in 3d object recognition: The importance of temporal context. *Neural Computation*, 11:347–374.

Beeck, H. P. O. d., Baker, C. I., DiCarlo, J. J., and Kanwisher, N. G. (2006). Discrimination Training Alters Object Representations in Human Extrastriate Cortex. *The Journal of Neuroscience*, 26(50):13025–13036.

Benson, P. J. and Perrett, D. I. (1991). Synthesising continuous-tone caricatures. *Image and Vision Computing*, 9(2):123–129.

Berkes, P. and Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision*, 5(6):9–9.

Bernacchia, A. and Amit, D. J. (2007). Impact of spatiotemporally correlated images on the structure of memory. *Proceedings of the National Academy of Sciences*, 104(9):3544–3549.

Bi, G.-q. and Poo, M.-m. (1998). Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type. *The Journal of Neuroscience*, 18(24):10464–10472.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2):115–147.

Biederman, I. and Kalocsai, P. (1997). Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 352(1358):1203–1219.

Blumenfeld, B., Preminger, S., Sagi, D., and Tsodyks, M. (2006). Dynamics of Memory Representations in Networks with Novelty-Facilitated Synaptic Plasticity. *Neuron*, 52(2):383–394.

Booth, M. C. and Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral cortex (New York, N.Y.: 1991)*, 8(6):510–523. PMID: 9758214.

Bourke, C., Douglas, K., and Porter, R. (2010a). Processing of facial emotion expression in major depression: A review. *Australian and New Zealand Journal of Psychiatry*, 44:681–696.

Bourke, C., Douglas, K., and Porter, R. (2010b). Processing of Facial Emotion Expression in Major Depression: A Review. *Australian & New Zealand Journal of Psychiatry*, 44(8):681–696.

Brincat, S. L. and Connor, C. E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature neuroscience*, 7(8):880–886. PMID: 15235606.

Browning, M., Holmes, E. A., Charles, M., Cowen, P. J., and Harmer, C. J. (2012). Using Attentional Bias Modification as a Cognitive Vaccine Against Depression. *Biological Psychiatry*, 72(7):572–579.

Bruce, V. and Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, 77(3):305–327.

Bruce, V. and Young, A. (2011). *Face Perception*. Psychology Press, London ; New York.

Cadieu, C., Kouh, M., Pasupathy, A., Connor, C. E., Riesenhuber, M., and Poggio, T. (2007). A model of v4 shape selectivity and invariance. *Journal of neurophysiology*, 98(3):1733–1750. PMID: 17596412.

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., and Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research*, 41(9):1179–1208.

Carlbring, P., Apelstrand, M., Sehlin, H., Amir, N., Rousseau, A., Hoffmann, S. G., and Andersson, G. (2012). Internet-delivered attention bias modification training in individuals with social anxiety disorder: A double blind randomized controlled trial. *BMC Psychiatry*, 12:66.

Chance, J. E., Turner, A. L., and Goldstein, A. G. (1982). Development of differential recognition for own- and other-race faces. *The Journal of Psychology*, 112(1st Half):29–37.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204.

Clarke, P. J. F., Notebaert, L., and MacLeod, C. (2014). Absence of evidence or evidence of absence: Reflecting on therapeutic implementations of attentional bias modification. *BMC Psychiatry*, 14.

Cowey, A. and Rolls, E. T. (1974). Human cortical magnification factor and its relation to visual acuity. *Experimental Brain Research*, 21(5):447–454.

Cox, D. D., Meier, P., Oertelt, N., and DiCarlo, J. J. (2005). 'Breaking' position-invariant object recognition. *Nature neuroscience*, 8(9):1145–1147. PMID: 16116453.

Craft, E., Schtze, H., Niebur, E., and von der Heydt, R. (2007). A Neural Model of Figure-Ground Organization. *Journal of Neurophysiology*, 97(6):4310–4326.

Cristea, I. A., Kok, R. N., and Cuijpers, P. (2015). Efficacy of cognitive bias modification interventions in anxiety and depression: Meta-analysis. *British Journal of Psychiatry*, 206:7–16.

Cumming, B. G. and Parker, A. J. (1999). Binocular neurons in v1 of awake monkeys are selective for absolute, not relative, disparity. *The Journal of Neuroscience*, 19(13):5602–5618. PMID: 10377367.

Damon, F., Quinn, P. C., Heron-Delaney, M., Lee, K., and Pascalis, O. (2016). Development of category formation for faces differing by age in 9- to 12-month-olds: An effect of experience with infant faces. *British Journal of Developmental Psychology*, 34(4):582–597.

Daniel, P. M. and Whitteridge, D. (1961). The representation of the visual field on the cerebral cortex in monkeys. *The Journal of Physiology*, 159(2):203–221.

Deco, G. and Lee, T. (2002). A unified model of spatial and object attention based on inter-cortical biased competition. *Computer Science Department*.

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

Deger, M., Helias, M., Rotter, S., and Diesmann, M. (2012). Spike-Timing Dependence of Structural Plasticity Explains Cooperative Synapse Formation in the Neocortex. *PLOS Computational Biology*, 8(9):e1002689.

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience*, 3:1–8.

Diekman, C., Dasgupta, K., Nair, V., and Unnikrishnan, K. P. (2014). Discovering Functional Neuronal Connectivity from Serial Patterns in Spike Train Data. *Neural Computation*, 26(7):1263–1297.

Duncan, J. and Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, 96(3):433–458.

Eguchi, A., Humphreys, G. W., and Stringer, S. M. (2016). The visually-guided development of facial representations in the primate ventral visual pathway: a computer modelling study. *Psychological Review*, 123(6).

Eguchi, A., Isbister, J., Ahmad, N., and Stringer, S. (2017a). The emergence of polychronization and feature binding in a spiking neural network model of the primate ventral visual system. *Manuscript submitted for publication.*

Eguchi, A., Mender, B. M. W., Evans, B., Humphreys, G., and Stringer, S. (2015). Computational modeling of the neural representation of object shape in the primate ventral visual system. *Frontiers in Computational Neuroscience*, 9(100):100.

Eguchi, A., Neymotin, S. A., and Stringer, S. M. (2014). Color opponent receptive fields self-organize in a biophysical model of visual cortex via spike-timing dependent plasticity. *Frontiers in Neural Circuits*, 8(16):16.

Eguchi, A. and Stringer, S. M. (2016). Neural Network Model Develops Border Ownership Representation through Visually Guided Learning. *Neurobiology of Learning and Memory*.

Eguchi, A., Walters, D., Peerenboom, N., Dury, H., Fox, E., and Stringer, S. (2017b). Understanding the Neural Basis of Cognitive Bias Modification as a Clinical Treatment for Depression. *Journal of Consulting and Clinical Psychology*, 85(3).

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., and Rasmussen, D. (2012). A Large-Scale Model of the Functioning Brain. *Science*, 338(6111):1202–1205.

Elliffe, M. C. M., Rolls, E. T., and Stringer, S. M. (2002). Invariant recognition of feature combinations in the visual system. *Biological Cybernetics*, 86(1):59–71.

Emmelkamp, P. (2012). Attention bias modifcation: The Emperor's new suit? *BMC Medicine*, 10:63.

Engell, A. D. and Haxby, J. V. (2007). Facial expression and gaze-direction in human superior temporal sulcus. *Neuropsychologia*, 45(14):3234–3241.

Evans, B. D. and Stringer, S. M. (2012). Transformation-invariant visual representations in self-organizing spiking neural networks. *Frontiers in Computational Neuroscience*, 6:46.

Evans, B. D. and Stringer, S. M. (2013). How Lateral Connections and Spiking Dynamics May Separate Multiple Objects Moving Together. *PLoS ONE*, 8(8):e69952.

FaceGen (2013). Version 3.c.1 [computer software].

Fares, T. and Stepanyants, A. (2009). Cooperative synapse formation in the neocortex. *Proceedings of the National Academy of Sciences of the United States of America*, 106(38):16463–16468.

Findlay, J. M. and Gilchrist, I. D. (2003). Active vision: The psychology of looking and seeing. *Journal of Neuro-ophthalmology - J NEURO-OPHTHALMOL*, 26(1).

Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200.

Fox, E., Mackintosh, B., and Holmes, E. (2014). Travellers' tales in cognitive bias modification research: A commentary on the special issue. *Cognitive Therapy Research*, 38:239–247.

Freeman, J. and Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, 14(9):1195–1201.

Freeman, J. B., Rule, N. O., Adams, R. B., and Ambady, N. (2010). The neural basis of categorical face perception: graded representations of face gender in fusiform and orbitofrontal cortices. *Cerebral Cortex (New York, N.Y.: 1991)*, 20(6):1314–1322.

Freiwald, W. A. and Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, 330(6005):845–851. PMID: 21051642.

Freiwald, W. A., Tsao, D. Y., and Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nature Neuroscience*, 12(9):1187–1196.

Frewen, P. A., Dozois, D. J. A., Joanisse, M. F., and Neufeld, R. W. J. (2008). Selective attention to threat versus reward: Meta-analysis and neural-network modeling of the dot-probe task. *Clinical Psychology Review*, 28(2):307–337.

Friedman-Hill, S. R., Robertson, L. C., and Treisman, A. (1995). Parietal Contributions to Visual Feature Binding: Evidence from a Patient with Bilateral Lesions. *Science*, 269(5225):853–855.

Friesen, W. V. and Ekman, P. (1976). *Pictures of Facial Affect*. Consulting psychologists Press.

Fujii, H., Ito, H., Aihara, K., Ichinose, N., and Tsukada, M. (1996). Dynamical Cell Assembly Hypothesis - Theoretical Possibility of Spatio-temporal Coding in the Cortex. *Neural Networks*, 9(8):1303–1350.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.

Galeazzi, J. M., Minini, L., and Stringer, S. (2015). The development of hand-centred visual representations in the primate brain: a computer modelling study using natural visual scenes. *Frontiers in Computational Neuroscience*, 9:147.

George, D. and Hawkins, J. (2005). A hierarchical bayesian model of invariant pattern recognition in visual cortex. In *IEEE International Joint Conference on Neural Networks*, volume 3, pages 1812–1817. IEEE.

Giersch, A. (2001). The effects of lorazepam on visual integration processes: How useful for neuroscientists? *Visual Cognition*, 8(3):549–563.

Gillebert, C. R., Op de Beeck, H. P., Panis, S., and Wagemans, J. (2008). Subordinate Categorization Enhances the Neural Selectivity in Human Object-selective Cortex for Fine Shape Differences. *Journal of Cognitive Neuroscience*, 21(6):1054–1064.

Goda, N., Yokoi, I., Tachibana, A., Minamimoto, T., and Komatsu, H. (2016). Crossmodal Association of Visual and Haptic Material Properties of Objects in the Monkey Ventral Visual Cortex. *Current Biology*, 26(7):928–934.

Gosselin, F. and Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research*, 41(17):2261–2271.

Grey, S. and Mathews, A. (2000). Effects of training on interpretation of emotional ambiguity. *The Quarterly Journal of Experimental Psychology Section A*, 53(4):1143–1162.

Grey, S. J. and Mathews, A. M. (2009). Cognitive bias modification - Priming with an ambiguous homograph is necessary to detect an interpretation training effect. *Journal of Behavior Therapy and Experimental Psychiatry*, 40(2):338–343.

Gross, C. G., Bender, D. B., and Rocha-Miranda, C. E. (1969). Visual receptive fields of neurons in inferotemporal cortex of the monkey. *Science*, 166(3910):1303–1306. PMID: 4982685.

Gross, C. G., Rocha-Miranda, C. E., and Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of Neurophysiology*, 35(1):96–111.

Hakamata, Y., Lissek, S., Bar-Haim, Y., Britton, J. C., Fox, N. A., Leibenluft, E., Ernst, M., and Pine, D. S. (2010). Attention Bias Modification Treatment: A Meta-Analysis Toward the Establishment of Novel Treatment for Anxiety. *Biological Psychiatry*, 68(11):982–990.

Hallion, L. S. and Ruscio, A. M. (2011). A meta-analysis of the effect of cognitive bias modification on anxiety and depression. *Psychological Bulletin*, 137:940–958.

Hansen, K. A., Kay, K. N., and Gallant, J. L. (2007). Topographic organization in and near human visual area v4. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 27(44):11896–11911.

Harmer, C. J. (2012). Emotional Processing and Antidepressant Action. In Cowen, P. J., Sharp, T., and Lau, J. Y. F., editors, *Behavioral Neurobiology of Depression and Its Treatment*, number 14 in Current Topics in Behavioral Neurosciences, pages 209–222. Springer Berlin Heidelberg.

Harmer, C. J., Mackay, C. E., Reid, C. B., Cowen, P. J., and Goodwin, G. M. (2006). Antidepressant Drug Treatment Modifies the Neural Processing of Nonconscious Threat Cues. *Biological Psychiatry*, 59(9):816–820.

Harmer, C. J., OSullivan, U., Favaron, E., Massey-Chase, R., Ayres, R., Reinecke, A., Goodwin, G. M., and Cowen, P. J. (2009). Effect of Acute Antidepressant Administration on Negative Affective Bias in Depressed Patients. *American Journal of Psychiatry*, 166(10):1178–1184.

Hasselmo, M. E., Rolls, E. T., and Baylis, G. C. (1989a). Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research*, 75(3):415–429. PMID: 2713076.

Hasselmo, M. E., Rolls, E. T., and Baylis, G. C. (1989b). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural Brian Research*, 32:203–218.

Haxby, Hoffman, and Gobbini (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, 4(6):223–233. PMID: 10827445.

Henriksson, L., Mur, M., and Kriegeskorte, N. (2015). Faciotopy - a face-feature map with face-like topology in the human occipital face area. *Cortex*, 72:156–167.

Hertz, J. A., Krogh, A. S., and Palmer, R. G. (1991). *Introduction To The Theory Of Neural Computation, Volume I.* Westview Press, first paperback edition edition.

Hirsch, C. and Mathews, A. (1997). Interpretative inferences when reading about emotional events. *Behavior Research and Therapy*, 35:1123–1132.

Holmes, E. A., Lang, T. J., and Sham, D. M. (2009). Developing interpretation bias modification as a "Cognitive Vaccine" for depressed mood: Imagining positive events makes you feel better than thinking about them verbally. *Journal of Abnormal Psychology*, 118:76–88.

Homola, G. A., Jbabdi, S., Beckmann, C. F., and Bartsch, A. J. (2012). A brain network processing the age of faces. *PLoS ONE*, 7(11):e49451.

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):106–154.2. PMID: 14449617 PMCID: PMC1359523.

Hung, C., Carlson, E. T., and Connor, C. E. (2012). Medial Axis Shape Coding in Macaque Inferotemporal Cortex. *Neuron*, 74(6):1099–1113.

Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science*, 310(5749):863–866.

Isik, L., Leibo, J. Z., and Poggio, T. (2012). Learning and disrupting invariance in visual recognition with a temporal association rule. *Frontiers in Computational Neuroscience*, 6:37.

Ito, M. and Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area v2 of macaque monkeys. *The Journal of Neuroscience*, 24(13):3313–3324. PMID: 15056711.

Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, 73:218–226.

Izhikevich, E. M. (2006). Polychronization: computation with spikes. *Neural Computation*, 18(2):245–282.

Izhikevich, E. M., Gally, J. A., and Edelman, G. M. (2004). Spike-timing Dynamics of Neuronal Groups. *Cerebral Cortex*, 14(8):933–944.

Jeanson, F. (2011). Coincidence detection: Towards an alternative to synaptic plasticity. In *Proceedings of CogSci 2011*.

Jehee, J. F. M., Lamme, V. A. F., and Roelfsema, P. R. (2007). Boundary assignment in a recurrent network architecture. *Vision Research*, 47(9):1153–1165.

Jones, J. P. and Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, 58(6):1187–1211. PMID: 3437330.

Jones, S. (2016). 50 million years of work could be lost to anxiety and depression. *The Guardian*.

Joormann, J., Hertel, P. T., Brozovich, F., and Gotlib, I. H. (2005). Remembering the Good, Forgetting the Bad: Intentional Forgetting of Emotional Material in Depression. *Journal of Abnormal Psychology*, 114(4):640–648.

Julian, K., Bear, C., Schmidt, N. B., Powers, M. B., and Smits, J. A. J. (2012). Attention training to reduce attention bias and social stressor reactivity: An attempt to replicate and extend previous findings. *Behaviour Research and Therapy*, 50:350–358.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 17(11):4302–4311. PMID: 9151747.

Kayaert, G., Biederman, I., Op de Beeck, H. P., and Vogels, R. (2005). Tuning for shape dimensions in macaque inferior temporal cortex. *The European journal of neuroscience*, 22(1):212–224. PMID: 16029211.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Comput Biol*, 10(11):e1003915.

Kiani, R., Esteky, H., Mirpour, K., and Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, 97(6):4296–4309.

Kietzmann, T. C., Swisher, J. D., Knig, P., and Tong, F. (2012). Prevalence of selectivity for mirror-symmetric views of faces in the ventral and dorsal visual pathways. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 32(34):11763–11772.

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. PMID: 8201425.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69.

Kreiter, A. K. and Singer, W. (1996). Stimulus-dependent synchronization of neuronal responses in the visual cortex of the awake macaque monkey. *Journal of Neuroscience*, 16(7):2381–2396.

Kriegeskorte, N., Mur, M., Bandettini, P. A., Kriegeskorte, N., Mur, M., and Bandettini, P. (2008a). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P. A. (2008b). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141.

Lades, M., Vorbruggen, J., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R., and Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311.

Larsson, J. and Heeger, D. J. (2006). Two retinotopic visual areas in human lateral occipital cortex. *The Journal of Neuroscience*, 26(51):13128–13142. PMID: 17182764.

Lawrence, S., Giles, C., Tsoi, A. C., and Back, A. (1997). Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113.

Layton, O. W., Mingolla, E., and Yazdanbakhsh, A. (2012). Dynamic coding of border-ownership in visual cortex. *Journal of Vision*, 12(13):8–8.

Leder, H. and Bruce, V. (1998). Local and relational aspects of face distinctiveness. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, 51(3):449–473.

Lescroart, M. D. and Biederman, I. (2013). Cortical Representation of Medial Axis Structure. *Cerebral Cortex*, 23(3):629–637.

Li, N. and DiCarlo, J. J. (2008). Unsupervised Natural Experience Rapidly Alters Invariant Object Representation in Visual Cortex. *Science*, 321(5895):1502–1507.

Lindsey, B. G., Morris, K. F., Shannon, R., and Gerstein, G. L. (1997). Repeated Patterns of Distributed Synchrony in Neuronal Assemblies. *Journal of Neurophysiology*, 78(3):1714–1719.

Lisetti, C. L. and Rumelhart, D. E. (1998). Facial Expression Recognition Using a Neural Network. In *FLAIRS Conference*, pages 328–332.

MacLeod, C. (2012). Cognitive bias modification procedures in the management of mental disorders:. *Current Opinion in Psychiatry*, 25(2):114–120.

MacLeod, C. and Cohen, I. L. (1993). Anxiety and the interpretation of ambiguity: A text comprehension study. *Journal of Abnormal Psychology*, 102:238–247.

MacLeod, C. and Mathews, A. (2012). Cognitive Bias Modification Approaches to Anxiety. *Annual Review of Clinical Psychology*, 8(1):189–217.

MacLeod, C., Rutherford, E., Campbell, L., Ebsworthy, G., and Holker, L. (2002). Selective attention and emotional vulnerability: Assessing the causal basis of their association through the experimental manipulation of attentional bias. *Journal of Abnormal Psychology*, 111(1):107–123.

Mangini, M. C. and Biederman, I. (2004). Making the ineffable explicit: estimating the information employed for face classifications. *Cognitive Science*, 28(2):209–226.

Mao, B.-Q., Hamzei-Sichani, F., Aronov, D., Froemke, R. C., and Yuste, R. (2001). Dynamics of Spontaneous Activity in Neocortical Slices. *Neuron*, 32(5):883–898.

Markram, H., Lubke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science (New York, N.Y.)*, 275(5297):213–215.

Martin, A. B. and von der Heydt, R. (2015). Spike Synchrony Reveals Emergence of Proto-Objects in Visual Cortex. *The Journal of Neuroscience*, 35(17):6860–6870.

Martinez, R. and Paugam-Moisy, H. (2009). Algorithms for Structural and Dynamical Polychronous Groups Detection. In Alippi, C., Polycarpou, M., Panayiotou, C., and Ellinas, G., editors, *Artificial Neural Networks - ICANN 2009*, number 5769 in Lecture Notes in Computer Science, pages 75–84. Springer Berlin Heidelberg.

Masquelier, T. and Thorpe, S. J. (2007). Unsupervised Learning of Visual Features through Spike Timing Dependent Plasticity. *PLoS Comput Biol*, 3(2):e31.

Mathews, A. and Mackintosh, B. (2000). Induced emotional interpretation bias and anxiety. *Journal of Abnormal Psychology*, 109:602–615.

Mathews, A. and MacLeod, C. (2005). Cognitive Vulnerability to Emotional Disorders. *Annual Review of Clinical Psychology*, 1(1):167–195.

Maurer, D., Grand, R. L., and Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, 6(6):255–260.

McCormick, D. A., Connors, B. W., Lighthall, J. W., and Prince, D. A. (1985). Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. *Journal of Neurophysiology*, 54(4):782–806.

Morin, E. L., Hadj-Bouziane, F., Stokes, M., Ungerleider, L. G., and Bell, A. H. (2014). Hierarchical encoding of social cues in primate inferior temporal cortex. *Cerebral cortex (New York, N.Y.: 1991)*. PMID: 24836688.

Motter, B. C. (2009). Central V4 Receptive Fields Are Scaled by the V1 Cortical Magnification and Correspond to a Constant Sized Sampling of the V1 Surface. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(18):5749–5757.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics*, 66(3):241–251.

Nikoli, D., Fries, P., and Singer, W. (2013). Gamma oscillations: precise temporal coordination without a metronome. *Trends in Cognitive Sciences*, 17(2):54–55.

Nishimura, H. and Sakai, K. (2004). Determination of border ownership based on the surround context of contrast. *Neurocomputing*, 58-60:843–848.

Olshausen, B. A., Anderson, C. H., and Essen, D. V. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience*, 13(11):4700–4719.

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

Pasupathy, A. (2006). Neural basis of shape representation in the primate brain. *Progress in brain research*, 154:293–313. PMID: 17010719.

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.

Paugam-Moisy, H., Martinez, R., and Bengio, S. (2008). Delay learning and polychronization for reservoir computing. *Neurocomputing*, 71(7-9):1143–1158.

Penton-Voak, I. S., Thomas, J., Gage, S. H., McMurran, M., McDonald, S., and Munafo, M. M. (2013). Increasing recognition of happiness in ambiguous facial expressions reduces anger and aggressive behavior. *Psychological Science*, 24:688–697.

Perrett, D. and Oram, M. (1993). Neurophysiology of shape processing. *Image and Vision Computing*, 11(6):317–333.

Perrett, D. I., Hietanen, J. K., Oram, M. W., and Benson, P. J. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 335(1273):23–30. PMID: 1348133.

Perrett, D. I., Rolls, E. T., and Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Experimental brain research. Experimentelle Hirnforschung. Exprimentation crbrale*, 47(3):329–342. PMID: 7128705.

Perrinet, L., Delorme, A., Samuelides, M., and Thorpe, S. J. (2001). Networks of integrate-and-fire neuron using rank order coding A: How to implement spike time dependent Hebbian plasticity. *Neurocomputing*, 38-40:817–822.

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.

Perry, G., Rolls, E. T., and Stringer, S. M. (2010). Continuous transformation learning of translation invariant representations. *Experimental brain research. Experimentelle Hirnforschung. Experimentation cerebrale*, 204(2):255–270. PMID: 20544186.

Petkov, N. and Kruizinga, P. (1997). Computational models of visual neurons specialised in the detection of periodic and aperiodic oriented visual stimuli: bar and grating cells. *Biological cybernetics*, 76(2):83–96. PMID: 9116079.

Pettet, M. W. and Gilbert, C. D. (1992). Dynamic changes in receptive-field size in cat primary visual cortex. *Proceedings of the National Academy of Sciences*, 89(17):8366–8370. PMID: 1518870.

Pitcher, D., Walsh, V., and Duchaine, B. (2011). The role of the occipital face area in the cortical face perception network. *Experimental Brain Research*, 209(4):481–493.

Preminger, S., Blumenfeld, B., Sagi, D., and Tsodyks, M. (2009). Mapping dynamic memories of gradually changing objects. *Proceedings of the National Academy of Sciences*, 106(13):5371–5376.

Preminger, S., Sagi, D., and Tsodyks, M. (2007). The effects of perceptual history on memory of visual objects. *Vision Research*, 47(7):965–973.

Prut, Y., Vaadia, E., Bergman, H., Haalman, I., Slovin, H., and Abeles, M. (1998). Spatiotemporal Structure of Cortical Activity: Properties and Behavioral Relevance. *Journal of Neurophysiology*, 79(6):2857–2874.

Qiu, F. T., Sugihara, T., and von der Heydt, R. (2007). Figure-ground mechanisms provide structure for selective attention. *Nature Neuroscience*, 10(11):1492–1499.

Quinn, P. C. (2010). The acquisition of expertise as a model for the growth of cognitive structure. *Neoconstructivism: The new science of cognitive development*, pages 252–273.

Quinn, P. C., Yahr, J., Kuhn, A., Slater, A. M., and Pascalis, O. (2002). Representation of the Gender of Human Faces by Infants: A Preference for Female. *Perception*, 31(9):1109–1121.

Renart, A., Parga, N., and Rolls, E. T. (1999). Associative memory properties of multiple cortical modules. *Network: Computation in Neural Systems*, 10(3):237–255.

Rhodes, G. (1997). *Superportraits: Caricatures and Recognition*. Psychology Press, Hove, East Sussex, UK, 1 edition edition.

Richards, A. and French, C. C. (1992). An anxiety-related bias in semantic activation when processing threat/neutral homographs. *The Quarterly Journal of Experimental Psychology A*, 45:503–525.

Richards, A., French, C. C., Calder, A. J., Webb, B., and Fox, R. (2002). Anxiety-related bias in the classification of emotionally ambiguous facial expressions. *Emotion*, 2:273–287.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025.

Rodrguez-Snchez, A. J. and Tsotsos, J. K. (2012). The roles of endstopped and curvature tuned computations in a hierarchical representation of 2D shape. *PloS one*, 7(8):e42058. PMID: 22912683.

Roelfsema, P. R., Lamme, V. A. F., and Spekreijse, H. (2004). Synchrony and covariation of firing rates in the primary visual cortex during contour grouping. *Nature Neuroscience*, 7(9):982–991.

Roiser, J. P., Elliott, R., and Sahakian, B. J. (2012). Cognitive Mechanisms of Treatment in Depression. *Neuropsychopharmacology*, 37(1):117–136.

Rolls, E. (2008). *Memory, Attention, and Decision-Making: A unifying computational neuroscience approach*. Oxford University Press, 1 edition edition.

Rolls, E. T. (2000). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron*, 27(2):205–218. PMID: 10985342.

Rolls, E. T., Aggelopoulos, N. C., and Zheng, F. (2003). The Receptive Fields of Inferior Temporal Cortex Neurons in Natural Scenes. *The Journal of Neuroscience*, 23(1):339–348.

Rolls, E. T. and Baylis, G. C. (1986). Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Experimental Brain Research*, 65:68–48.

Rolls, E. T., Baylis, G. C., and Hasselmo, M. E. (1987). The responses of neurons in the cortex of the superior temporal sulcus of the monkey to bandpass spatial frequency filtered faces. *Vision Research*, 27:311–326.

Rolls, E. T., Baylis, G. C., and Leonard, C. M. (1985). Role of low and high spatial frequencies in the face-selective responses of neurons in the cortex in the superior temporal sulcus. *Vision Research*, 25:1021–1035.

Rolls, E. T., Cowey, A., and Bruce, V. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal coritcal visual areas. *Philosophical Transactions of the Royal Society*, 335(1273):11–21.

Rolls, E. T. and Deco, G. (2002). *Computational Neuroscience of Vision.* Oxford University Press, USA, Oxford, 1 edition.

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(11):2547–2572. PMID: 11110127.

Rolls, E. T. and Treves, A. (1998). *Neural Networks and Brain Function.* Oxford University Press, Oxford, 1 edition.

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. *Journal of computational neuroscience*, 4(4):309–333.

Rosenblatt, F. (1961). *Principles of neurodynamics; perceptrons and the theory of brain mechanisms.* Spartan Books, Washington.

Royer, S. and Paré, D. (2003). Conservation of total synaptic weight through balanced synaptic depression and potentiation. *Nature*, 422(6931):518–522. PMID: 12673250.

Rubin, E. (1915). *Synsoplevede Figurer.* PhD thesis, University of Copenhagen, Copenhagen.

Rugulies, R. (2002). Depression as a predictor for coronary heart disease: a review and meta-analysis1. *American Journal of Preventive Medicine*, 23(1):51–61.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Rumelhart, D. E. and Zipser, D. (1985). Feature Discovery by Competitive Learning*. *Cognitive Science*, 9(1):75–112.

Schneider, S. and Moyer, A. (2010). Depression as a predictor of disease progression and mortality in cancer patients. *Cancer*, 116(13):3304–3304.

Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., and Poggio, T. (2005). *A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex.* MIT CSAIL, MA, USA.

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104(15):6424–6429.

Shafritz, K. M., Gore, J. C., and Marois, R. (2002). The role of the parietal cortex in visual feature binding. *Proceedings of the National Academy of Sciences*, 99(16):10917–10922.

Silver, M. A. and Kastner, S. (2009). Topographic maps in human frontal and parietal cortex. *Trends in cognitive sciences*, 13(11):488–495.

Softky, W. R. (1995). Simple codes versus efficient codes. *Current Opinion in Neurobiology*, 5(2):239–247.

Sormaz, M., Watson, D. M., Smith, W. A. P., Young, A. W., and Andrews, T. J. (2016). Modelling the perceptual similarity of facial expressions from image statistics and neural responses. *NeuroImage*, 129:64–71.

Spoerer, C. J., Eguchi, A., and Stringer, S. M. (2016). A computational exploration of complementary learning mechanisms in the primate ventral visual pathway. *Vision Research*, 119:16–28.

Stork, D. (1989). Is backpropagation biologically plausible? In *, International Joint Conference on Neural Networks, 1989. IJCNN*, pages 241–246 vol.2.

Stringer, S. M., Perry, G., Rolls, E. T., and Proske, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biological Cybernetics*, 94(2):128–142. PMID: 16369795.

Stringer, S. M. and Rolls, E. T. (2008). Learning transform invariant object recognition in the visual system with multiple stimuli present during training. *Neural networks: the official journal of the International Neural Network Society*, 21(7):888–903. PMID: 18440774.

Stringer, S. M., Rolls, E. T., and Tromans, J. M. (2007). Invariant object recognition with trace learning and multiple stimuli present during training. *Network (Bristol, England)*, 18(2):161–187. PMID: 17966074.

Sugihara, T., Qiu, F. T., and von der Heydt, R. (2011). The speed of context integration in the visual cortex. *Journal of Neurophysiology*, 106(1):374–385.

Surcinelli, P., Codispoti, M., Montebarocci, O., Rossi, N., and Baldaro, B. (2006). Facial emotion recognition in trait anxiety. *Anxiety Disorders*, 20:110–117.

Susskind, L. (1995). The world as a hologram. *Journal of Mathematical Physics*, 36(11):6377–6396.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708.

Tanaka, J. W. and Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, 46(2):225–245.

Tanaka, J. W. and Gordon, I. (2011). Features, Configuration, and Holistic Face Processing.

Tanaka, K., Saito, H., Fukada, Y., and Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of neurophysiology*, 66(1):170–189. PMID: 1919665.

Thomas, C. M. and Morris, S. (2003). Cost of depression among adults in England in 2000. *The British Journal of Psychiatry*, 183(6):514–519.

Torralba, A., Murphy, K. P., and Freeman, W. T. (2007). Sharing Visual Features for Multiclass and Multiview Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):854–869.

Tovee, M. J., Rolls, E. T., and Azzopardi, P. (1994). Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque. *Journal of Neurophysiology*, 72:1049–1060.

Trappenberg, T. P., Rolls, E. T., and Stringer, S. M. (2002). Effective size of receptive fields of inferior temporal visual cortex neurons in natural scenes. *Advances in neural information processing systems*, 1:293–300.

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136.

Tromans, J. M. (2012). *Computational neuroscience of natural scene processing in the ventral visual pathway*. Ph.D., University of Oxford.

Tromans, J. M., Harris, M., and Stringer, S. M. (2011). A computational model of the development of separate representations of facial identity and expression in the primate visual system. *PLoS ONE*, 6(10):e25616.

Tromans, J. M., Page, H. J., and Stringer, S. M. (2012). Learning separate visual representations of independently rotating objects. *Network: Computation in Neural Systems*, 23(1-2):1–23.

Troyer, T. W., Krukowski, A. E., Priebe, N. J., and Miller, K. D. (1998). Contrast-invariant orientation tuning in cat visual cortex: thalamocortical input tuning and correlation-based intracortical connectivity. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 18(15):5908–5927.

Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., and Tootell, R. B. H. (2003). Faces and objects in macaque cerebral cortex. *Nature Neuroscience*, 6(9):989–995.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., and Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science (New York, N.Y.)*, 311(5761):670–674. PMID: 16456083 PMCID: PMC2678572.

Tsotsos, J. K. (1993). An Inhibitory Beam for Attentional Selection. In *Proceedings of the 1991 York Conference on Spacial Vision in Humans and Robots*, pages 313–331, New York, NY, USA. Cambridge University Press.

Tsunoda, K., Yamane, Y., Nishizaki, M., and Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience*, 4(8).

Ustn, T. B., Ayuso-Mateos, J. L., Chatterji, S., Mathers, C., and Murray, C. J. L. (2004). Global burden of depressive disorders in the year 2000. *The British Journal of Psychiatry: The Journal of Mental Science*, 184:386–392.

van Rossum, M. C., Bi, G. Q., and Turrigiano, G. G. (2000). Stable Hebbian learning from spike timing-dependent plasticity. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 20(23):8812–8821.

Vanrullen, R. (2007). The power of the feed-forward sweep. *Advances in cognitive psychology / University of Finance and Management in Warsaw*, 3(1-2):167–176. PMID: 20517506.

Vogels, R. and Biederman (2002). Effects of illumination intensity and direction on object coding in the macaque inferior temporal cortex. *Cerebral Cortex*, 12:756–766.

von der Heydt, R., Zhou, H., and Friedman, H. S. (2003). Neural coding of border ownership: Implications for the theory of figure-ground perception. *Perceptual organization in vision: Behavioral and neural perspectives*, pages 281–304.

von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14(2):85–100.

von der Malsburg, C. (1981). The Correlation Theory of Brain Function. Departmental Technical Report, MPI.

von der Malsburg, C. (1999). The What and Why of Binding: The Modeler's Perspective. *Neuron*, 24(1):95–104.

von der Malsburg, C. and Schneider, W. (1986). A neural cocktail-party processor. *Biological Cybernetics*, 54(1):29–40.

Wagatsuma, N., Oki, M., and Sakai, K. (2013). Feature-Based Attention in Early Vision for the Modulation of Figure-Ground Segregation. *Frontiers in Psychology*, 4.

Wallis, G. (2013). Toward a unified model of face and object recognition in the human visual system. *Frontiers in Psychology*, 4:497.

Wallis, G. and Blthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences*, 98(8):4800–4804.

Wallis, G. and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, 51(2):167–194.

Wang, G., Tanifuji, M., and Tanaka, K. (1998). Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. *Neuroscience research*, 32(1):33–46.

Wang, P. S., Simon, G., and Kessler, R. C. (2003). The economic burden of depression and the cost-effectiveness of treatment. *International Journal of Methods in Psychiatric Research*, 12(1):22–33.

Wassle, H., Grunert, U., Rohrenbeck, J., and Boycott, B. B. (1990). Retinal ganglion cell density and cortical magnification factor in the primate. *Vision Research*, 30(11):1897–1911.

Wegrzyn, M., Riehle, M., Labudda, K., Woermann, F., Baumgartner, F., Pollmann, S., Bien, C. G., and Kissler, J. (2015). Investigating the brain basis of facial expression perception using multi-voxel pattern analysis. *Cortex*, 69:131–140.

Wiskott, L. and Sejnowski, T. J. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Computation*, 14(4):715–770.

Wolfe, J. M. and Cave, K. R. (1999). The Psychophysical Evidence for a Binding Problem in Human Vision. *Neuron*, 24(1):11–17.

Xu, X., Biederman, I., and Shah, M. P. (2014). A neurocomputational account of the face configural effect. *Journal of Vision*, 14(8):9.

Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z., and Connor, C. E. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature neuroscience*, 11(11):1352–1360. PMID: 18836443.

Yankouskaya, A., Humphreys, G. W., and Rotshtein, P. (2014). Differential interactions between identity and emotional expression in own and other-race faces: effects of familiarity revealed through redundancy gains. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 40(4):1025–1038.

Yarbus, A. L. (1967). *Eye Movements During Perception of Complex Objects*. Springer US, NY, USA.

Yau, J. M., Pasupathy, A., Brincat, S. L., and Connor, C. E. (2013). Curvature processing dynamics in macaque area v4. *Cerebral Cortex*, 23(1):198–209.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, 81(1):141–145.

Young, M. P. and Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science*, 256(5061):1327–1331. PMID: 1598577.

Yue, X., Tjan, B. S., and Biederman, I. (2006). What makes faces special? *Vision research*, 46(22):3802–3811.

Zeki, S., Aglioti, S., McKeefry, D., and Berlucchi, G. (1999). The neurological basis of conscious color perception in a blind patient. *Proceedings of the National Academy of Sciences of the United States of America*, 96(24):14124–14129.

Zhang, J., Li, X., Song, Y., and Liu, J. (2012). The fusiform face area is engaged in holistic, not parts-based, representation of faces. *PLoS ONE*, 7(7):e40390.

Zhaoping, L. (2005). Border Ownership from Intracortical Interactions in Visual Area V2. *Neuron*, 47(1):143–153.

Zhou, H., Friedman, H. S., and von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *The Journal of Neuroscience*, 20(17):6594–6611.