Computational neuroscience of natural scene processing in the ventral visual pathway

James Matthew Tromans

Submitted for the Degree of DPhil in Experimental Psychology

The University of Oxford

Trinity Term 2012

Acknowledgements

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

Firstly, I would like to thank my supervisor, Dr. Simon Stringer, for his continued encouragement, enthusiasm and contributions. Even when presented with the greatest of challenges, you have remained upbeat and optimistic.

Secondly, I must thank my research colleagues who have provided a constant source of entertainment and academic consultation, in equal measure. I would like to thank Bedeho for his significant contribution to the improvements to the VisNet model; Ben, for his efforts on the Gabor filtering routines; Dan, for his wisdom and advice, having started his DPhil an academic year ahead of me; and Irina, Hector and Mitchell, for being enthusiastic students and satisfying to mentor.

Thirdly, I offer my thanks to the Oxford Foundation for Theoretical Neuroscience and Artificial Intelligence (OFTNAI), Prof. Peter McLeod and Dr. Mark Buckley. In addition, I offer special thanks to Prof. Brian Rogers for having faith in my abilities, and Prof. Edmund Rolls for igniting my interest in computational neuroscience.

Finally, I offer sincere gratitude to my girlfriend, Sarah, who has been both rock steady and understanding even when the inevitable frustration from a failed simulation has fallen upon her shoulders. Last, but certainly not least, I offer immense thanks 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 this thesis without you!

Short Abstract

Computational neuroscience of natural scene processing in the ventral visual pathway
James Matthew Tromans
Pembroke College
Submitted for the Degree of DPhil in Experimental Psychology
Trinity Term 2012

Neural responses in the primate ventral visual system become more complex in the later stages of the pathway. For example, not only do neurons in IT cortex respond to complete objects, they also learn to respond invariantly with respect to the viewing angle of an object and also with respect to the location of an object. These types of neural responses have helped guide past research with VisNet, a computational model of the primate ventral visual pathway that self-organises during learning. In particular, previous research has focussed on presenting to the model one object at a time during training, and has placed emphasis on the transform invariant response properties of the output neurons of the model that consequently develop. This doctoral thesis extends previous VisNet research and investigates the performance of the model with a range of more challenging and ecologically valid training paradigms. For example, when multiple objects are presented to the network during training, or when objects partially occlude one another during training. The different mechanisms that help output neurons to develop object selective, transform invariant responses during learning are proposed and explored. Such mechanisms include the statistical decoupling of objects through multiple object pairings, and the separation of object representations by independent motion. Consideration is also given to the heterogeneous response properties of neurons that develop during learning. For example, although IT neurons demonstrate a number of differing invariances, they also convey spatial information and view specific information about the objects presented on the retina. A updated, scaled-up version of the VisNet model, with a significantly larger retina, is introduced in order to explore these heterogeneous neural response properties.

Long Abstract

Computational neuroscience of natural scene processing in the ventral visual pathway
James Matthew Tromans
Pembroke College
Submitted for the Degree of DPhil in Experimental Psychology
Trinity Term 2012

Neurons in the latter stages of the ventral visual pathway learn to respond to visual stimuli with different types of invariances. For example, it has been shown that neurons within inferotemporal cortex may respond to the same object regardless of viewpoint, while other IT neurons may respond to the same object regardless of retinal location. These response properties have helped guide past research with VisNet, a computational model of the primate ventral visual pathway. VisNet consists of a hierarchical series of four feedforward processing layers. Each layer loosely maps to a different processing stage of the primate ventral visual system, with the output VisNet layer reflecting anterior IT cortex, cytoarchitecture TE.

Previous research with VisNet has focussed on the development of invariant responses in the output layer of the model. These simulations usually comprised the presentation of only one object at a time during learning. There are a number of important and related limitations to this approach. For example, the real primate visual system must learn about individual objects in a potentially complex multi-object environment, and presenting only one object at a time to the VisNet model during training is ecologically unrealistic. In an attempt to create a carefully controlled and systematic study of object segmentation from within a scene, other VisNet simulations have presented one object at a time on a noisy background during training. However, as will be shown, it is possible that this type of training environment may actually make object segmentation and learning more challenging. A real visual scene does not usually consist of just one object on a noisy background, but instead consists of many

objects that together create the complete scene. The approach taken within this doctoral thesis is to present the VisNet model with pairs of objects that together comprise a scene. This avoids considering a scene as the object to be learned, plus a noisy background. In this way, this thesis investigates the computational neuroscience of natural scene processing in the ventral visual pathway, using a more realistic training paradigm than previous VisNet research.

The first research chapter presents results when the VisNet model has simulated a reduced training paradigm. Previous research has suggested that a large, ecologically implausible, amount of training may be necessary in order for VisNet output neurons to develop object specific, transform invariant, representations. The simulations presented as part of this first research chapter reveal an important and interesting finding; the amount of training necessary to produce statistical decoupling between pairs of objects presented during training is relatively small. Only three pairings per object are necessary for the VisNet output neurons to develop object selective, invariant representations, during training. This important finding verifies that the underlying computational principles helping to produce these responses are extremely robust, and could exist in the primate visual system.

Following the theme of ecological validity, the second research chapter investigates the performance of the VisNet model when tasked with learning about individual objects during a partial occlusion training paradigm. In reality, the primate visual system is able to learn about novel objects even when they are partially occluded. Past VisNet research does not explain how the model might learn to solve this problem during training. The results of this chapter show that VisNet is able to use the same principles discussed in the first research chapter to help build separate representations of individual objects during training, even when one object partially occludes the other.

The first two research chapters present various object pairings to the network during training.

This is an important progression from only presenting individual objects, or individual objects on a noisy background, as was common in past VisNet research. The third research chapter of this doctoral thesis explores whether VisNet is able to make use of independent motion when developing separate representations of individual stimuli. In this case, only one pair of objects is presented during simulation. The simulation results show that independent motion, when combined with CT learning, enables VisNet to develop object specific, and transform invariant, object representations.

The penultimate research chapter builds on the simulation results obtained in the third research chapter. More specifically, it explores whether the VisNet model is able to develop separate representations of facial identity and facial expression using similar principles to those developed during the independent motion paradigm. A continuum of forty facial expressions were combined with a continuum of forty facial identities. This created a total of 1600 different faces with which the model was trained. VisNet output neurons learned to respond to a subset of each continuum. For example, after training, some output neurons learned to respond to just the first few facial expressions, irrespective of the facial identity. A different set of output neurons learned to respond to the next contiguous region of facial expressions in a similar manner, such that all facial expressions were represented in this way. Similarly, VisNet output neurons also learned to represent the different facial identities.

The final research chapter implements a substantially increased retina size in a scaled-up version of VisNet. Previous computing limitations have precluded research with a larger VisNet architecture; it was too computationally expensive to simulate the model. Accordingly, this is the first time such a development has been possible. The benefits afforded by the new model, in particular the increase in the retina size, allowed far greater object detail to be processed within the layers of the VisNet model. It was observed that a more heterogeneous population of cells formed in the output layer.

For example, the simulations within this final chapter used a realistic human face stimulus rotating in depth, created using FaceGen software. Within the same network, some output neurons learned to respond to different views of the face stimulus, while other neurons learned to respond invariantly. Furthermore, it is shown how the same output neurons could learn to develop view specific, and also location invariant responses.

The primary focus of this doctoral thesis is to investigate how VisNet, as a biologically plausible model of the primate ventral visual pathway, may learn to recognise individual objects, even when multiple objects are presented during learning. The research provides a foundation for modelling how natural scene processing may occur in the primate ventral visual system and explores different mechanisms that can support learning about individual objects in a multi-object environment. Future research must place emphasis on the heterogeneous response properties of neurons within different layers of the VisNet model. Adjustments to the training paradigms, model architecture and analytical techniques, may be necessary and are discussed where appropriate.

Contents

Chapter 1

Introduction

The primary focus of this thesis is to develop our theoretical understanding of how the visual cortex functions; specifically, how the visual cortex is able to build internal representations of individual objects despite the fact that objects in the environment are rarely seen in isolation. The thesis is further concerned with how invariant representations of individual objects can form and how they are encoded. That is, how does the primate visual cortex learn to recognise the same object from different viewpoints and retinal positions.

The primate visual system is a complex and remarkable network of interacting components that together create our visual perception. The ability to see is something that is seemingly effortless, a process that every healthy human is able to perform, yet visual perception recruits more of the primate cortex than any other sense. Our ability to use sight to guide a wide range of behaviours, from social interactions through to planning and strategy, is particularly impressive and appears deceptively easy; 'seeing' does not seem to be the most difficult aspect when negotiating a busy shopping centre for the most efficient route. To the average person, how we learn to see and the staggering complexities that are involved in just recognising an object from different viewpoints, barely sparks a thought in their

mind. And this is the way it should be. A well designed system should perform its primary functions seamlessly and mask the complexities of its operation through the elegant and consistent results it produces. However, as Richard Gregory famously stated, given the types of inputs that our visual system receives, it is 'nothing short of a miracle' that it is able to perform as it does [1]. Understanding how the visual system functions is undoubtedly a challenging task.

Ongoing research reveals detailed information about the functional and structural properties of the primate visual system and future investigations will provide more anatomical and physiological detail and also psychophysical information. However, new data of this type alone will not be enough; it is a necessity to build new theoretical frameworks and continue developing existing ones. Theoretical models should be developed alongside biological research, each to complement one another. Models can be used to explore, evaluate, and help explain biological findings and thus guide the direction of the future experimental work. In addition, artificial approximations of the biological mechanisms can be developed.

Despite a wealth of different methodologies, improving technologies, and varying approaches to studying vision, the ways in which the primate visual system is able to learn about individual objects in a cluttered, multi-object environment, are still not well understood. Indeed, even more challenging questions remain, such as the specific neural code for how different visual objects in the environment are encoded within the primate cortex to form part of an internal representation.

The core of this research is a continuation of the computational modelling work of Wallis and Rolls (1997) and more recently Stringer and Rolls (2008) in their development and investigation of "VisNet", an artificial neural network model of the ventral visual processing stream of the macaque monkey brain. Each research chapter within this thesis is self contained and independent. However, as

[1] Quotation by Gregory (1997)

will become clear, each chapter may build on the previous by exploring a related problem, or utilising the previous findings. Chapter 2 extends previous work by Stringer et al. (2007) and explores the extent to which training must occur in order for the VisNet model to develop separate representations of individual objects. The research conducted as part of Chapter 3 investigates whether the same model is able to develop a holistic invariant representation of an object that is always partially occluded during learning. The next research chapter touches on a limitation highlighted in previous studies. More specifically, the research in Chapter 4 investigates how independent motion of just two objects can be combined with Continuous Transformation learning to allow the computational model to form separate representations of the objects. Chapter 5 presents research that explores the role that these mechanisms may play in learning to recognise facial expression and identity, as well as a number of limitations. Finally, Chapter 6 investigates how the VisNet model may be able to replicate important experimental results in which certain neurons in inferotemporal cortex were found to be view selective but translation invariant (Wang et al., 1998).

The contribution of this thesis is diverse and sometimes it is necessary to include new introductory material within each chapter. The remainder of this chapter introduces the relevant aspects of the primate ventral visual pathway and provides a review of the existing VisNet research literature. Consideration is given to different theoretical approaches to invariant object recognition. The VisNet paradigm is included as part of the introduction because the fundamental principles are well established in previously published work.

1.1 Object Recognition in the Primate Visual System

As discussed above, primates demonstrate a remarkable and robust ability to develop visual perception with apparent ease. Despite the fact that there is a great deal of variation in the input, the primate visual cortex is able to take retinal output and learn to build internal invariant representations of individual objects in a multi-object cluttered environment. The following sections provide an introduction to the neurophysiology that underpins these abilities.

1.1.1 The Retina

Light enters the eye via the cornea, before passing through a lens to form an image on the retina at the back of the eye. The retina consists of a number of layers. Human retina photoreceptors are positioned at the back of the retina and therefore located behind a number of other cell bodies through which the light must pass before being processed.

There are two primary forms of photoreceptor cells: cones and rods. Cone cells are sensitive to the wavelength of the light (colour), and function best in brighter light. There are three primary types of cone cells which respond to red (long wavelength), green (medium wavelength) and blue (short wavelength). Rod cells are more sensitive to luminosity than cone cells and are responsible for our ability to see in low light, as well as largely being responsible for peripheral vision because they are concentrated on the outer edges of the retina. The human eye contains roughly 100 million rod photoreceptors and 6 million cone photoreceptors. Photoreceptor density varies with respect to the distance from the fovea, the point of central focus on the retina. Visual acuity decreases with eccentricity from the fovea, and receptive field sizes increase accordingly (Bailey et al., 1985).

Photoreceptors project to ganglion cells in the retina via bipolar and amacrine cells. There are ap-

proximately 1.3 million retinal ganglion cells in the human retina (Hecht, 1987), suggesting that each ganglion cell would receive input from up to 100 photoreceptors. However, the level of connectivity is not evenly distributed; it varies considerably and is a function of the location of the ganglion cell in the retina. In the fovea, a ganglion cell will connect with a very small number of photoreceptors, whereas at the periphery of the retina, a cell may receive afferent synapses from thousands of rod and cone cells.

Nonetheless, the receptive fields of retinal ganglion cells are small, ensuring that they are only able to integrate information over a limited part of the visual field. This property reduces the cells' ability to respond constantly when the stimulus shifts over a relatively large spatial area.

Early research investigating retinal ganglion cell receptive field properties showed that, unlike photoreceptor cells, retinal ganglion cells have a receptive field with a specific centre-surround structure (Kuffler, 1953). That is, for each cell there is a central region of the receptive field that, when stimulated with light, will cause the cell to become active and fire. Correspondingly, there is also an outer ring to the receptive field which, when stimulated, will inhibit the cells' firing rate. These types of cells are called 'on-centre' and it should be noted that there are also retinal ganglion cells that have the reverse pattern of connectivity, aptly named 'off-centre' cells. This means that retinal ganglion cells are sensitive to differences in the pattern of light across the retina, not merely to the stimulation of the photoreceptors by light alone.

De Monasterio and Gouras (1975) sampled 460 monkey retinal ganglion cells with receptive fields positioned centrally in the animal's field of view. In particular, they noted their discovery of 'colour-opponent' cells and 'broad-band' cells. These cell types are now more commonly called P and M cells respectively, because they make connections with 'parvocellular' and 'magnocellular' layers

of the dorsal lateral geniculate nuclei (LGN) (Shapley and Perry, 1986). There are a number of differences between these cell types: M cells have significantly higher contrast sensitivity; they have larger receptive fields (Croner and Kaplan, 1995) and therefore relatively fewer M cells are required to cover the retinal surface (Perry et al., 1984); they have thicker axons and so conduct impulses more quickly; they produce a phasic response if presented with a stimulus that does not change. By contrast, P cells have a lower contrast sensitivity; they have smaller receptive fields; they have slow conduction velocity; they produce a tonic response to a non-changing stimulus. The receptive fields of both cell types increase in size with eccentricity from the fovea.

Since these original findings, we now understand that not all retinal ganglion cells fall into these two groups. Hendry and Reid (2000) discovered that some (bistratified) retinal ganglion cells form a third pathway, known as the koniocellular (or K) pathway. Only about 10% of retinal ganglion cells are bistratified and they have very large receptive fields with no surround, only responding to blue cone cells. These retinal ganglion cells project to koniocellular layers, located between the magnocellular and parvocellular layers of the LGN. This neural connectivity allows the retina to perform linear spatial and temporal filtering of light intensity patterns. However, supporting studies with non-primate species suggest that the retina may be able to perform far more complicated calculations. Research performed using different types of fish retina shows that retinal ganglion cells may service more complicated multiple gain control mechanisms, which may help to maintain sensitivity to small changes in light intensity over time and space (Sakai et al., 1995). In addition, Smirnakis et al. (1997) showed that retinal ganglion cells of the salamander display adaptation to a mean level of contrast maintained over a relatively long period of time. More recently, Schwartz and Berry (2008) discovered that the retina is capable of highly sophisticated temporal pattern recognition. Soo et al. (2011) have shown

that retinal ganglion cell receptive field responses do not necessarily follow a Gaussian profile, and that the cells are able to encode 33% more information than predicted by a Gaussian receptive field model. These recent advances in our understanding of the receptive field properties of retinal ganglion cells demonstrate that we are still developing our knowledge in this area.

Despite the fact that the retina is arguably one of the most well understood parts of the visual processing system, there is still a growing body of research investigating the properties of cell responses of the retina in various species. The retina transforms varying patterns of light into neural activity in retinal ganglion cells, which in turn provide stabilised and normalised output represented by action potentials that travel along axons which together form the optic nerve. The majority of retinal ganglion cells project to the visual cortex via the dorsal Lateral Geniculate Nucleus (LGN) and to the superior colliculi. A discussion of these projections, and the site of their terminations, follows.

1.1.2 The Lateral Geniculate Nucleus

The dorsal part of the Lateral Geniculate Nucleus (LGN) in the thalamus relays visual information, received from the ganglion cells in the retina, to the primary visual cortex.

Both the left and right hemispheres of the brain contain an LGN. Each LGN receives information from both eyes, but only from one half of the visual field. In animals with forward facing eyes and binocular overlap, there is a partial crossing over, or decussation, of the optic nerves at the optic chiasm. The axons of ganglion cells in the left half of each retina (carrying information about the right half of the visual field) project to the left LGN, while the axons of ganglion cells in the right half of each retina (carrying information about the left half of the visual field) project to the right LGN. Although this arrangement creates significant overlap of visual fields between the eyes and, therefore, produces a relatively high level of redundancy, it is very common among predators because it enables

a better perception of the depth of field of view, necessary for hunting.

The LGN is arranged into six three-dimensional layers where each layer contains a retinotopic map of half of the visual field. The visual field is mapped onto the surface of the LGN so that neighbouring cells have receptive fields that are also nearby in the visual world. Cells in the LGN are uniformly distributed and thus the foveal regions of the retina project onto a relatively large part of the LGN. Cells in layers 2,3, and 5 of the LGN receive their inputs from the ipsilateral eye while layers 1, 4, and 6 receive input from the contralateral eye. This connectivity means that images of the same object formed on the right and the left retinas can be processed together in the same part of the brain. Schroeder et al. (1990) found that there are binocular interactions between layers, suggesting that visual processing in the LGN is more than a simple relay of monocular visual responses generated in the retina.

The retinal ganglion cells primarily project to three different types of cells in the LGN. The first two layers of the LGN are called magnocellular layers. The remaining four superior layers are called the parvocellular layers, and comprise smaller cells. Most M retinal ganglion cell axons project to the magnocellular layers, and the rest to the superior colliculi, while all P ganglion cells project to the parvocellular layers (Leventhal et al., 1981). In the parvocellular layers, cells have colour-opponent receptive fields: some are excited by red light and inhibited by green (and vice versa) and others are excited by blue light and inhibited by yellow (and vice versa). In addition, a population of 'K' cells projects to six koniocellular layers, each one positioned just below one of the magno- or parvocellular layers, and containing cells even smaller than the P cells (Hendry and Reid, 2000). If an electrode were gradually inserted through all six layers of the LGN, the cells in this column would be found to respond to the same point in visual space (Hubel and Wiesel, 1961).

The responses of LGN cells generally reflect the responses of retinal ganglion cells to which they are connected. The differences between M and P ganglion cells in the retina are mirrored by the properties of M and P cells in the LGN (Derrington and Lennie, 1984), supporting the view that the different magno- and parvocellular (and koniocellular) pathways are functionally distinct. Lesion studies have been performed to try to identify these functions. Lesions of the P layers suggest that they support colour, texture, and pattern discrimination (Schiller et al., 1990) while M layers support rapid visual motion (Merigan and Maunsell, 1990). However, lesions of the P layers, but not of the M layers, caused a deficit in detecting contrast even though further research has shown that M cells have higher contrast sensitivity than P cells. One explanation could be that there are more P cells than M cells in any given area of the retina, and therefore this larger volume of P cells could convey greater contrast sensitivity than the total M population. For a more detailed overview of the distinct contributions of the magnocellular and parvocellular visual streams, see Denison and Silver (2011).

In addition to processing information from the retinal ganglion cells, the LGN also receives a significant number of corticofugal connections. In fact, most of the afferent projections to the LGN are from the visual cerebral cortex. These connections are argued to be an integral part of the neuronal circuitry responsible for analysing the visual image (Murphy et al., 1999). For example, although the LGN cells will respond to stimuli at any orientation, Sillito et al. (1993) found that feedback from the visual cortex makes them sensitive to orientation context. Suffice it to say, even at this relatively early stage in the visual processing system, the LGN is heavily organising and regulating the information that flows through it and is part of a complex feedback system with the visual cortex.

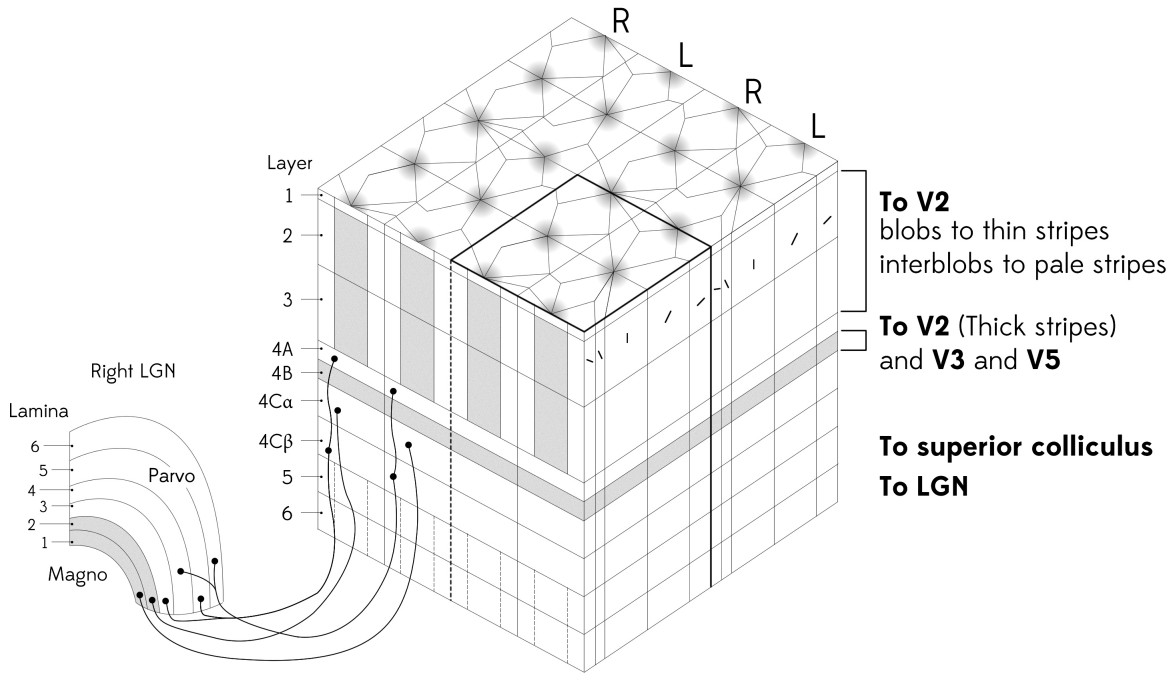

Figure 1.1: LGN and primary visual cortex (V1). It can be seen that V1 is divided into six distinct layers. As depicted in the figure, the fourth layer receives most visual input from the LGN. This layer can be further subdivided where sublamina 4Cα receives most magnocellular input from the LGN, while layer 4Cβ receives parvocellular input. Figure adapted from Levine (2000).

1.1.3 Early Stage Cortical Processing in V1

Located in the occipital lobe at the back of the brain, the primary visual cortex (V1) comprises six functionally distinct areas and together these are anatomically equivalent to Brodmann area 17. It is also known as striate cortex due to its appearance under a microscope, which is caused by afferent myelinated axons from the LGN, terminating in layer 4 of V1. Figure 1.1 shows a diagram of both the LGN and example V1 cortex of the adult monkey.

The fourth layer of V1 is subdivided into 4A, 4B, 4Cα, and 4Cβ. Sublamina 4Cα receives most magnocellular input from the LGN, while layer 4Cβ receives input from parvocellular pathways. 4Cβ ultimately projects to other areas of the cortex via V1 layers 2 and 3. Sublamina 4Cα is further

connected to 4B and from there to extrastriate visual areas. The efferent koniocellular connections terminate in layers 1 and 3. Neurons within V1 form a more complex network than those observed in the retina. The number of different cells recorded and the nature of their connections, both within layers and between them, serves to highlight the increasing complexity of processing as information flows through the visual system (Callaway, 1998).

The primary visual cortex has around 200 million cells, compared with 1.5 million cells in the LGN. The receptive fields of the LGN cells have a similar organisation to the retinal ganglion cells that feed them, but V1 cell responses are very different. Single cell studies in V1 were successfully conducted in cats (Hubel and Wiesel, 1959, 1962) and later in monkeys (Hubel and Wiesel, 1968). The different types of receptive fields observed in V1 are numerous: some have concentric receptive fields like those in the LGN; some have side-by-side parallel excitatory and inhibitory areas; some have three or more side-by-side parallel antagonistic areas.

Hubel and Wiesel discovered that, at their most basic, cells with these receptive fields perform a simple linear spatial summation of light intensity. Due to their elongated shape and antagonistic areas, these cells respond more when the stimulus, a line or bar, is orientated along the length of the receptive field, and less when it is orientated across the antagonistic areas. This orientation selectivity and associated tuning profile is a property not found in the retina or LGN. Hubel and Wiesel referred to these cells as 'simple cells'. See Ferster and Miller (2000) for a review on neural mechanisms of orientation selectivity in the visual cortex.

Another type of cell observed in the visual cortex by Hubel and Wiesel was classified as a 'complex cell'. Complex cells do have similar qualities to simple cells, but differ with respect to their phase (translation) invariance properties. That is, complex cells will produce a response regardless of

where the stimuli lies within the receptive field provided it is at the correct orientation. If one 'drifts' a grating across the receptive field of a simple cell, its response will rise and fall, whereas with a complex cell, it will remain constant (De Valois et al., 1982). Complex cells receive input from a number of simple cells, emphasising the hierarchical nature of the visual processing system. Even more sophisticated cells (hypercomplex) respond significantly when a stimulus is a 'corner' moving in a preferred direction across the receptive field. Importantly, they will not respond if the stimulus is too long and as such, these types of cells are referred to as 'end-stopped'

In addition to preserving the parvo- and magnocellular efferent pathways from the LGN in V1, described above, the retinotopic map from the LGN is also maintained in V1. Cells in a particular location on the surface of V1 can be directly mapped to a location on the retina. As in the LGN, cells in and around the fovea are mapped onto relatively more of the cortex to account for the higher density of retinal ganglion cells, a process known as cortical magnification. Some cells in V1 will only respond to one eye, while others respond to a stimulus when it is presented to either eye. In the former case, these cells often demonstrate 'occular dominance', that is, they respond preferentially to a stimulus presented to one eye compared to when it is presented to the other. The organisation of occular dominance columns is such that they form alternating bands of right or left eye preference across the primary visual cortex. Evidence for these findings is presented next.

In their studies, Hubel and Wiesel showed that when recording from an electrode perpendicular to the cortex surface, cells demonstrate a preference for the same orientated stimuli. However, when an electrode was inserted obliquely through a relatively short distance, the cells' selectivity to orientation gradually and systematically changed. Over larger distances, there are breaks in this type of continuity. With the advent of new technology, Obermayer and Blasdel (1993) performed an optical imaging

study which imaged the pattern of cortical activity in response to stimuli comprising gratings of different orientations presented to each eye separately. This confirmed a highly organised and systematic functional architecture in the monkey cortex.

Through the use of a staining technique with cytochrome oxidase, an enzyme involved in the energy system of cells, Livingstone and Hubel (1988) revealed that there are 'blobs' of cells that are almost exclusively devoted to processing colour (including shades of grey) and not orientation. These blob regions receive parvocellular input as opposed to magnocellular input. In addition, there are 'inter-blob' regions that are selective for orientation but insensitive to colour. Specifically, 'inter-blob' regions also receive parvocellular input but are sensitive to boundaries, including boundaries of colour, rather than colour itself. This subdivision of the parvocellular pathway would suggest that there are three partially separated neural circuits: a steropsis and motion pathway (magno V1 neurons); a colour pathway (parvo 'blob' V1 neurons); and a form pathway (parvo 'inter-blob' V1 neurons).

But what does the complex organisation of V1 cells achieve in terms of processing information derived from light to build visual representations of whole objects and scenes which ultimately guide movement and behaviour? Neurons in V1 have been described as 'feature detectors' (Barlow, 1972) and this approach led to the theory that V1 cells may act as spatiotemporal filters. Specifically, at the retinal level, ganglion cells respond to a broad range of filters as measured by bandwidth and as such may be referred to as 'low-pass' filters, while cells in V1 respond to more narrow bandwidths and are referred to as 'baseband' filters. V1 simple cell receptive field profiles can be modelled by multiplying a sinusoidal function by a Gaussian function to create a Gabor filter (Jones and Palmer, 1987). Similarly, a difference of two Gaussian functions also produces a suitable fit (Hawken and Parker, 1987;

Parker, 1996) and is described later within this chapter as part of the VisNet model. Work by Bell and Sejnowski (1997) supports the view that natural scenes can be represented using an 'alphabet' of small local patches of light intensities, which in turn remove redundancy in the information received via the feedforward efferent connections from the LGN.

It becomes clear that as information flows through the visual system, it undergoes increasingly complex processing at each additional stage. Although the view is well established that V1 cells may act as spatiotemporal filters driven by feedforward connectivity from the LGN, there is also a role played by lateral and feedback connections from higher cortical areas. Lamme and Roelfsema (2000) show that lateral and feedback connections allow neurons to encode different information over time. Specifically, the responses of V1 cells are modulated by information occurring outside of their receptive fields. This effect must rely on connections other than those that are feedforward.

This complex reorganisation by V1 of input from the LGN is still an early step in the complete visual process. From here, there are two primary streams, dorsal and ventral, into extrastriate cortical areas (Ungerleider et al., 1982). This thesis is largely concerned with object recognition in the ventral visual pathway, which extends from V1 to V2 $\rightarrow$ V4 $\rightarrow$ posterior IT cortex (TEO) $\rightarrow$ anterior IT cortex (TE). As such, a discussion of V2 follows.

1.1.4 Cortical Processing Beyond V1

The efferent connections of the primary visual cortex (V1 or striate cortex) project primarily to V2, which is the second major area of the visual cortex. V2 is part of the extrastriate cortex (Brodmann Areas 18 and 19), which also includes other cortical areas such as V3, V4 and V5/MT.

The structure of V2 was documented by Tootell et al. (1983) who used a number of different markers, including cytochrome oxidase, to reveal stripes running parallel to the cortical surface. These

stripes were labelled ‘thick’, ‘thin’ and ‘pale’. Subsequent research investigated the connectivity from V1 to V2, revealing that thick stripes, thin stripes and pale stripes receive input from V1 layer 4B, V1 ‘blobs’, and V1 ‘interblobs’ respectively (Figure 1.1) (Livingstone and Hubel, 1988). Each stripe performs specialised processing for stereoptic depth (thick stripe), colour (thin stripe), and form (pale stripe). Direct projections from the magnocellular pathway bypass V2 by synapsing directly onto neurons in the medial temporal (MT) area, or V5 in humans. This area is specialised for processing motion (Hess et al., 1989).

Anzai et al. (2007) have shown that macaque V2 neuronal responses build on those of V1. In their study, they examined the structure of receptive fields within area V2 by recording their responses to different combinations of orientated bars and edges. They showed that 70% of neurons have similar orientation tuning throughout their receptive fields, but others contain subregions tuned to different orientations, roughly 90° apart. Some receptive fields showed moderate and gradual shifts in preferred orientation and some demonstrated bimodal responses, that is, to two different orientations. Furthermore, they demonstrated selectivity for combinations of orientations that would be helpful for analysing contours and textures.

This functional segregation has led to a theory of two streams of visual processing: the dorsal stream and ventral stream (Ungerleider et al., 1982; Ungerleider and Haxby, 1994). The popularity of this theory is widespread and may in part be attributed to its simplicity. Goodale and Milner (1992) provided an intriguing dissociation between the dorsal and ventral streams whereby a subject, who had extensive damage to the temporal lobe, demonstrated a severe impairment in object shape recognition, but did not suffer from an inability to interact with the same objects and also showed robust performance at visuomotor tasks. These observations caused Goodale and Milner to conclude

that the ventral pathway is required for conscious awareness about objects while the dorsal pathway supports 'vision for action'.

However, it is very well documented that there is a vast number of connections between the dorsal and ventral streams of the visual system suggesting that a strict two stream hypothesis of visual perception is an over simplification (Martin, 1992). Nonetheless, as highlighted above, the functional segregation provided by the selective cell response properties of neurons in the retina, LGN and V1 is somewhat maintained in V2. Connections carrying depth and motion information predominantly project to V3 and MT/V5 (dorsal stream), while connections conveying information about form and colour predominantly project to V4 (ventral stream).

This thesis is concerned with how the primate visual system learns to develop separate representations of objects in cluttered scenes. Object recognition is subserved by the ventral visual pathway and as such, the dorsal pathway is given no further consideration within this thesis. With this in mind, extrastriate area V4 of the macaque monkey is considered next.

Area V4 comprises at least four subregions and is anatomically located anterior to V2 and posterior to TEO. Cells within V4 show a number of preferential sensitivities, including colour (Zeki, 1973) and orientation (Desimone and Schein, 1987). The area is highly convergent, receiving large amounts of afferent synapses from a number of cortical regions. This makes it well suited for processing object identity and it is widely reported that this area is modulated by attentional mechanisms. For example, receptive fields of V4 neurons shrink around an attended stimuli (Moran and Desimone, 1985; Reynolds et al., 1999). Even if attention is directed to an area outside of the receptive field, the response profile of the cell is shifted towards the attended location space (Connor et al., 1997).

The response properties of cells in V4 tend to reflect the area's intermediate position within the

ventral visual stream. One example of this is the receptive field size of V4 neurons which are around five times the size of those found in V1, but are smaller than those recorded from neurons in the inferior temporal cortex (Desimone and Schein, 1987; Kobatake and Tanaka, 1994). Some of the cell response properties of V4 neurons are similar to those in V2. For example, using optical recording Ghose and Ts'O (1997) found orientation-specific columns in V4 comparable in size and arrangement to those in V2 (for a summary, including original research on the organisation of orientation selectivity throughout the macaque visual cortex, see Vanduffel et al. (2002)). However, one may expect to find neurons in V4 that demonstrate sensitivity to more complex form given that processing appears more sophisticated at each functional stage of the visual system.

Pasupathy and Connor have performed a series of studies investigating the population code for shape in area V4. Using specially designed parameterised 'blob' stimuli, they found that neurons in this area are also tuned for complex patterns including curvature, as well as orientation. More specifically, they found that neurons encode 'object-relative position of boundary fragments' such as 'concavity on the right' (Pasupathy and Connor, 2002). Different neurons in V4 preferred different curvatures and positions, suggesting that a population of V4 neurons can encode whole stimuli in this way. For a review, see Pasupathy (2006).

However, V4 neurons still respond to simple orientated stimuli, and studies reveal that some V1 cells respond to more complex combinations of bars in certain arrangements than was originally thought. Also, there are feedback connections at every step of the visual system with the only exception being to the retina. Nonetheless, evidence would suggest that the ventral visual pathway is strongly feedforward in nature, providing a hierarchy of increasingly complex representations, the sophistication of which can be attributed to the stage of processing along the pathway. Further pro-

cessing is still required until representations of whole objects or faces can be achieved, and as such, the inferior temporal cortex (IT) is considered next.

1.1.5 Higher Level Visual Processing in the Inferior Temporal Cortex

The inferior temporal (IT) cortex is part of the temporal lobe, and includes the inferotemporal gyrus and the lower bank of the superior temporal sulcus. It receives significant visual input primarily from V4, but also input from other brain regions; efferent synapses from the IT cortex project to the limbic system (Amaral and Price, 1984; Yukie and Iwai, 1988) and frontal cortex (Barbas, 1988). The majority of experimental studies investigating the role of IT cortex have been performed with (rhesus macaque) monkeys, and unless otherwise stated, this is the case in the studies presented below. More recent studies have used fMRI to investigate the role of IT cortex in human primate visual processing, and this is made clear where appropriate.

A distinction is often made between the posterior (TEO) and anterior (TE) parts of the IT cortex (Baylis et al., 1987); area TEO is thought to come before area TE in the context of a hierarchical feedforward ventral visual stream. Figure 1.2 shows the approximate locations of the different ventral visual cortical areas. It may be noted that TE may be further subdivided, but for the purposes of this section these areas will be grouped together unless otherwise stated.

As a whole, IT cortex is argued to be the last exclusively visual processing area, completing the ventral visual pathway (Ungerleider et al., 1982). There is clear evidence for top-down influences upon lower level processing (Gilbert and Sigman, 2007). However, the traditional view that the ventral visual system is organised into a feedforward feature hierarchy is still appealing and evidence for this organisation is presented next. In particular, the following section is concerned with understanding the role of IT cortex in the visual system, providing special consideration for IT neuron response

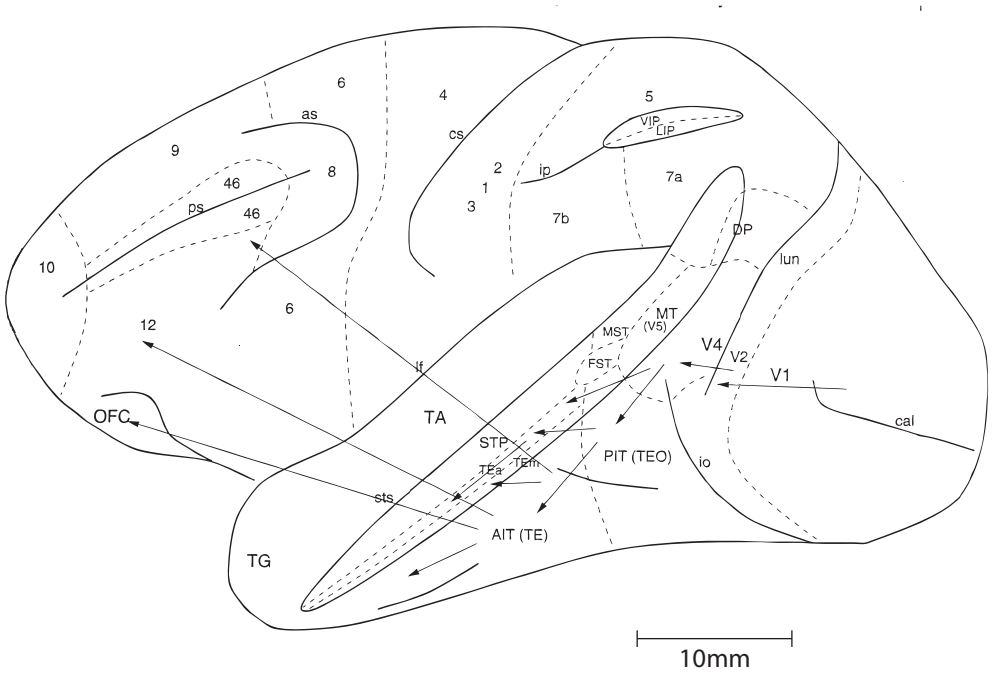

Figure 1.2: Lateral view of the macaque brain. The figure emphasises the connections in what is considered to comprise the ventral visual pathway, from V1 to V2, V4, the inferior temporal visual cortex etc. Figure adapted from Rolls and Deco (2002).

properties and their receptive fields.

Over the last decade, numerous advances have been made in understanding the general role of IT cortex in visual processing. It is well established that lesions to IT cortex impair the ability to discriminate visual shapes (Mishkin and Pribram, 1954; Dean, 1976). The removal of IT cortex from both hemispheres of the monkey causes a deficit in previously learned discrimination tasks exclusively in the visual domain, as well as long-lasting deficits in learning new visual shapes. Also, in humans, lesions to the temporal lobe can cause various forms of visual agnosias. One such example is prosopagnosia: the inability to recognise the faces of familiar people (Damasio et al., 1990) despite relatively intact recognition of other objects. However, a significant difficulty when interpreting human lesion studies is that they are uncontrolled, so this must be taken into consideration when interpreting the results. Nevertheless, lesion studies do confirm that IT cortex is fundamental for object recognition.

The receptive fields of neurons in IT cortex often include the foveal region of the retina (Gross et al., 1972). Specifically, TEO neurons usually have large receptive fields representing the contralateral visual field and still exhibit a weak level of retinotopic organisation (Blumberg and Kreiman, 2010). On the other hand, TE neurons have binocular receptive fields and do not appear to be retinotopically organised (Gross et al., 1972). Although large, the exact size of receptive fields in TE may vary from 2° (DiCarlo and Maunsell, 2003) up to 11° (Rolls, 1991). It is worth noting that in experiments where objects are presented on a completely blank background, receptive field sizes can be as large as 39° when one object is presented at a time, and still as large as 24° for an attended stimulus when two objects are presented simultaneously.

Two possible explanations for this apparent variation in receptive field size include the manner in

which the stimuli are presented to the monkey, as well as the size of the stimuli used. For example, as mentioned above, it has been shown that receptive fields become smaller when the preferred stimulus is presented on a cluttered 'natural' scene, when compared to being presented on a blank background (Rolls et al., 2003). Also, it would appear that studies using the largest stimuli also reveal the largest receptive fields (Tovee et al., 1994). It is possible that the receptive fields of each stimulus inhibit one another, perhaps serviced by lateral (or feedback) inhibitory competition (Wang et al., 2002; Aggelopoulos and Rolls, 2005).

The cell response properties of IT neurons continue to support a feature hierarchy of increasingly complex representations in the ventral visual pathway. Neurons in IT cortex respond preferentially to relatively complex stimuli, responding more strongly to some shapes than to others (Tanaka et al., 1991; Gross et al., 1972). Some neurons recorded in TE actually show selectivity to complex stimuli because they fail to respond to simple orientated bars or edges and only respond to combinations of features (Kobatake and Tanaka, 1994).

Furthermore, in order to better understand IT neuronal responses to complex objects, Brincat and Connor (2004) extended the V4 'blob stimuli' study conducted by Pasupathy and Connor (2002). Specifically, they recorded from TEO and posterior region of TE when the monkey was presented with similar stimuli to those used in the original V4 study. They discovered that the recorded neurons responded best to approximately three (but up to six) combinations of curvatures present as part of the stimulus. It was concluded that representations of retinal input in IT are produced by combining curved edges in particular spatial combinations in order to define part of a larger object boundary.

IT neurons also respond to biologically relevant stimuli, such as hands and faces (Desimone et al., 1984; Rolls, 1984). In some cases, neurons in the lower bank of the superior temporal sulcus respond

exclusively to faces (Tsao et al., 2006). Evidence from single cell recording in the human brain shows that IT cells can demonstrate similar preferences (Quiroga et al., 2005).

It may be argued that IT cortex comprises two main types of neuronal cell response properties: neurons that respond to complex visual features (but usually not whole objects), and neurons that respond to behaviourally relevant features especially hands and faces (Gross et al., 1972; Desimone et al., 1984; Perrett et al., 1984; Tanaka et al., 1991; Kobatake and Tanaka, 1994). Recent research suggests that there may be a further subdivision whereby object, face, and facial expression are processed through independent neural mechanisms (Hadj-Bouziane et al., 2008).

It has been shown that some cell responses in IT are topographically organised (Tanaka, 1993b, 1996) and it has been proposed that these responses may form part of a 'visual alphabet' (Tanaka, 1993a). An early systematic study of area TE revealed that there is a columnar organisation (Fujita et al., 1992). When an electrode was inserted at right-angles (normally) into the cortex, it was found that neurons over a distance of 0.7 to 2.1 mm responded well to stimuli which were identical or similar to the optimal stimulus for the first cell tested in each penetration. When an electrode was inserted at a oblique angle, neurons within this penetration showed similar visual preferences across 0.2 to 0.7mm but then stopped responding. Similar stimulus preferences were shown once again after a gap of around 0.4 to 1mm, thus indicating that stimulus preference was vertically elongated across cortical depth and was patchy across the cortical surface. In this study, a 'stimulus simplification technique' was used. This technique identifies the simplest visual features to which a cell responds and then generates a stimulus set to include optimal, suboptimal, and inefficient stimuli for that cell. Although useful, this method is subjective as it relies upon the researcher's choice of visual features to which it is thought the cell might respond. More recently, similar results have shown a columnar organisation

in IT cortex without using stimulus simplification (Sato et al., 2009); instead, similarity in object selectivity of nearby cells was investigated.

The finding that cells in IT cortex are highly organised is compatible with results from Logothetis et al. (1995), who concluded that each neuron was tuned to a slightly different aspect of the object. It is well established that IT neurons demonstrate significant neuroplasticity, learning to respond to novel stimuli after training. Logothetis et al. (1995) showed that IT neurons can be trained to recognise and classify different computer generated 3D shapes. Specifically, a population of IT neurons responded selectively to views of these previously unfamiliar novel objects. Their results suggest that, after training a monkey to discriminate and recognise different 3D objects, a population of IT neurons can develop a complex receptive field organisation. Although the cell response properties and organisational structure of IT cortex shows that different neurons respond preferentially to different shapes, the authors proposed that a population of neurons may be necessary to encode whole objects.

Rolls et al. (1997) measured the amount of information available from a population of fourteen IT neurons in response to twenty monkey and human faces. They found that individual cells carried a small amount of information about the faces presented. As each additional cell was incorporated into the analysis, the amount of information grew linearly. However, the total number of stimuli that can be encoded in such a system grows exponentially as a function of the number of cells in the population. The authors concluded that representations in IT are 'sparse' and 'distributed'; neurons respond to relatively few stimuli, and stimuli are encoded by a population of neurons. For a review, see Rolls and Treves (2011).

Supporting evidence is provided by Hung et al. (2005) who showed that it is possible to derive information regarding position and scale from a population of cells recorded in IT cortex (100 cells,

randomly sampled). This emphasises that position information does not have to be encoded by individual neurons; it may be encoded across a larger population instead. Furthermore, it was shown that information about both object 'identity' and 'category' was also present. Local field potentials (LFP) also support the view that neurons in IT cortex take part in population encoding and are spatially organised (Kreiman et al., 2006). An LFP constitutes an ensemble measure that pools information over a relatively large cortical region, and is well correlated to the average of individual spikes over a similar sized area of the same cortex. The authors demonstrate that LFPs show strong stimulus selectivity and are tolerant to changes in position or size of an effective stimulus, reinforcing the view that IT neurons are partially invariant to their effective stimuli.

It is widely reported that a number of neurons in IT cortex respond with a significant level of tolerance to various changes in an object's appearance. A central component of this doctoral thesis is to investigate how the ventral visual pathway learns to develop highly tolerant (and invariant) representations of objects in natural scenes. Therefore, a review of the invariant cell response properties of neurons in the IT cortex is presented next.

When a cell responds selectively to the same object over a number of different views (be they rotational depth or retinal position), it is often referred to as an invariant response. However, a distinction should be made between neurons that respond selectively to their preferred object, firing equally to all views, and neurons that still respond selectively to their preferred object, but do not maintain the same level of response amplitude for every view. The former is true invariance while the latter is termed tolerance. The computational studies within this doctoral thesis refer to invariance as it is defined above, i.e. maximal firing rates to all views, unless otherwise stated.

A number of IT neurons maintain object selectivity despite changes in rotational depth (view-

point) (Logothetis and Sheinberg, 1996) and size (Rolls and Baylis, 1986; Ito et al., 1995; Brincat and Connor, 2004; Schwartz et al., 1983). Other neurons respond consistently to faces and hands despite changes in retinal position (Desimone et al., 1984), and comparable results have been demonstrated for non-face objects (Ito et al., 1995; Rolls et al., 2003). Also, Sry et al. (1993) have shown that IT neurons demonstrate tolerance to luminance, motion and texture without altering the shape selectivity of the cells. Translation (position) and viewpoint (rotational depth) invariances are particularly central to this doctoral thesis, so these are discussed next in further detail.

Experimental evidence has demonstrated that there are position invariant neurons. Some neurons respond invariantly to their preferred stimulus despite significant changes of the location of the stimulus within the receptive field (Ito et al., 1995). However, subsequent research has shown that neurons can also be sensitive to the location of the preferred stimulus within the receptive field, displaying position tolerance. In some cases, these neurons stopped responding altogether (DiCarlo and Maunsell, 2003). Both studies claim to have recorded from the same anterior part of the monkey IT cortex, area TE. As mentioned above, stimulus size directly influences receptive field size (Tovee et al., 1994; Rolls et al., 2003), and this may contribute to differences between the results. Regardless, there would appear to be a heterogeneous population of cells with some cells responding to individual views of an object, others responding to small contiguous regions and some responding to all views.

Op De Beeck and Vogels (2000) performed a systematic study of IT neuronal response profiles from three monkeys during exposure to computer-generated images of different sizes. The authors were particularly interested in the position invariant responses of these neurons, or lack of responses. They reported a wide range of receptive field sizes, but in most cases neurons demonstrated a peak activation when the preferred stimulus was presented in a certain location within its receptive field.

Neurons usually showed a Gaussian like decline in activation when the stimulus was moved away from the 'hot spot'. This would suggest that IT neurons can encode spatial information even in larger receptive fields based on their level of activation. The large variation in receptive field size may have been due to the four different object sizes presented to the monkeys during the experiment.

As mentioned above, Hung et al. (2005) were able to show that position information can also be encoded by a population of neurons. Additional studies suggest that IT neuronal populations encode the relative positions of several objects presented simultaneously in a scene. Aggelopoulos and Rolls (2005) demonstrated that different neurons have different shape receptive fields and furthermore, sometimes respond optimally when the effective object is not at the fovea. It was shown that these asymmetric receptive fields had different responses when the stimuli were presented at different locations around the fovea. Although some neurons do respond with a high degree of position tolerance and invariance, there is strong evidence supporting the fact that position information is present in the population response of neurons in IT.

An element of invariance to rotational depth (viewpoint) is necessary for recognising the same object. An object varies not only with position, but also with the angle at which it is observed. A number of studies have shown that IT cells respond to consecutive viewpoints of the same object (Logothetis et al., 1995), or face (Hasselmo et al., 1989b), when it is rotating in depth. Specifically, it has been demonstrated that monkeys exhibit view-dependent recognition to wire-frame objects and perform worse when viewpoint was rotated away from the familiar views. However, after continued exposure they learned to respond to different rotations, eventually producing view-invariant responses (Logothetis et al., 1995).

Complementary research has also been performed with faces. Freiwald and Tsao (2010) used

fMRI to investigate face selective areas in the macaque monkey. They built on existing research by Tsao et al. (2003) and later Tsao et al. (2008), who showed that monkeys and humans share similar neural architecture for visual object processing, especially for faces. More specifically, Tsao et al. (2003) showed that face selective cells are not just scattered randomly throughout IT cortex, but are in fact "embedded within a large swath of object-selective cortex extending from V4 to rostral TE". Tsao et al. (2008) went on to show that macaques appear to have six regions of face-selective cortex arranged along the temporal lobe. Humans appeared to have an additional three areas specialised for faces. The results presented by Freiwald and Tsao (2010) reveal that the two middle face response patches of IT cortex, middle lateral (ML), and middle fundus (MF), produce functionally different response properties compared to the two anterior patches, anterior lateral (AL) and anterior medial (AM). Neurons in the middle face patches were view specific. However, the anterior patch neurons achieved partial and almost full invariance; AL neurons were tuned to identity across a number of contiguous views (and responded equally well to the reflected viewpoint) while neurons in AM, the most anterior face patch, achieved almost full view invariance.

1.1.6 Summary

In summary, there are a number of specialised brain regions within IT cortex, specifically for biologically relevant stimuli such as faces. Neuronal response properties within these areas are heterogeneous; some neurons respond to specific views of their preferred stimulus, while others respond to contiguous views and in some cases, all views. The evidence presented above strongly suggests that IT cortex is responsible for our ability to recognise objects under many different conditions, such as varying retinal position or viewing angle and also represent the retinal location or orientation of these objects.

1.2 Computational Approaches to Object Recognition

A number of different theoretical approaches have been suggested for how the brain may develop representations of individual objects. The biological plausibility of these approaches varies, as does the ecological validity of the training and/or the testing stimuli. This section presents a short summary of the benefits and limitations of a select set of different approaches. It is then argued, especially in relation to the neurophysiological evidence previously summarised, that a feature hierarchy approach appears to be closest to the biological processing within the primate ventral visual system.

It is possible to describe an object in a number of ways, for example in terms of its features: texture, area, colour, length and width. An object recognition system that relies on this type of object feature encoding has been proposed (Gibson, 1950; Mel, 1997). In such a system, the mere presence of a feature causes the associated feature detector to become active. For example, if a mouth, nose and two eyes are present in an image, a feature driven system would respond as if a face has been presented. However, the spatial arrangement of these features is not encoded in this system and as such, these features might be jumbled and fail to constitute a real face. Neurons in the IT cortex of monkeys respond less to faces where the features have been rearranged, revealing that they rely on the spatial arrangement of these features (Rolls et al., 1994).

There is some limited evidence that animals with less developed visual systems (compared to primates) may encode object representations similar to a feature space representation. For example, it was shown that pigeons could not distinguish intact cubes from distorted cubes when these different cubes comprised the same local features. It is hypothesised that the pigeons have learned to respond to the presence of these local features regardless of their spatial arrangement (Cerella, 1986). However, notwithstanding this limited evidence, a model of object recognition that completely relies on this

type of system is unlikely to exist in the primate visual system.

Another approach to object recognition suggested that a description of an object structure will not be affected by viewpoint if it is encoded within an object-centered frame of reference, as opposed to view-centered (Marr and Nishihara, 1978). The object must be broken down into its component parts (elements), and a structural relationship between the parts is then stored. Specifically, the authors believe that it may be easier to build invariant object representations once the individual parts have been identified. This type of encoding requires an internal 3D representation, or 3D structural description. In order to recognise an object, the scene must first be parsed and subsequently decomposed into objects. This process is sometimes referred to as recovering a description from the image. This description is then used within a look-up system. The look-up system must identify whether this recovered structural description matches that of any known stored object.

A similar approach, called the 'Recognition by Components Theory' (Biederman, 1987), receives revision by Hummel and Biederman (1992) in which a neural network model is presented. In the original theory, objects were divided into basic geometric components called 'geons'. Object recognition requires that representations be matched against those representations stored in memory which are most similar. The revision presents a neural network model that is successfully able to represent the structural description specifying the object's parts and the relationship between them.

A syntactic relationship between object elements is necessary, otherwise the approach of Marr and Nishihara (1978) and Biederman (1987) fails in a similar fashion to that outlined for 'feature spaces'; that is, the spatial arrangement of the object's elements would be lost. Furthermore, the relative size of the elements must also be present in the description. One notable drawback to this approach to vision is that the scenes must first be decomposed into objects, and objects must then

be decomposed into their elements. Natural scenes are likely to pose a particular problem to this requirement. For example, the structural description of a cat and a dog would be quite similar, and a very detailed structural description would have to be provided to the look-up system in order to tell the two apart. This might not be possible when the natural scene comprises partially occluded objects that also look very similar to one another. Aside from encoding a 3D syntactic description, a biological implementation of this system must be able to encode a dynamic real-time structural description. This may be achieved with temporal synchronicity of neuronal responses across a population, but such a solution would be particularly challenging (Rolls and Treves, 2011).

As previously stated, there are a vast number of back-projections within the ventral visual pathway. Some back-projections travel from 'adjacent' layers, for example, TEO to V4, while others may bypass entire regions of cortex, for example, TEO to V2. It has been suggested that computation within the visual system may be 'invertable', whereby neurons in V1 may be activated through feedback in exactly the same manner as when receiving feedforward input from the LGN (Hinton et al., 1995; Hinton and Ghahramani, 1997). The authors argued that backwards activation suffers no loss of information when reconstituting the activity of lower level cortical regions. More recently, Hinton (2010) suggested that actual output provided from the retina may be used simultaneously with feedback information to adjust synaptic weights during learning. However, despite the back-projections present in the ventral visual pathway, the biological mechanisms for such a learning regime are not apparent. Information loss occurs in the ventral visual pathway as a result of the complex reorganisation of retinal input throughout the ventral visual stream, and a precise reconstruction from top-down influence would therefore seem unlikely.

When considering computational models for processing visual data together with the neurophys-

iology data, a feature hierarchy approach appears to be the best fit. Specifically, the neurophysiology seems to suggest that despite very significant numbers of back-projections, lateral connections, and connectivity between the dorsal and ventral streams, the features that are represented along the ventral visual pathway increase in complexity. This approach must start with low-level representations of a scene, such as those found in V1: orientated bars and edges. A hierarchy is then built in subsequent layers based on what is represented in the preceding layer. For example, if the first layer responds to orientated bars and edges, a second layer may respond to combinations of the same features, and a third layer may respond to increasingly complex combinations and fail to respond to simpler ones. This is very similar to experimental data discussed previously, and as such, this doctoral thesis assumes a feature hierarchy approach to vision both in terms of how the primate visual system operates, and therefore in terms of the computational models used to study it.

A successful computational feature hierarchy approach to modelling vision must overcome a number of potential challenges. Firstly, the number of feature combination neurons must be bound within realistic parameters in order to avoid 'combinatorial explosion'. For example, if a separate neuron is required to encode every possible combination of features, the number of neurons necessarily increases exponentially as a function of the number of features to be encoded. Secondly, a feature hierarchy must be able to build invariant representations of individual objects, even in a cluttered scene. Thirdly, this invariance learning process and the recognition process must be very fast. Finally, the spatial information present within the stimuli must be maintained so that the representations of observed objects contain the encoded spatial configuration of their component parts.

Through the use of a number of simple competitive neural networks, a computational feature hierarchy approach to vision may overcome a number of these challenges by virtue of its inherent

design. For example, the potential for combinatorial explosion may be addressed if the neurons of a competitive network system respond to the statistically regular low-order feature combinations. Then, the total number of neurons required may be kept within a biologically realistic range. Even though this is still a significant number, it does correspond with the primate physiology, where roughly half of the total surface of the primate neocortex comprises visual processing areas (Sereno et al., 1995). Fast position invariance learning can occur as a natural property of a self-organising competitive network feature hierarchy model. If the network receives training with intermediate level feature combinations in all possible locations, then learning of a novel object can occur rapidly because it is simply a new combination of the medium order features.

Competitive networks can use unsupervised, local, biologically plausible learning rules to update the synaptic connections of neurons between layers. These synaptic weights are driven by the training stimuli that comprise the learning environment; under the correct conditions, competitive networks will learn to represent only the statistics present in their training stimuli including their spatial arrangement and not all possible combinations of all possible features. In this manner, they can take advantage of the statistical regularities that exist in the natural visual world. Also, competitive networks will make use of all available features in the stimuli to ensure that an output is always produced, even if the stimuli are partially occluded.

1.3 VisNet

VisNet is an artificial neural network model that self-organises during learning to provide a biological approximation of the ventral visual processing stream of the macaque temporal cortex. VisNet was originally developed by Wallis and Rolls (Wallis et al., 1993; Wallis and Rolls, 1997) and later revised

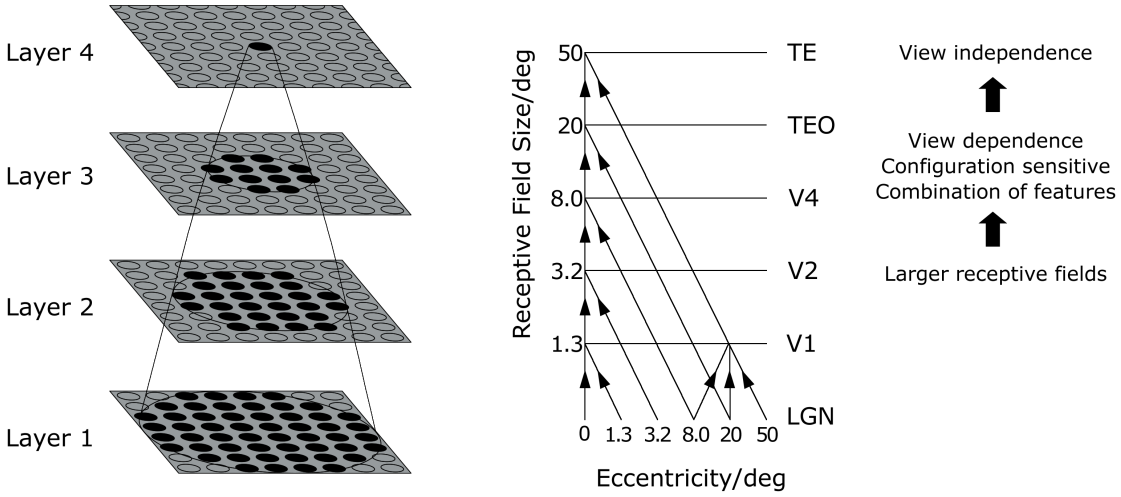

Figure 1.3: VisNet model overview. Left: Stylised image of the four layer network developed by Wallis et al. (1993) and Wallis and Rolls (1997). Convergence through the network is designed to provide fourth layer neurons with information from across the entire input retina. Right: Convergence in the visual system V1: visual cortex area V1; TEO: posterior inferior temporal cortex; TE: inferior temporal cortex (IT).

by Rolls and Milward (2000). As part of this doctoral thesis, VisNet has been substantially revised to incorporate a number of enhancements that allow for improved flexibility, improved computational speed and alternative image filtering routines. Many of the core and original VisNet principles are still instantiated in this new version and therefore, the generic term "VisNet" shall be used where the differences between the versions of the model are irrelevant. The design of the VisNet model was originally based on neurophysiological findings within the ventral visual stream (Rolls et al., 1992) and it has a number of important features.

Firstly, it comprises a hierarchy of four distinct processing layers that are intended to relate to V2 $\rightarrow$ V4 $\rightarrow$ posterior IT (TEO) $\rightarrow$ anterior IT (TE), respectively. Each layer of the VisNet model operates as a competitive network (Rumelhart and Zipser, 1985; Hertz et al., 1991) or a self-organising map (Kohonen, 1982). The different implementations of lateral connectivity within each layer are discussed in Section 1.3.2. The first layer of the model receives input from a separate pre-processing

stage which mimics V1. Past VisNet research uses a difference of two Gaussian functions to model V1 cell responses and provide a suitable fit to these responses (Hawken and Parker, 1987; Parker, 1996), and Chapter 6 uses both the difference of two Gaussian functions, as well as the performance of multiplying a sinusoidal function by a Gaussian function to create a Gabor filter that also models V1 cell responses (Jones and Palmer, 1987).

Secondly, neurons receive connections from a topologically corresponding region of the preceding layer. This leads to an increase in the receptive field size of neurons: the receptive fields of neurons in the first layer of the model are relatively small, while the receptive fields of neurons in the fourth (output) layer of the model are relatively larger. In theory, this level of convergence means that information from any part of the network's input layer has the potential to influence firing behaviour in any single neuron in the output layer of the model.

Thirdly, there is synaptic plasticity between neurons from different layers based on a Hebbian associative learning rule (Hebb, 1950). The earliest versions of VisNet focused on a modified version of this rule incorporating a temporal trace of the neurons' previous activity (Foldiak, 1991). This trace learning rule is based on the assumption that views of real world objects are not constant, but vary continuously over time. If either the object or the observer move with respect to one another, different views of the object will be seen. A central tenet of this theory is that views of a particular object will vary more rapidly than will the identity of the object being observed. The temporal proximity of views that belong to the same object underpin the invariant recognition of that object rather than binding together views from different objects. More recent VisNet research has focussed on Continuous Transformation learning (Stringer et al., 2006) as an alternative method for developing invariant object representations in feedforward competitive networks. The following section provides an overview of

the VisNet literature and a full model description is provided later in Section 1.3.3.

1.3.1 Literature Review

Hierarchical Model of the Ventral Visual Pathway

The very first publication describing experiments with VisNet was written by Wallis et al. (1993). In this pioneering research, the VisNet model demonstrated an ability to learn translation invariant properties to simple shapes such as 'T', 'L' and '+'. A more detailed study followed, which included a number of varied experiments providing a thorough account of VisNet's performance (Wallis and Rolls, 1997). The majority of this research focused on translation invariance, that is, the ability to recognise an object when presented at different retinal locations. Two sets of different training stimuli were used, each presented as part of a separate experiment. In the first experiment, the training set comprised the same simple shapes as in the first VisNet study by Wallis et al. (1993), while in the second experiment more complex face stimuli were used. A third and final study investigated rotational view invariance. These studies marked a milestone in computational feature hierarchy object recognition. They are discussed in further detail next.

The stimuli used in the first simple shapes study were selected based on the pre-existing neurophysiological evidence that has already been presented in this chapter: namely, that neurons in the ventral visual pathway are responsive to features, such as a horizontal/vertical bar conjunction. The network was trained over 800 epochs, where one epoch comprised a single presentation of each object in nine possible locations. Together, the nine different locations formed an equally spaced 3x3 grid over the retinal canvas, which was 128x128 pixels. All four layers of the VisNet model were trained simultaneously and stimuli were selected randomly, but once selected they were presented across the

nine locations in a predefined order. For neurons in the output layer to learn to respond to a single stimulus at all locations, the trace rule was required to operate over a suitably long interval; it spanned a period during which the features seen by a particular neuron in the hierarchy were likely to come from the same object. Conversely, the trace rule was required to operate over a short enough period so that it did not learn to produce associations between features from different objects.

The network testing followed exactly the same regime, but for only one epoch and without learning. Neurons in the output layer of the VisNet model had learned to respond to one of the three objects across all nine locations, and all three objects were represented in this way. Neurons in the earlier layers had learned to respond to lower-order feature combinations, and had intermediate levels of translation invariance. This was largely due to the receptive field size of neurons in the intermediate layers; that is, they do not cover the whole retina and will therefore never be able to respond to the stimuli in all locations.

The second translation invariance study investigated whether VisNet was able to learn invariant representations to real human faces, in contrast to the simple shapes outlined above. Instead of just the three objects, seven faces were used across all nine locations. The authors made performance comparisons between a trace learning paradigm, a Hebbian learning paradigm with no temporal trace, and an untrained case where no learning occurs. They found that the trace rule was required in order for VisNet to develop translation invariant representations similar to those in the first experiment.

Wallis and Rolls (1997) claimed that this result highlighted the relative power of the trace learning rule for building translation invariant representations in comparison to the Hebbian learning rule. As will be made clear in the following chapters detailing the original research contributions of this doctoral thesis, this claim was possibly over emphasised. The poor performance of the Hebbian

learning rule was actually a direct consequence of the way in which the stimuli were presented to the network during training and testing.

The third and final experiment by Wallis and Rolls (1997), presented as part of this initial study, investigated VisNet's ability to learn view invariant responses to three different faces 'rotating' in depth over seven views. During training, each face was presented in isolation to the network over the seven consecutive views, centrally on the retina. This training paradigm meant that, for the first time, all the views of each stimuli used during training occurred in the same location on the retina. The results obtained demonstrated that VisNet was able to successfully address the challenge of building view invariant representations when trained with the trace rule, but failed to do so when trained with a simple Hebbian learning rule, or was untrained altogether. However, there are a number of caveats to this interpretation.

Firstly, due to the similarity between the features across different face views, some neurons (even in the first layer of VisNet) showed invariance to changes in the viewing angle of the face. The authors originally argued that this was to be expected, as slightly rotated views of the same face will share many of the same basic features in the same locations as in other views. A simple Hebbian learning effect will cause the same features to become represented by the same output neurons during training. However, consider the following. Although the degree of rotation was not published by the original authors, careful inspection reveals that each face varied over roughly 126°, in seven 18° increments. In reality, the facial features (e.g. eyes, nose, mouth) varied quite considerably, and future research showed that steps of 18° may be too large for the original authors' claims to be correct (Stringer et al., 2006). Therefore, it is more likely that the early layers of VisNet did not learn to respond to the facial features, and instead learned to respond to the contour boundaries at the edges

of the face stimuli. These remained constant, and many contour boundaries appeared in the same location across different views of the same face. Therefore, this would imply that VisNet failed to respond to the important components of a face, such as the eyes and the mouth, but instead merely learned to respond to the most salient contour boundaries, which in this case comprise the participant's hair. Nonetheless, despite these considerations, the network did successfully classify the faces, where neurons in the output layer of the model responded exclusively to their preferred face across all seven rotational views.

As part of the translation (location) invariance simulations, Wallis and Rolls (1997) also investigated a key parameter of the trace learning rule, η. They concluded that the optimal length of this trace term was dependent upon the number of stimulus views, or rather, the number of views that should be associated together in order to form an invariant representation with respect to location. It was also concluded that η should vary with respect to the receptive field size of neurons within a layer. Specifically, if a neuron receives input from a small part of the retina (small receptive field), it cannot respond to large shifts in the input stimuli and as such, the number of transforms to which it responds will be relatively small. Accordingly, η may also be smaller. Conversely, if a neuron receives input from the whole of the retina, as in the latter layers of the VisNet model, the neuron is capable of responding to much larger shifts in the input stimuli, and therefore more views. The η value should then be larger because the neurons can respond to all possible views of a stimulus.

As part of their conclusions, the authors declared that VisNet was able to maintain the spatial information present within the object, that is the relative spatial relationships between the parts of an object. Based on the evidence provided, this claim should be investigated further. It must be shown that VisNet neurons (perhaps in the intermediate layers) respond exclusively and preferentially to a

specific combination of object features when they are presented in a particular spatial arrangement. The same neurons should not respond maximally when only a single feature is present. If they do, these neurons merely signal the presence of an object feature, and do not encode spatial information. The authors do not provide enough supporting evidence to justify the claim and it is not until later that the more appropriate experiments were performed (Elliffe et al., 2002).

In summary, this initial publication by Wallis and Rolls (1997) reveals that a computational feature hierarchy model of object recognition, VisNet, performs well when learning and recognising invariant representations. Although the mapping between the four layers of VisNet and the ventral visual pathway is loose, it may serve as an approximation. Over successive stages, from layer 1 to layer 4, VisNet learned translation and viewpoint invariant representations using a biologically plausible local learning rule and self-organising competitive network architecture.

Model Revision and Performance Limits

In a detailed follow up study (Rolls and Milward, 2000), an updated translation paradigm was implemented. This time there were seven different faces presented over 25 equidistant locations arranged in a 5x5 grid formation over the input retina. Presentation across location was random, such that faces were not presented in a predefined order and the order varied from epoch to epoch. In addition, many different aspects of VisNet were investigated and revised, including the introduction of new network performance measures. The main contribution of this research was to finalise a number of aspects of the VisNet model and to improve model performance and biological plausibility.

A number of different trace learning rules were investigated. Firstly, the trace formula was modified to incorporate a temporal trace of neuronal activity from the preceding trial, as opposed to the current trial (n.b. a trial comprises a stimulus view presentation). This alteration was investigated for

the post-synaptic cell, as well as the pre-synaptic cell, and for both simultaneously. Secondly, new information analysis measures were created to investigate how much individual cells had learned about the stimuli to which they were exposed during training. Furthermore, multiple cell information analysis measures were developed to quantify how much information could be encoded by a population of cells. Thirdly, a revised activation function was introduced along with modified lateral inhibition within a layer. The relevant details of these revisions are discussed next.

The modification of the trace learning rule to incorporate a temporal trace of neuronal activity during the preceding trial produced better results than the original trace rule adopted by Wallis and Rolls (1997). Improved performance was sustained over all variations of the new trace rule, including the pre- and post-synaptic variations, as well as both together. Importantly, this modified trace rule did not incorporate any contribution of the firing produced during the current trial.

The information theory analysis measures introduced later in Equation 1.13 and Equation 1.15 were developed as part of the study by Rolls and Milward (2000). The equations are based on the same measures applied as part of experimental single cell recording analysis (Rolls and Treves, 1998). They provide a principled and quantitative measure of network 'invariance' performance, that is, how well the network learns to represent individual stimuli invariantly. These measures contribute to the standard VisNet results section in numerous subsequent publications. By design, these measures are particularly well suited for investigating invariant object recognition.

A revised activation function was implemented, along with updated lateral inhibition. The original lateral inhibition procedure implemented by Wallis and Rolls (1997) only operated over a few pixels. In this updated study, lateral inhibition operated as described later in Equation 1.3 and over a larger region based on approximately half of the radius of convergence from the preceding layer. Next,

contrast enhancement was adjusted by means of a new activation function; instead of raising the activation to a fixed power and normalising the resulting firing within a layer to have an average firing rate equal to 1.0, a more biologically plausible form of the activation function is implemented in the form of a sigmoid function (Equation 1.5). One particular advantage of this sigmoid activation function is that it allows the sparseness level to be accurately controlled and captures the threshold non-linearity of neurons.

In summary, Rolls and Milward (2000) provided useful evidence that the VisNet architecture captures some aspects of invariant visual object recognition as performed by the primate visual system. This study implemented an updated trace learning rule, new information analysis measures, an improved biologically plausible activation function and showed that a more appropriate range of lateral inhibition can improve network performance. Unless otherwise stated, these adjustments form the basis for subsequent VisNet research.

Feature Combinations

As noted above, Elliffe et al. (2002) extended the original VisNet research (Wallis and Rolls, 1997) to further investigate whether the model was able to provide a solution to the 'spatial binding problem'. Specifically, the updated version of the model (Rolls and Milward, 2000) was used to investigate whether VisNet can encode the local spatial configuration information between object features such that output neurons are able to respond to stimuli even in novel locations. Although the original authors claimed to have addressed this challenge in their study (Wallis and Rolls, 1997), necessary additional experiments were performed by Elliffe et al. (2002) and are discussed next.

VisNet was trained using a number of highly controlled stimuli comprising up to four orientated bars, positioned either vertically or horizontally: top horizontal bar (T); left vertical bar (L); bottom

horizontal bar (B); and right vertical bar (R). These component features were independent of one another and were combined in different ways to create a total of thirteen stimuli. The stimuli consisted of individual features, feature pairs, feature triples and all four features. The network was trained using a trace learning rule and in a translation invariant paradigm identical to the first VisNet study (Wallis and Rolls, 1997). During training, each one of the thirteen stimuli was presented individually to the network, sequentially across all nine possible locations. The results of this study show that VisNet was able to develop separate, translation invariant, representations for some of the individual stimuli presented during training. The stimuli comprising object triple and quadruple features were not represented invariantly in the output layer of the VisNet model. The authors argued that this study demonstrates that VisNet is able to develop translation invariant discrimination of objects constructed from a common pool of features, which have a subset-superset relationship. Specifically, they note that a VisNet output cell responding invariantly to feature combination TL may genuinely signal the presence of exactly that combination, and will not necessarily be activated by T or L alone, or by TLB.

However, explicit testing of this claim was not performed by Elliffe et al.. To successfully make this claim, at least one cell that responds invariantly to TL must not respond to T, L, or TLB in the same manner. Cell response plots are not provided and there is no evidence to suggest that these tests were performed as part of this study. As such, it is not yet clear that VisNet is able to maintain the spatial information present within the stimulus in order to develop specific feature combination cells that respond exclusively to their preferred combination. Furthermore, it was not demonstrated that VisNet is able to generalise to novel locations, that is, respond invariantly to the same stimulus in a location at which it was not presented during training.

Generalisation to Novel Views

In 2002, Stringer and Rolls conducted a different study to investigate whether VisNet could learn to correctly recognise novel views of 3D objects, where the novel views comprised a combination of previously learned features (Stringer and Rolls, 2002). The 3D objects were rotating in depth, and as such this study was a view-invariant paradigm, as opposed to a translation invariant paradigm. To perform correctly, VisNet needed to generalise to transforms of objects on which it had not been trained. Specifically, it needed to develop representations useful for the recognition of 3D objects by forming view-invariant representations to allow generalisation within a generic view. It was hypothesised that VisNet could achieve this when provided with invariant training on the low-level features from which the new object view would be composed.

In this experiment Stringer and Rolls (2002) adopted two training stages. Initially, only the first two layers of VisNet were trained, using feature-pairs. There were three different individual features: a horizontal; a diagonal; and a vertical line. These three individual features could occur in three possible locations, and overall this meant that when combined, there were 18 possible feature pairs presented during training. During the first stage of training, each pairing was presented across five different rotations that varied by 30° steps over a total of 120°. During the final stage of training, a total of six possible feature-triples were presented and only the last two layers of the VisNet model were trained. Importantly, the feature-triples were built using the original three simple features outlined above. Furthermore, these feature-triples were only presented across four orientations, instead of all five. This meant that VisNet could be tested using each feature-triple at a novel rotation, unseen during the training phase.

The results of this experiment were presented as cell response plots as well as information analysis

measures. Together they showed that VisNet learned to respond exclusively to each feature-triple, across all five possible viewpoints, despite the fact that each feature-triple was only presented during training at four viewpoints. This showed that VisNet, a feature hierarchy model of visual processing, was able to successfully generalise to novel views of an object and maintain the spatial relationship presented between features. It was shown that, in order to recognise feature-triples presented at novel views (transforms) during testing, it was not necessary to provide training with every possible feature-triple transform. Importantly, it should be noted that the untraind transform was only novel in so much that the complete feature-triple had not been presented there before. Learning had already occurred in the first two layers of the VisNet model to all possible feature-pairs in the novel location. This suggests that learning of a feature 'alphabet' can support invariant recognition of more complex stimuli in novel transforms.

Position Invariant Recognition with Cluttered Backgrounds

A second strand of research began to study model performance in more interesting training environments that were designed to capture some of the complexities of the real world.

Stringer and Rolls (2000) investigated VisNet's multilayer hierarchical approach to invariant object recognition when learning about stimuli in 'cluttered' environments and when they were partially occluded. Both experiments used a trace learning rule during training to update the synaptic weights between neurons. These different experiments form the basis and inspiration for a number of original research experiments presented as part of this doctoral thesis and as such, they are discussed in detail next.

A developing visual system learns to see by viewing complex natural scene made up of many objects. The underlying system experiences little or no outside feedback and is completely self-

organising. A model of the ventral visual pathway must be able to capture how we learn and build independent and invariant representations of individual objects despite the fact that these objects are never viewed in isolation. In the first study, Stringer and Rolls (2000) investigated whether VisNet was able to build invariant representations to seven different translating face stimuli, presented across 9 locations, and across three different backgrounds.

This translation study was very similar to the original VisNet research (Wallis and Rolls, 1997). Instead of training all four network layers simultaneously for 800 epochs, the model was now trained layer-by-layer for fewer epochs: 50 epochs for the first layer; 100 epochs for the second layer; 100 epochs for the third layer; and 75 epochs for the fourth layer. This alteration provided qualitatively identical results to the original training procedure, but had the advantage of decreasing the processing time of the overall simulation. As such, unless otherwise stated, this training procedure was adopted for all subsequent VisNet research, including the work presented as part of this doctoral thesis.

The inclusion of different backgrounds during training and testing was one of the major novel contributions of this study. Specifically, there were three possible backgrounds upon which the face stimuli could be presented: blank, and two 'cluttered'. The cluttered backgrounds were meant to be representative of a real world environment where one background was of a tree and the other was of plant foliage. During training, faces were superimposed over one of the three backgrounds and the translation paradigm was then completed as normal. This was repeated for each background in turn, and the introduction of a different background marked a different experiment in the overall process.

After training and testing VisNet, network performance was measured in one of two ways: testing on a blank background or testing on the two cluttered backgrounds. In both methods, the response profiles of neurons in the output layer of the model were used as an indicator of model performance.

Specifically, neurons should respond exclusively to a preferred face across all possible nine locations. VisNet was able to successfully build independent translation invariant representations of each of the seven faces when presented during training and testing on the blank background. This result confirms the original VisNet research (Wallis and Rolls, 1997). VisNet was also able to successfully recognise faces presented on the cluttered backgrounds during testing on condition that they were trained first on the blank background. However, VisNet was unable to learn about the faces when they were presented during training on a cluttered background and then tested on the blank background, and furthermore, VisNet failed to learn about individual faces when they were presented during training and testing on the same cluttered background. VisNet had in fact learned about the background, and not about the faces.

These results suggest that it is challenging for VisNet to build position invariant representations when stimuli are presented in cluttered scenes. The authors argue that these difficulties may be ameliorated if the appropriate feature-detecting neurons are set-up through previous exposure to the relevant class of stimuli. However, the manner in which VisNet is trained and the concept behind the 'cluttered' environment could both serve to complicate the challenge that VisNet has to overcome if it is to successfully build invariant object representations.

In the second study, Stringer and Rolls (2000) investigated how VisNet may learn to recognise objects that are partially occluded. In natural visual environments, objects are often partially occluded by other objects and yet the primate visual system is still able to overcome this challenge and correctly recognise different objects. To investigate whether VisNet can cope with such a situation, a new translation paradigm was conducted. The model was trained on seven different complete faces presented across all possible nine locations on a blank background. Testing was performed using partial views

of these faces: the top half of the face with the bottom half occluded; or the bottom half of the face with the top half occluded. The occluded part of the face was completely removed and replaced with the solid background shade.

VisNet successfully recognised faces during testing when presented with either the top half, or the bottom half of the face. Recognition was slightly improved in the top-half scenario (where the bottom half of the face was occluded). A possible explanation for this result is that there is more visual information available in the top half of the face images, including significant changes in contrast around the hair. This successful performance suggests that VisNet does not need to rely on key markers embedded within the stimulus in order to complete recognition tasks. Relying on key markers can become fragile when these markers are removed or occluded. As this experiment demonstrates, VisNet can instead generalise and complete the object representations from partial views as a result of the distributed representations stored within the competitive neural network architecture.

The ecological validity of the partially occluded object training paradigm can be improved. This would provide a more realistic challenge for the VisNet model to overcome; a challenge with which the primate visual system can clearly cope. During learning, objects are not presented in isolation, nor are they presented with the type of occlusion demonstrated in the previous study. Although careful consideration was given to the training stimuli in order to avoid confounding occlusion with cluttered backgrounds, this type of control also led to a somewhat impoverished training environment. Objects are usually occluded by other objects that together make up a scene. The same may be said for the experiments within the first study by Stringer and Rolls (2000); that is, backgrounds do not simply comprise of 'noise' that can be represented by nondescript plant foliage or trees. Instead, backgrounds comprise many other objects that together make up a scene. This approach of using increasingly

realistic scenes is very different to that of Stringer and Rolls (2000), and serves as a foundation for the occluded object experiments in Chapter 3 and object separation by rotation in Chapter 4.

Continuous Transformation Learning: View Invariance

The research discussed so far has investigated the performance of VisNet while using a trace learning rule, introduced later in Equation 1.8 and Equation 1.9. As previously stated, the preferred version of the trace learning rule incorporates a temporal trace of neuronal activity from the preceding stimulus presentation trial. Stringer et al. (2006) present an alternative learning mechanism that relies purely on a Hebbian learning rule without the need for a temporal trace. This alternative learning mechanism is called Continuous Transformation (CT) learning and forms a foundation for a number of studies performed as part of the original research in this doctoral thesis. In brief, CT learning relies on spatial continuity between transforms of the same object in the visual environment. This learning mechanism is described in detail later in this chapter, and a summary of the original research is presented next.

As part of their publication, Stringer et al. (2006) described four separate experiments investigating the performance of CT learning with a Hebbian learning rule when building invariant representations of two different objects rotating in depth. In all the experiments, the number of training epochs was 50, 100, 100, and 75 epochs for layers 1 - 4 respectively. In the initial experiment, VisNet was trained using a cube and a tetrahedron, each rotating through 180 ° in 1 ° steps. All 180 views were presented consecutively for each object. In the second experiment, the performance of CT learning with a Hebbian learning rule was compared to trace learning; a total of four different step-sizes (1 °, 2 °, 9 °, and 36 °) were investigated in order to assess relative performance between the two learning algorithms. In the third experiment, rather than presenting each object in turn, where all views of the selected object are presented consecutively, the views of each object were interleaved. That is, train-

ing occurred with temporally alternating transforms from each object. Specifically, the cube and the tetrahedron were again rotated over 180 ° with 1 ° rotations between individual views, but in contrast to the previous experiments, the different transforms of the two objects were not shown successively for one object and then the other, but were interleaved. In the fourth and final experiment, the views of each object were split up into six blocks. Within each block, the rotation was shown continuously in thirty 1 ° steps. However, the presentation order of the blocks within and between objects was randomised.

The results of the first experiment show that VisNet can use CT learning to build separate representations of each object, where neurons in the output layer of the model learn to respond exclusively to either the cube, or the tetrahedron. In the case of the cube, many neurons learned to respond to all views and were therefore regarded as fully invariant. In the case of the tetrahedron, neurons learned to respond to large contiguous portions of the view-space, and the entire view-space could be represented by two output neurons. It should be noted that the cube was presented face-on to the retina, such that after the first initial 90 ° rotation, the following 90 ° looked identical to the first. This makes the challenge of developing a completely view invariant representation easier. This type of rotational symmetry was not present in the tetrahedron, and so the network had to bind together 180 genuine different views of the same object, which was made particularly challenging by the 'catastrophic' change in object appearance when a new face of the tetrahedron came into view.

Results from the second experiment reveal that CT learning can become fragile if the change in viewing angle from trial to trial is greater than 1 °. Specifically, when the viewing angle step-size was 1 °, invariant representations formed, but when the step-size was 2 ° or more, CT learning began to break down. Incidentally, because the trace learning rule has a basic Hebbian component, it also

performed very well at a 1 ° step-size. In this situation, the Hebbian component to the trace learning rule enabled CT learning to take place. If there is significant overlap between stimulus presentations, then the Hebbian component of the trace learning rule will allow CT learning to take place even if the temporal trace itself contributes very little to learning. At 2 ° step-sizes, both trace and CT learning broke down, while at 9 ° and 36 °, trace learning began to increasingly outperform CT learning. The reason for the superior trace performance over the larger step-sizes is because of two factors. Firstly, the similarity between consecutive views becomes smaller as the step-size increases and this degrades CT performance directly. Secondly, because the total viewing range was always 180 °, as the step-sizes increased, the total number of views decreased. For example, where the step-size was 1 °, there were 180 views. However, when the step-size was 36 °, there were only 5 views. Trace learning is most robust when there are relatively fewer transforms per object. The results reveal that trace learning was able to produce invariant representations of each object individually when the step-size was 36 °.

Results from the third experiment reveal that CT learning can still operate usefully even when the views from different objects are interleaved during training. Neurons in the output layer of VisNet learned to respond exclusively to either the cube or the tetrahedron. Furthermore, the results presented suggest that the output neurons had also learned to respond invariantly to their preferred object: cube or tetrahedron. This is in contrast to the first experiment, where the output neurons learned to respond to large contiguous portions of the 180 tetrahedron views, but not all of them. Although it is implied, the results presented by Stringer et al. (2006) do not explicitly demonstrate if truly invariant tetrahedron representations formed in the output layer of the model. Given the results of the first experiment, it is unlikely. Nonetheless, the network learned to respond exclusively to each object over large portions (if not all) of the 180 views presented during training, despite the fact that views of each object

were interleaved with one another.

CT learning is a useful alternative to the trace learning rule because the spatially closest views of an object need not be presented close together in time. For example, if there are frequent saccades between different objects, temporal continuity and thus trace learning might be compromised. However, the results of this study show that CT learning can operate successfully in this type of environment.

Results from the fourth and final experiment demonstrate that VisNet is able to develop exclusive, view invariant, representation of each object even when trained on random blocks comprising 30 contiguous views of each object. Neurons in the output layer of the model learned to respond exclusively to each object, and also learned to bind together the different training blocks into a completely invariant representation. This is an important result because, in the real world, it is very likely that all the views of an object will not be presented together consecutively. It is more likely that there are discontinuous segments of transforms in which the different segments are seen in a random order.

In summary, Stringer et al. (2006) described VisNet experiments demonstrating that CT learning, an alternative learning mechanism to trace learning, can develop rotational view invariant representations using a simple Hebbian learning rule. In the experiments discussed above, view invariance was investigated, but CT learning may also build invariant representations when objects translate across the retina.

Continuous Transformation Learning: Translation Invariance

In a recent publication, Perry et al. (2010) describe experiments with VisNet and CT learning during a translation training paradigm. The performance of CT learning was investigated with respect to the development of location invariant responses. CT learning theory suggests that translation invariance is not only possible, but arguably an optimal training paradigm for building location invariant stimulus

representations. This study contained four separate experiments.

The first experiment explored the effect of stimulus spacing, where the distance between the different consecutive stimuli presentations was increased from one to five pixels. The second experiment explored how CT learning operates as the number of training locations, over which translation invariant representations of images must be formed, increases. The third experiment explored whether randomising the order in which the different views of each stimulus are presented has a qualitative effect on learning. The fourth and final experiment explored how many different stimuli neurons in the output layer of VisNet are capable of discriminating with translation invariance.

The results from these experiments show that VisNet is able to build translation invariant responses to individual stimuli using only a simple Hebbian learning rule. Results from the first experiment confirmed that, as the spacing between consecutive transforms of the same object increased from one to five pixels, the CT learning effect became progressively weaker. Although the spacing between stimulus views will be heavily dependent on the relative size of the stimulus, it still serves to highlight the underlying principle of CT learning and corroborates previous research (Stringer et al., 2006). More specifically, if the pattern of activation produced after a stimulus presentation within the first VisNet layer does not significantly match that of a comparable (but different) stimulus presentation, then the different stimulus presentations will fail to become mapped to the same neuron in the next layer of the VisNet hierarchy.

Results from the second experiment reveal that, in the case of this particular training paradigm, CT learning could support invariance learning of up to 225 different transforms of the same object. However, there are a number of limitations to this result that must be considered in context. As in the first experiment, this result is specific to the training paradigm, that is, the particular retinal size, and

the size of the stimuli used during training.

The third experiment confirms that the presentation order of stimulus views is not a crucial factor for developing translation invariance. A comparable number of neurons in the VisNet output layer learned to respond to the training stimulus in both a traditional consecutive training paradigm and in a 'permuted' (randomised) stimulus view presentation paradigm. In the fourth experiment, the authors investigated how many different translation invariant stimulus representations VisNet is capable of discriminating between. Similar to the other studies, the results are specific to the particular training paradigm used. In this case, VisNet could discriminate perfectly up to 3 faces when translation occurred over 25 locations.

In summary, Perry et al. (2010) demonstrated that VisNet is able to build translation invariant representations with a CT learning mechanism using only a basic Hebbian learning rule. Although many of the experiments performed as part of this study are specific to particular training paradigms and model parameters, an important general result emerged: CT learning can successfully operate when the different translating views of a stimulus are presented randomly during training. This finding supports the previous research outlined above, demonstrating that the randomisation of stimulus views during presentation in a rotational view invariance training paradigm did not prevent the development of invariant representations (Stringer et al., 2006).

Past research has also shown that VisNet is able to develop lighting invariant responses (Rolls and Stringer, 2006). In this study, the lighting source was provided from the left during training, but during testing it was presented from both the right and then the left, creating two testing phases. VisNet performed equally well to both lighting sources during these recognition phases demonstrating that it had developed lighting invariant responses.

The VisNet literature reviewed so far has established that a feedforward feature hierarchy model of the ventral visual pathway can build invariant representations to different stimuli (objects and faces). As outlined above, these stimuli may vary in terms of retinal location, size, and rotational view. In order to achieve these invariances, the model can use temporal continuity in an associative synaptic learning rule with a short term memory trace, and/or it can use spatial continuity in Continuous Transformation learning.

Up to this point, the general approach to VisNet research had been to investigate model development when learning with one object presented at a time. However, in the real world objects are not seen in isolation during learning and backgrounds comprise many other objects that together make up a scene. A new approach when investigating object recognition is to train VisNet with multiple objects presented simultaneously during training. Although VisNet must now learn to build separate object representations despite the fact that objects are never seen in isolation during training, it also serves to increase the ecological validity of the training environment and also provides VisNet with more useful visual information.

Multiple Objects: Trace Learning

Stringer et al. (2007) published the first results after successfully training VisNet with multiple objects presented simultaneously. The publication describes two separate studies, the first of which comprises two experiments which demonstrate that a simple one-layer competitive network is able to build independent representations of individual objects despite the fact that three objects were always present during training. The second study demonstrates that VisNet can develop separate translation invariant representations for different objects, despite the fact that they are always presented in pairs during training.

In the second study, a trace learning rule is implemented so VisNet can associate together different translating views of the same object despite the fact that objects are always seen as part of a pair, and never in isolation. The VisNet retina was divided into a 2x2 grid thus creating four separate non-overlapping locations. Ten different objects (e.g. pyramid, cube, cylinder) were used, which together formed a total of ${}_{10}C_2$ (45) different pairings. During training, each pairing was presented to the network shifting clockwise around the four locations in the grid. Each object occupied its own part of the 2x2 grid such that during any given stimuli presentation, two of the grid locations were kept empty. As is usual in VisNet experiments, the individual layers were trained for 50, 100, 100, and 75 epochs from layers 1 - 4, respectively. The presentation of all the object pairs in all four possible locations constitutes one epoch of training.

The results demonstrated that VisNet was able to develop neurons that responded exclusively to their preferred object across all four possible transforms. All objects were represented in this way. Together, the two studies provided evidence supporting an important observation; the most important factor influencing the development of individual representations for each object appeared to be the number of stimuli in the total training set. For example, in the case of the second study with VisNet, this was set to 10 objects. As previously stated, with 10 objects in the training pool, there were a total of 45 pairs. Over the course of one epoch of training, each object was presented once with all 9 remaining objects. An object comprises features, such as bars and edges, and those features that constitute a given object are therefore highly correlated with one another because each time an object is presented, so are all of its features. This is in contrast to the features of different objects. Features of different objects are presented together less frequently. As a result of the object pairings, the features of a given object were seen together nine times more often than they were seen with the features from

any other object. Thus, representations of individual objects are learned. In contrast, results from the first study with a one-layer competitive network revealed that, as the number of objects in the parent training pool was decreased, the network failed to learn separate representations of individual objects and learned about object combinations instead. This observation forms the basis for Chapter 2.

In summary, these experiments mark the first time that VisNet was trained with multiple objects presented simultaneously during learning with a trace rule. Furthermore, it is the first time that VisNet was presented with multiple objects at the same time during a translation invariance training paradigm. VisNet successfully developed invariant representations of individual objects by utilising the statistics of the training environment.

Multiple Objects: Continuous Transformation Learning

In a similar study, Stringer and Rolls (2008) investigated the performance of VisNet when learning to develop rotational view invariant representations, with two objects always presented during training. This study investigated CT learning, as opposed to the trace learning rule used in the experiments of Stringer et al. (2007).

The VisNet retina was divided into a 3x3 grid providing nine non-overlapping locations. The stimuli used to train the network comprised nine objects, each of which was assigned to one of the nine locations. During training, VisNet was presented with object-pairs and never with objects in isolation. In total, there were ${}_9C_2$ (36) possible stimuli pairings on which to train the network. For each pairing presentation, the two stimuli were simultaneously rotated about their vertical axis over a range of 90 °, with a step-size of 1 ° between successive views. As in other VisNet studies, training comprised 50, 100, 100, and 75 epochs for layers 1 - 4, respectively. Importantly, a simple Hebbian learning rule was used to isolate and explore the effect of CT learning.

The results of this study showed that VisNet successfully learned to develop completely invariant representations of individual objects. Specifically, neurons responded exclusively to their preferred object over all 90 views. This demonstrated that CT learning can operate usefully even when multiple objects are always presented during training. This problem is non-trivial since VisNet must learn to separate objects from one another and build invariant representations simultaneously. The results corroborate those of the previous trace learning experiments (Stringer et al., 2007).

Summary

The experiments presented as part of this section have incorporated the majority of VisNet research published to date. It has been shown that this feedforward hierarchy model of the ventral visual pathway is able to learn invariant representations with respect to view (rotating in depth), translation (retinal position), and lighting. Depending on the training paradigm, these invariant representations may be formed using two different learning mechanisms: trace learning and Continuous Transformation (CT) learning. Furthermore, invariant representations of individual objects can be formed even when multiple objects are always presented together during training. VisNet has also demonstrated an ability to successfully generalise to novel views of an object, providing it has received the appropriate training using an 'alphabet' of lower-order feature combinations.

A general approach to VisNet research had been to investigate model development during training with simple training environments, predominantly comprising only one object at a time. It is necessary to provide highly controlled training environments with which to train VisNet so that network performance is systematically investigated and understood. However, it is also important to investigate increasingly complex training environments that begin to approach something comparable to a complex scene.

1.3.2 Competition and Self Organisation

This section provides an overview of the different learning mechanisms within the VisNet model. Competitive learning is implemented in VisNet through competitive layers of neurons. Competitive learning has the general function of classifying different input stimuli into different output categories that are represented by neurons in the output layer of the model. Continuous Transformation (CT) learning plays a central role in the majority of experiments presented as part of this doctoral thesis and as such, it receives a detailed explanation in the following section. The final part of this section provides a description of a self-organising map (SOM) architecture. Originally proposed by von der Malsburg (1973), the SOM model relies on Hebbian learning in forward connections, short-range lateral excitation, and long-range lateral inhibition.

Competitive Learning

A competitive network (Rumelhart and Zipser, 1985; Hertz et al., 1991) is a self-organising neural network that learns to classify different input patterns according to their similarity, causing one or more neurons in the output layer to become active. Competition can be graded such that a subset of output neurons respond to input stimuli or 'winner-takes-all', whereby only one output neuron is able to become active after presentation of a particular stimulus, due to the increased levels of competition. Competitive learning is implemented in the VisNet model using a spatial filter discussed in Section 1.3.3, and may be described as follows.

Consider two layers of neurons within a VisNet-style model, connected by one layer of feedforward synapses. These synapses may be updated according to a simple Hebbian learning rule and are initially set to random values to prevent biases in learning. When input neurons fire, they will cause a

subset of neurons in the output layer to receive varying levels of activation. Mutual lateral inhibition operates within this output layer of neurons such that every cell may inhibit every other cell within the same layer, and vice versa. The level of inhibition that a cell provides to its neighbours is proportional to the amount of activation it initially receives from cells in the preceding layer. Also, in the case of VisNet, neurons will inhibit cells that are spatially proximal rather more strongly than those that are relatively far away. The resulting effect causes cells that receive the most input activation from the preceding layer to inhibit other cells close by in the same layer more strongly. As such, only a small number of output neurons will be able to fire and consequently have their synaptic connections with the preceding layer strengthened. This is in accord with a standard Hebbian learning rule, whereby co-firing cells between layers will have their synapses strengthened. When the same input pattern is presented to the network again, it is more likely to cause the same output neuron(s) to become active once more due to the previously strengthened connections.

Competitive networks have a number of properties that are well understood and documented, such as redundancy removal and sparsification (Rolls and Treves, 1998). In particular, orthogonalisation and categorisation are very useful properties that cause new input patterns that are similar to previously learned input patterns to activate the same neurons in the output layer of the model. This is because the novel, similar, input patterns activate some of the same input neurons to previously learned input patterns, increasing the likelihood that the same output neurons will fire and subsequently have their connections to the newly recruited input neurons strengthened. This forms the basis for categorisation. Furthermore, dissimilar input patterns are orthogonalised and will be represented by the firing of different output neurons.

More specifically, after learning, an individual output neuron (or subset of output neurons) will

represent a cluster of similar input patterns (categorisation). The input patterns within this cluster will be more similar to one another than they are to the cluster of input patterns represented by any other output neurons (or subset of output neurons). That is, the within-cluster similarity of the input patterns in any given cluster is always larger than the between-cluster similarity between all of the clusters. Hence, the competitive network performs orthogonalisation.

The combined process is illustrated in Figure 1.4. The left-hand plot shows the state of the competitive network prior to learning while the right-hand plot shows the state of the network after learning. In each plot, the grey circles represent input vectors while the black crosses represent the synaptic weight vectors of three output neurons. After training, the only difference between the two network states is the position of the three weight vectors; they have moved to the center of the three clusters of inputs. After enough training, the weight vector is updated to become the prototypical weight vector for the category of inputs to which it has learned to best respond. Notably, this prototypical weight vector may never have actually been presented as an input during training. The level of mutual inhibition within a layer will control the level of competition, and help define the size of the clusters of input patterns to which neurons in the output layer learn to best respond, as well as the number of output cells that are active.

Continuous Transformation (CT) learning

A computational theory of how the ventral visual pathway in the brain may develop neurons that respond to objects with transform (e.g. view or location) invariance is Continuous Transformation (CT) learning. CT learning can cause output neurons to respond invariantly to a range of input patterns. It uses a biologically plausible associative (Hebbian) synaptic modification rule that can exploit image similarity across successive transforms (e.g. views) of a continuously transforming object in order

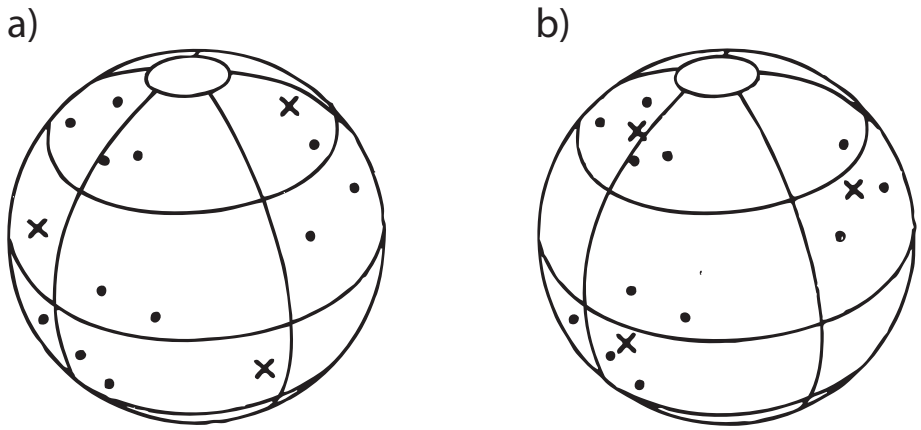

Figure 1.4: Competitive Learning. The figure shows input vectors, represented by the dots, and weight vectors for each of the three neurons, each represented by an 'x'. The figure (a) before learning is presented on the left. The figure (b) after learning is presented on the right. Figure adapted from Rumelhart and Zipser (1985).

to develop output neurons that respond invariantly. Specifically, it makes use of the spatial correlation between object transforms, such as translation across the retina, in order to develop invariant responses. This is how CT learning differs from trace learning: trace learning relies on the notion that views presented close together in time typically belong to the same object (Foldiak, 1991). Furthermore, CT learning is not the same as the orthogonalisation and categorisation process described above. Specifically, CT learning can cause the same output neuron to learn to respond to two very different input patterns, such as two markedly different views of the same object.

Consider the following: two distinct views of the same object may cause different output neurons to fire due to the orthogonalisation process, as the dot product between the two corresponding input vectors is very small. The two views are too different from one another to be associated together when presented to the network during training. The entire problem of learning invariant representations is concerned with grouping together inputs that are no more similar to one another, in terms of the dot product between their vectors, than they are to other stimuli between which the network must learn

to differentiate. However, it is possible to produce an invariant output representation for these two patterns so long as there are other intermediary patterns; that is, if the two extreme views are also accompanied by a significant number of other in-between views during training. The underlying assumption is that the transforms we wish the model to learn cause the image to change in a gradual and continuous fashion, and therefore small changes in the stimulus produce only small changes in the dot product. This notion plays a recurring role within the research chapters of this doctoral thesis. As such, a simplified example of the CT learning process is illustrated in Figure 1.5 and a description of how it operates is provided next.

The network shown in Figure 1.5 has an input layer where stimuli are presented, and an output layer where transform invariant representations develop through learning. The output layer operates as a competitive network, where individual cells send inhibitory projections to the other cells within this layer and thereby compete with one another. Initially, the weights of the feedforward synaptic connections are set to random values. Then, during learning, a stimulus is first presented in position 1 (shown in Figure 1.5a) and is represented by three active neurons in the input layer (neurons 1, 2, and 3). Activity propagates through the random feedforward connections to the output layer, where one of the neurons (Figure 1.5a, neuron 8) wins the competition due to mutual inhibition within the layer. The simultaneous co-activation of neurons in the input and output layers causes the synaptic connections between them to become strengthened according to a Hebbian learning rule, Equation 1.7.

As the stimulus moves from position 1 to position 2 (shown in Figure 1.5b), it causes activation in the input layer to also move along one neuron. Therefore, when the stimulus is in position 2, it causes neurons 2, 3 and 4 to become active and also causes neuron 1 to become inactive. The overlap in the input space allows two neurons in the input layer to remain active (neurons 2 and 3) during

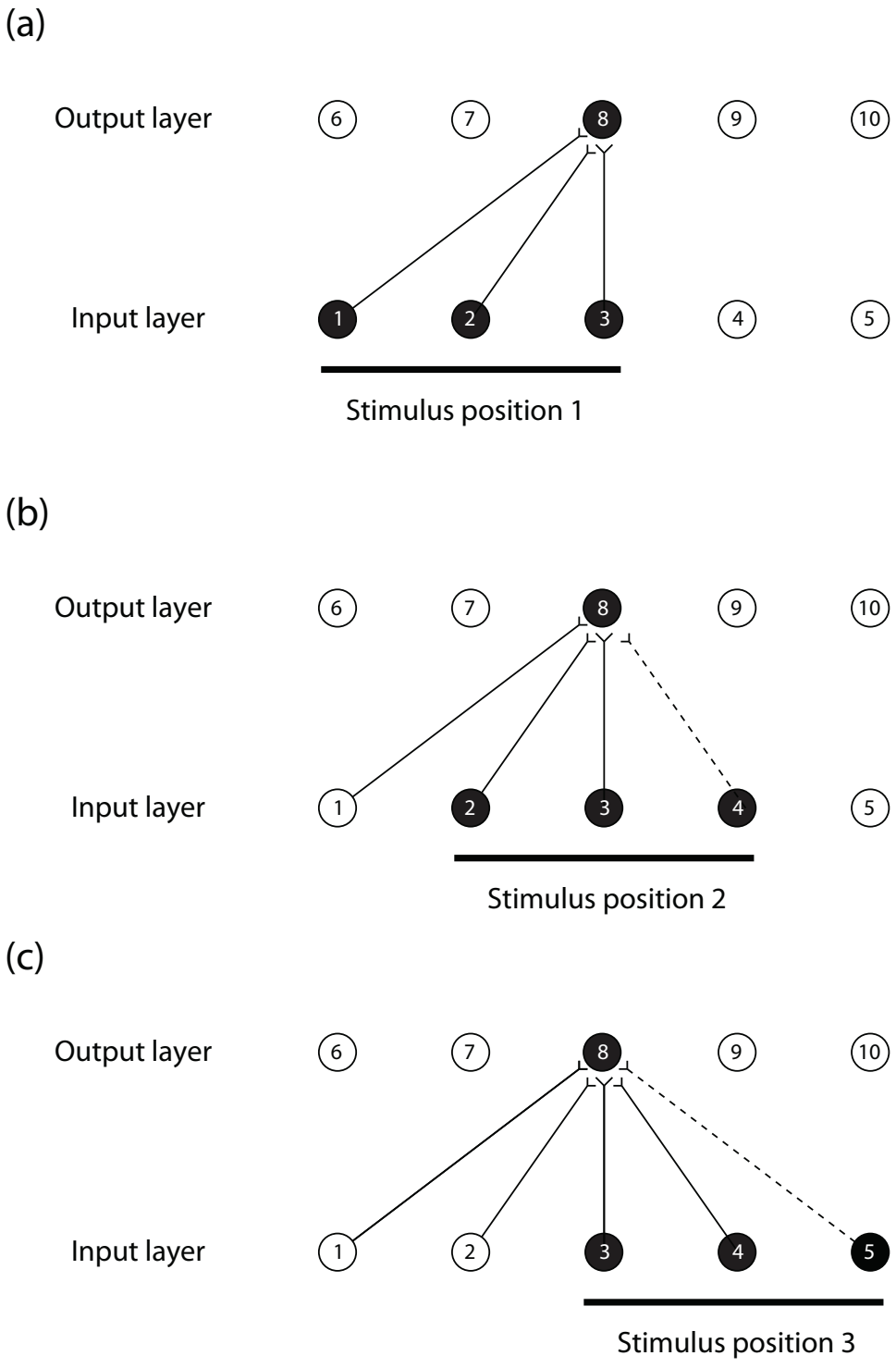

Figure 1.5: An illustration of how CT learning functions in a feedforward one-layer network. Activation of overlapping neurons during the transformation of the object from position to position leads to the activation of the same neuron in the output layer. Connections are strengthened according to a Hebbian learning rule after each presentation of the stimulus. Figure adapted from Stringer et al. (2006)

both transformations. The activation of the same neurons in the input layer substantially increases the likelihood that the same neuron in the output layer (neuron 8) will become active again because the connections have already been strengthened when the stimulus was presented in position 1. The simultaneous activation of the output neuron with input neurons 2, 3 and the additional input neuron 4 causes their synaptic connections to become strengthened according to the Hebbian leaning rule. Therefore, the activation of neuron 8 will now become associated with the activation of neurons 2, 3 and 4. As the stimulus continues to move, as shown in Figure 1.5c, from one position to the next, the process repeats itself and the same neuron in the output layer remains activated. This output neuron becomes a position invariant neuron.

In the case of a more complex model such as VisNet, input layer activation represents the output from the pre-processing input filter stage. The Hebbian learning rule only strengthens the synaptic connections from those V1 filters that are activated by the particular visual form of the object view currently presented. The weight vector of each of the first layer neurons gradually shifts during learning to point in the same direction as the firing rate vectors of the V1 input pattern(s) to which it is learning to respond. After training, each neuron will respond in proportion to the similarity between the current input pattern and the input pattern(s) to which the neuron learned to respond during training; each first layer neuron computes its activation (Equation 1.2) according to the dot product of its weight vector and the current input pattern from the V1 input layer (Hertz et al., 1991).

Self-Organising Map

A number of experiments presented as part of this doctoral thesis make use of a self-organising map (SOM) network architecture. A SOM is a form of intra-layer connectivity whereby neurons excite their nearby neighbours and inhibit neurons that are relatively farther away (von der Malsburg, 1973).

This type of long range inhibition coupled with short range excitation is a simple modification of the competitive learning architecture described above. This type of connectivity is often described as having a Mexican hat spatial profile shape (Kohonen, 1982), presented in Figure 1.6, and enables networks to develop topological maps; neurons that respond to similar features in the input space are encouraged to develop spatially close together in the network. In a similar fashion, neurons that respond to different features in the input space are encouraged to develop farther apart. The same simple Hebbian learning rule utilised in standard competitive learning may also be used in a SOM, thus making this type of self-organisation biologically plausible. However, it should be noted that von der Malsburg (1973) did not claim a direct relationship between cells in the actual cortex and the two different types of cells responsible for excitatory and inhibitory connections in a SOM. Nonetheless, a SOM architecture is particularly appealing to computational modellers working in the neuroscience fields because it can account for a number of key observations regarding how the brain self-organises.

For example, experimental findings support feature maps in the visual cortex, such as orientation tuning and ocular dominance columns in V1 (Miller, 1994; Harris et al., 1997). Also, neurons in IT demonstrate characteristics that would be compatible with a SOM-like architecture. Kiani et al. (2007) examined more than six hundred neurons in IT cortex while monkeys performed a passive fixation task with natural and artificial object images. It was reported that cells close to one another tended to have similar stimulus and category selectivity. Cells with similar category selectivity appeared in multiple small clusters distributed over the recorded region and the categorical structure of objects was represented by the pattern of activity distributed over this entire cell population. Other arguments have been made regarding the biological utility of developing topology-preserving feature maps. For example, Rolls and Deco (2002) have suggested that if neurons with similar responses are required to

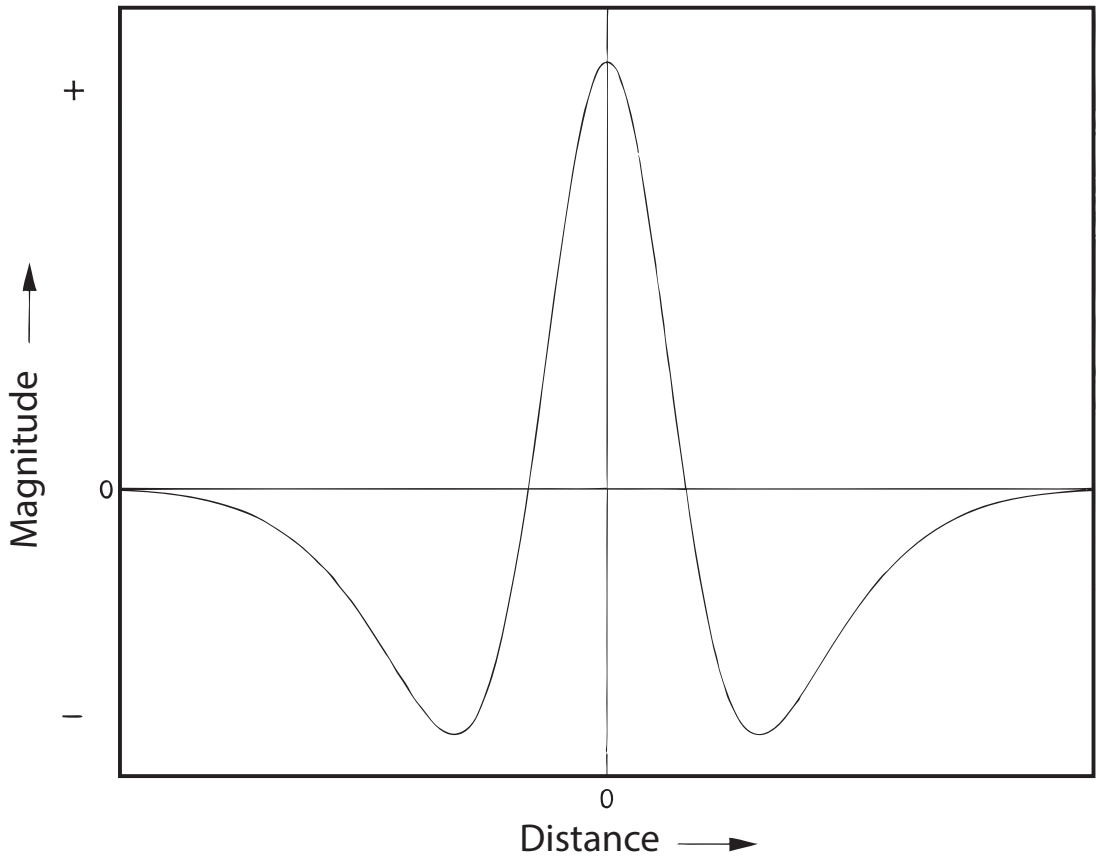

Figure 1.6: Mexican hat spatial profile. The figure presents the Mexican hat lateral spatial interaction profile used within SOMs (Kohonen, 1982). The profile helps show how neurons excite their nearby neighbours and inhibit neurons that are relatively farther away.

communicate information between one another more often than the neurons that respond differently from one another, then the total length of the neuronal connections will be minimised. It may also be argued that this type of organisation will also help minimise the complexity of rules required to specify cortical connectivity (Rolls and Stringer, 2000).

One such biologically realistic implementation of the SOM learning architecture is the LISSOM (Laterally Interconnected Synergetically Self-Organising Map) model. This model, and its variants (e.g. RF-SLISSOM, Receptive Field Spiking LISSOM), have repeatedly demonstrated the self-organisation of a variety of functional maps in V1, including that of ocular dominance columns (Sirosh and Miikkulainen, 1994), orientation and direction preference maps, and even long range excitatory collaterals (Miikkulainen et al., 2005) which accurately resemble those recorded *in vivo* (Bosking et al., 1997). However, current experimental evidence suggests that short range lateral cortical connections in V1 are inhibitory (Tucker and Katz, 2003) whilst long range excitatory collaterals connect functionally similar areas (Bosking et al., 1997). Although this evidence contradicts local excitation and longer range inhibition, an effective functional Mexican-hat profile can be captured nonetheless, provided that inhibition acts over a relatively short time scale (Kang et al., 2003). In summary, although we do not yet know the precise anatomical wiring of the visual cortex, suffice it to say that SOMs have so far successfully captured the important details of map formation in the visual cortex, including the formation of feature maps of abstract visual spaces, as presented in this doctoral thesis and by Michler et al. (2009).

1.3.3 Model Architecture

Section 1.1 provides a neurophysiological account of how object recognition in the ventral visual pathway could operate. Section 1.1.5 specifically implicates the role of IT in invariant recognition

of both objects and faces. The previous section, 1.2, argues for a feature hierarchy model of computational object recognition. The following text provides a detailed overview and model architecture description of VisNet, the computational model of the ventral visual system used to produce the results presented in this doctoral thesis.

To recap, VisNet is based on the following: (i) A series of hierarchical competitive networks or SOMs with local graded 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 Hebb-like associative learning rule. Figure 1.3 previously presented shows a schematic diagram of the VisNet hierarchy and the associated brain regions that each layer could model.

	Dimensions	Number of Connections	Radius
Layer 4	32x32	100	12
Layer 3	32x32	100	9
Layer 2	32x32	100	6
Layer 1	32x32	272	6
Retina	128x128x32	-	-

Table 1.1: Network dimensions. The number of connections per neuron and the radius in the preceding layer from which 67% of connections are received

The forward connections to individual cells are derived from a topologically corresponding region of the preceding layer, using a Gaussian distribution of connection probabilities. These distributions are defined by a radius which will contain approximately 67% of the connections from the preceding layer. Typical values are provided in Table 1.1. Prior to presenting the visual stimuli to the network's input layer, they are pre-processed by a set of input filters which accord with the general tuning profiles of simple cells in V1. The filters provide a unique pattern of filter outputs for each visual stimulus, which is passed through to the first layer of VisNet. Figure 1.7 outlines the filter sampling paradigm. The input filters used are computed by weighting the difference of two Gaussians by a third orthogonal

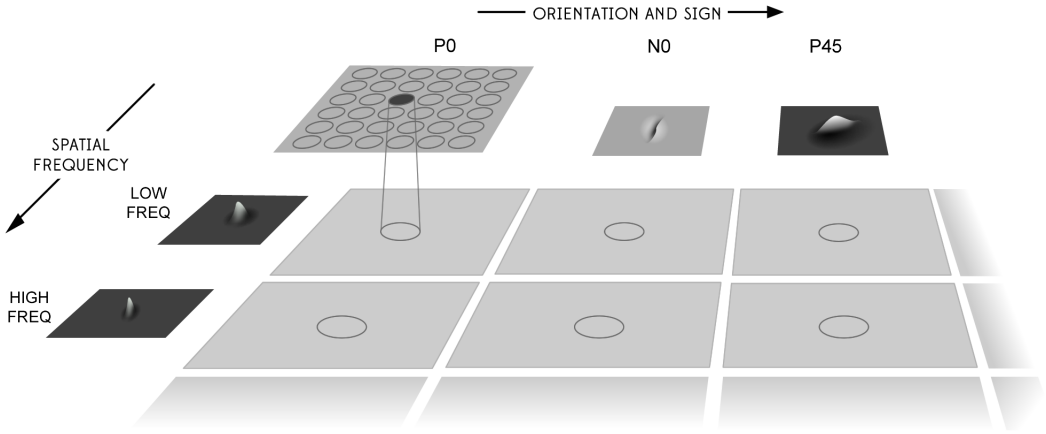

Figure 1.7: The filter sampling paradigm. Here each square represents the retinal image presented to the network after being filtered by a Difference-of-Gaussian filter of the appropriate orientation sign and frequency. The circles represent the consistent retinotopic coordinates used to provide input to a layer 1 cell. The filters double in spatial frequency towards the reader. Left to right the orientation tuning increases from 0 ° in steps of 45 °, with segregated pairs of positive (P) and negative (N) filter responses. Figure reproduced from Rolls and Deco (2002).

Gaussian according to the following:

$$\Gamma_{xy}(\rho,\theta,f) = \rho\left[e^{-(\frac{xcos\theta+ysin\theta}{\sqrt{2}/f})^2} - \frac{1}{1.6}e^{-(\frac{xcos\theta+ysin\theta}{1.6\sqrt{2}/f})^2}\right] e^{-(\frac{xsin\theta-ycos\theta}{3\sqrt{2}/f})^2} \tag{1.1}$$

where f is the filter spatial frequency, θ is the filter orientation, and ρ is the sign of the filter, i.e. $\pm$ 1. Individual filters are tuned to spatial frequency (0.0625 to 0.5 cycles/pixel); orientation (0^o to 135^o in steps of 45^o); and sign ($\stackrel{+}{-}1$). The number of connections to each cell in layer 1 from each spatial frequency filter group is given in Table 1.2. Past neurophysiological research has shown that models based on Difference-of-Gaussians functions can account for the variety of shapes of spatial contrast sensitivity functions observed in cortical cells similar to those based on the Gabor function or the second differential of a Gaussian (Hawken and Parker, 1987).

The activation h_i of each neuron i in the network is set equal to a linear sum of the inputs y_j from

Frequency	0.5	0.25	0.125	0.0625
Number of Connections	201	50	13	8

Table 1.2: Layer 1 connectivity. The numbers of connections to each layer 1 cell from each spatial frequency set of input filters are shown. The spatial frequency is in cycles per pixel.

afferent neurons j weighted by the synaptic weights w_{ij}. That is,

$$h_i = \sum_j w_{ij} y_j \tag{1.2}$$

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

The original VisNet model implemented only a competitive network within each layer. This competition is graded rather than winner-take-all, and is implemented in two stages. To implement lateral inhibition, the activation h of neurons within a layer are convolved with a spatial filter, I, where δ controls the contrast and σ controls the width, and a and b index the distance away from the centre of the filter

$$I_{a,b} = \begin{cases} -\delta e^{-\frac{a^2+b^2}{\sigma^2}} & \text{if } a \neq 0 \text{ or } b \neq 0, \\ 1 - \sum_{\substack{a \neq 0 \\ b \neq 0}} I_{a,b} & \text{if } a = 0 \text{ and } b = 0. \end{cases} \tag{1.3}$$

Typical lateral inhibition parameters are given in Table 1.3.

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

Table 1.3: Lateral inhibition parameters

In addition to lateral inhibition, some simulations within this doctoral thesis implement a self-organising map (SOM) within each layer. In the case of the SOM architecture, short-range excitation and long-range inhibition forms a Mexican-hat spatial profile and is constructed as a difference of two

Gaussians as follows:

$$I_{a,b} = -\delta_I e^{\left[-\frac{a^2+b^2}{\sigma_I^2}\right]} + \delta_E e^{\left[-\frac{a^2+b^2}{\sigma_E^2}\right]} \tag{1.4}$$

To implement the SOM, the activations h_i of neurons within a layer are convolved with a spatial filter, I , 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. a and b index the distance away from the centre of the filter. Typical lateral inhibition and excitation parameters are given in Table 1.4.

Layer	1	2	3	4
Excitatory Radius, σ_E	2.1	1.65	1.2	1.8
Excitatory Contrast, δ_E	5.35	33.15	117.57	120.12
Inhibitory Radius, σ_I	4.14	8.1	12.0	18.0
Inhibitory Contrast, δ_I	1.5	1.5	1.6	1.4

Table 1.4: Example lateral inhibition and excitation parameters for the SOM

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

$$y_i = f^{sigmoid}(r_i) = \frac{1}{1 + e^{-2\beta(r_i - \alpha)}} \tag{1.5}$$

where r_i is the activation (or firing-rate) after applying the lateral inhibition and/or excitatory filter, y_i is the firing-rate after contrast enhancement, and α and β are the sigmoid threshold and slope, respectively. The parameters α and β are constant within each layer, although α is adjusted to control the sparseness of the firing-rates. The sparseness a of the firing within a layer can be defined, by

extending the binary notion of the proportion of neurons that are firing, as

$$a = \frac{(\sum_{i=1}^{N} y_i/N)^2}{\sum_{i=1}^{N} y_i^2/N} \tag{1.6}$$

where y_i is the firing rate of the ith neuron in the set of N neurons (Rolls and Treves, 1990, 1998). For the simplified case of neurons with binarised firing rates $= 0/1$, the sparseness is the proportion $a \in [0, 1]$ of neurons that are active. For example, to set the sparseness to, say, 0.04, the threshold is set to the value of the 96th percentile point of the activations within the layer. Typical parameters for the sigmoid activation function are shown in Table 1.5. These are general robust values found to operate well across different types of VisNet experiment, having been previously optimised to provide reliable performance (Stringer et al., 2006, 2007; Stringer and Rolls, 2008). These are example values and are subject to change depending on the experiment. Specific values are mentioned accordingly where appropriate.

Layer	1	2	3	4
Percentile	96	96	96	96
Slope β	190	40	75	26

Table 1.5: Sigmoid parameters

The inhibitory part of the SOM filter and contrast enhancement stages of the VisNet model aim to simulate the function of inhibitory interneurons. In the brain, inhibitory interneurons affect direct competition between excitatory cells within each layer of the ventral visual pathway. The way in which contrast enhancement is currently implemented in VisNet allows us to control the sparseness of firing-rates within each layer. This is a useful aspect of the model, which allows us to explore the effects of sparseness on network performance.

1.3.4 Learning

An important property of the VisNet model is that the learning at synaptic connections between cells is unsupervised. That is, during training, the exact nature of the stimuli presented, or desired network output, is not explicitly specified to the network to guide learning. The co-activation of neurons in two successive layers causes their synaptic connection to become strengthened, according to a Hebbian learning rule,

$$\delta w_{ij} = k y_i y_j \tag{1.7}$$

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

In addition to Hebbian learning, trace learning is also implemented within simulations presented as part of this doctoral thesis. The trace learning rule is a modified version of the associative Hebbian learning rule, incorporating a temporal trace of activity in the post-synaptic neuron (Foldiak, 1991). Trace learning utilises the temporal continuity of objects in the real world (over short time periods) to help the learning of invariant representations. On the short time scale, of e.g. a few seconds, the visual input to a network is more likely to be from different transforms of the same object, rather than from a different object. As such, trace learning encourages neurons to respond to input patterns which occur close together in time. These input patterns are likely to represent different transforms (e.g. views) of the same object. The trace learning rule used within the experiments presented as part of this doctoral thesis is as follows

$$\delta w_{ij} = k \bar{y_i}^{\tau-1} x_j^{\tau} \tag{1.8}$$

where the trace ${\bar{y}_i}^{\tau}$ is updated according to

$$ {\bar{y}_i}^{\tau} = (1 - \eta)\, y_i^{\tau} + \eta {\bar{y}_i}^{\tau - 1} \tag{1.9} $$

and where x_j is the j^{th} presynaptic input to the postsynaptic neuron i; y_i is the current firing rate of neuron i; $\bar{y}^{\tau}$ is the trace value of neuron i at time-step τ; k is the learning rate (annealed to zero); w_{ij} is the synaptic weight from the j^{th} input to neuron i; η is the trace value. The parameter η may be set anywhere in the interval [0, 1], and is usually set to a typical value of 0.8. A discussion of this trace learning rule in relation to other versions of the trace learning rules is provided by Rolls and Milward (2000).

With both the Hebbian and trace learning rules, in order to restrict and limit the growth of each neuron's synaptic weight vector, $\mathbf{w}_i$ for the ith neuron, its length is normalised at the end of each time-step during training as is usual in competitive learning (Hertz et al., 1991).

Cell output is always positive. This means that the learning rules already outlined will only produce positive weight changes. Weight vector normalisation is required to ensure that the same set of neurons do not always win the competition. Only those synapses where inputs are highly correlated with the output neuron's response will receive high weight values while other weights will decrease in strength. This is a necessary procedure when working with competitive networks (Hertz et al., 1991). Neurophysiological evidence for synaptic weight normalisation is provided by Royer and Pare (2003) but the mechanisms for exactly how this process takes place are still not well understood. To implement weight normalisation the synaptic weights are rescaled to ensure that for each output cell i

$$ \sqrt{\sum_j \left(w_{ij}\right)^2} = 1 \tag{1.10} $$

where the sum is over all input cells j. Such a renormalisation process may be achieved in biological systems through heterosynaptic long-term depression that depends on the existing value of the synaptic weight (Rolls and Treves, 1998).

1.3.5 Information Analysis

Information analysis is used throughout this doctoral thesis as one of the primary methods for measuring network performance. Information in this thesis has the restricted definition of 'measure of surprise', where something 'surprising' conveys a lot of information.

Originally designed to measure the amount of information that can be reliably transmitted by communication channels (Shannon, 1948), information analysis has wide applicability in a number of research areas and has previously been used to quantify information transmitted by IT cells (Rolls et al., 1997; Panzeri et al., 1999; Panzeri and Schultz, 2001). Information analysis was first used to measure the performance of VisNet by Rolls and Milward (2000).

Neurons are inherently noisy and information theory is well suited to deal with this problem because it can measure how much information is available from one or more neurons where the transmission is uncertain. For example, how much information is conveyed by the responses of a set of neurons, R, about a set of stimuli S. More specifically, mutual information is a measure of information that can be obtained about one random variable by observing another, and this resulting quantity conveys the mutual dependence of the two random variables. When no information is transmitted, S and R are statistically independent of one another and there is no mutual dependence between the two random variables. Accordingly, information transmission is a function of the statistical dependence between R and S.

In summary, if the single response, r, from a single stimulus, s is considered, then $p(r, s) =$

$p(r)p(s)$ where r and s are statistically independent. It follows that when r and s are independent, then knowing r does not give any information about s and vice versa, so their mutual information is zero and no information is transmitted,

$$\log_2 \frac{P(r,s)}{P(r)P(s)} = 0 \tag{1.11}$$

where $P(r,s)$ is the joint probability distribution function of r and s, and $P(r)$ and $P(s)$ are the marginal probability distribution functions of r and s, respectively. Logarithms allow for information to be combined by addition rather than by multiplication and because information theory was developed for binary state systems, they are used to the base 2 and measured in bits.

If we sum across all combinations of s and r to ascertain the mutual information between sets S and R, and also weight combinations according to their frequency of occurrence such as when $p(s,r)$ is large, the following standard equation emerges,

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

To apply this measure to VisNet, there are considerations that must be taken into account. Unlike biological neurons, VisNet neurons do not produce noisy responses and as such, repeating the presentation of the same stimuli will cause the exact same responses each time and there is no uncertainty. However, at the time that the information measures were developed the general idea behind the VisNet model was to produce cells that learn to respond invariantly. For example, when the viewing angle of the object varies. This variation in viewing angle can be treated as a source of uncertainty, and $P(s,r)$ can be derived by comparing responses to s over the various transformations of the stimulus.

Therefore, if r is identical for all transforms of s, then $P(r, s) = 1$, and as r starts to vary over some variations of s, then $P(r, s)$ will decrease and therefore the amount of information measured will decrease accordingly.

Information analysis was intended to measure information carried by a set of discrete events that can be assigned different probabilities. In the case of VisNet, R comprises continuously varying values and as such, these values must first be binned into discrete categories. More specifically, if the response r produced by stimulus s is 0.95 on one occasion but 1.0 on another (e.g. after the stimulus has moved slightly on the retina), do they constitute the same or different responses? The answer to this question denotes how much information is carried and as such, the manner in which different responses are binned will have a significant impact on the information measures obtained. Neurophysiological studies also use neuronal firing rate as a measure and can therefore suffer from the same issues.

How to bin the continuous responses may be obvious when neural rates fall into clear discrete levels. Indeed, the binning procedures were originally developed with real neuronal data in mind, which has distributions described by Rolls and Treves (2011). VisNet, however, is normally run with parameters that tend to produce a rather binarised distribution of firing rates, with most neurons firing at zero or close to zero, and a few neurons firing at one or close to one (the maximum rate). The distribution of cell responses obtained from VisNet is clearly important when considering the type of binning that should be used. For example, if equipopulated bins were used with VisNet, then the bins would have a very unusual distribution. For this reason, it is preferable that equispaced bins are used when applying information analysis to VisNet responses.

The next challenge when using information analysis is deciding how many bins there should be.

It is important to avoid over-binning the data, that is, to use more bins that are reasonable given the standard deviation of the neuronal responses. Responses will be interpreted as different on every trial if too many bins are used and subsequently lead to an over-estimate of the amount of information. Bearing these issues in mind, a description is now given of how the specific information measures that are used to characterise VisNet performance are derived.

Single Cell Information Analysis

If all the responses to a single object fall within the same response bin, r_1, while all other responses to other objects fall in a different bin, r_0, then the cell performance is optimal. That is, the stimulus-specific information or 'surprise', $I(s, R)$, conveyed by this cell is maximal. This differs from the mutual information, $I(S, R)$, conveyed by the cell. Mutual information measures the cell's ability to discriminate between the full set of stimuli. When r and s are statistically independent, then $P(r|s) = P(r)$. Accordingly, the object-specific information $I(s, R)$ is the amount of information the set of responses R has about a specific object s, and is given by,

$$I(s, R) = \sum_{r \in R} P(r|s) \log_2 \frac{P(r|s)}{P(r)}, \tag{1.13}$$

where r is an individual response from the set of responses R. This measurement is applied to each stimulus. The maximum value obtained not only denotes the cell's preferred stimulus, but also the cell's maximum information capacity.

It should be noted here that the maximum amount of stimulus-specific information that any individual cell can represent depends on the number of N objects. Accordingly,

$$I(s, R) = log_2 N \tag{1.14}$$

For example, when $N = 9$ there is a maximum amount of 3.17 bits of stimulus specific information that can be conveyed by a neuron about its preferred stimulus. The table below provides an example to demonstrate that the maximum amount of information that can be conveyed by a cell varies with respect to the number of N objects.

	Response Bin 1	Response Bin 2	Response Bin 3	Total
Object 1	0	0	7	7
Object 2	7	0	0	7
Object 3	7	0	0	7
Object 4	7	0	0	7
Total	21	0	7	28

Table 1.6: Example frequency table. An example frequency table used as part of the stimulus specific single cell information analysis. The table shows the response of a specific individual cell to each of the seven different transforms from each of the four objects. Accordingly, there were 28 transforms in total. The three columns represent the three mutually exclusive response bins. Columns providing the grand totals are also presented.

The values presented in Table 1.6, which represent the frequencies with which a specific cell responds to each of the seven transforms per stimulus, are subject to single cell information analysis as outlined in Equation 1.13. Using the stimulus specific single cell information analysis routines, we wish to calculate the amount of information conveyed by this particular cell about 'Object 1'.

First, we calculate the probability of a particular response, 'Response Bin 1', given the particular stimulus, 'Object 1': $P(r|s) = 0/7$. This value equals zero, and the rest of the equation is multiplied by this value, so the first term in the summation loop equals zero. Second, we calculate the probability of a particular response, 'Response Bin 2', given the particular stimulus, 'Object 1': $P(r|s) = 0/7$. This also equals zero, so the second term in the summation loop equals zero. Finally, we calculate the probability of a particular response, 'Response Bin 3', given the particular stimulus, 'Object 1':

$P(r|s) = 7/7$. The second term of the equation is $log_2(p(r|s)/p(r)) = log_2((7/7)/(7/28)) = log_2 4 = 2$. Accordingly, completing the summation loop, there are $0 + 0 + 2 = 2$ bits of information conveyed about 'Object 1' by this specific cell. Furthermore, this is the maximum amount of information available. Importantly, if we were to change the number of objects from 4 to 5, we would obtain $log_2((7/7)/(7/35)) = log_2 5 = 2.32$ bits of information instead.

In summary, the single cell information measure is applied to individual cells and measures how much information is available from the response of a single cell about the stimuli with which it was presented. The single cell information measure for each cell shows the maximum amount of information that the cell conveys about it's preferred object. Further details are provided by Rolls et al. (1997) and Rolls and Milward (2000).

Multiple Cell Information Analysis

The single cell information measure cannot give a complete assessment of VisNet's performance with respect to invariant object recognition. If all output cells learned to respond to the same stimuli then there would in fact be relatively little information available about the set of stimuli, S, and single cell information measures alone would not reveal this. To address these issues, a multiple cell information measure is calculated, which assesses the amount of information that is available about the whole set of objects from a population of neurons.

Multiple cell information analysis measures the amount of mutual information $I(S, R)$ between the set of stimuli, S, and the set of output responses, R. With an increasing number of cells, the process becomes increasingly intractable. More specifically, the larger the numbers of cells measured, the greater the dimensionality of each individual r. Due to possible stochastic effects, the full range of possible values of r must be used but in practice it is usually not possible to collect enough data to

satisfy this requirement.

A solution to this problem exists by mapping the response matrix, R, onto a simpler variable that can then be sampled fully. It should be noted that in the case of VisNet, there is no variability between responses when considering the same stimuli and the same cell. However, the following approach used in neurophysiological studies (Rolls et al., 1997) is also used here to facilitate direct comparison. That is, any given response, r, is mapped to a scalar variable of stimulus identity, s'. This mapping from $S \rightarrow R \rightarrow S'$, allows for the development of probability tables that relate the actual s value to the values of s' thus enabling a practical measure of multiple cell information, $I(S, S')$. Past VisNet studies that make use of information analysis have used maximum likelihood estimation as the method of decoding for multiple cell data and the same method is used within this doctoral thesis. In short, for any given value of r, the probability that this response was generated by each stimulus s from the whole stimulus set, S, is obtained and the most probable stimulus is chosen as s'. Further details are provided by Rolls et al. (1997) and Rolls and Milward (2000).

In summary, from a single presentation of an object, the average amount of information obtained from the responses of all the cells regarding which object is shown is calculated. This is achieved through a decoding procedure that estimates which object s' gives rise to the particular firing rate response vector on each trial. A probability table of the real objects s and the decoded objects s' is then constructed. From this probability table, the mutual information is calculated as,

$$I(S, S') = \sum_{s,s'} P(s, s') \log_2 \frac{P(s, s')}{P(s)P(s')}. \tag{1.15}$$

Multiple cell information values are calculated for the subset of cells which, according to the single cell analysis, have the most information about which stimuli is shown. In particular, the multiple

cell information is calculated from the first five cells for each stimuli that had the most single cell information about that stimuli. For example, this results in a population of 20 cells when there are four stimuli. Previous research found this to be a sufficiently large subset to demonstrate that invariant representations of each object presented during testing were formed, and that each object could be uniquely identified (Stringer and Rolls, 2000).

1.4 Discussion

This chapter reviewed the literature on the ventral visual system. As part of this review, consideration was also given to the retina and LGN. It was suggested that the ventral visual system is primarily feedforward in nature and has a hierarchical structure whereby neurons in the early stages respond to low-order features while neurons towards the end of the ventral visual system respond to more complex objects and faces. Furthermore, in some cases, neurons in inferotemporal cortex learn to respond invariantly, but it was also shown that other neurons are more specific with their responses. For example, neurons respond maximally to an object when it is in a specific part of its receptive field (Op De Beeck and Vogels, 2000) and therefore convey spatial information.

It was acknowledged that communication occurs between the dorsal and ventral visual pathways, suggesting that each pathway is not purely feedforward. Furthermore, back-projections within the ventral visual system were discussed and it is clear that they may play an important role within the learning and self-organisation of the system.

Different computational approaches to object recognition were introduced, followed by a comprehensive review of the existing VisNet modelling literature. This is the first time such a VisNet review has been conducted. It is shown that VisNet models the feedforward nature of the ventral vi-

sual system, and is inspired by neurophysiological research. As part of the VisNet literature review, two important learning mechanisms were introduced: trace learning (Foldiak, 1991) and Continuous Transformation learning (Stringer et al., 2006). The general principles of competition and self-organisation that are implemented within the VisNet model were outlined and a precise account of the model architecture was provided. Finally, an overview of the information analysis measures that are used to assess network performance was given.

The following research chapters build on past VisNet research. They explore the performance of the VisNet model when presented with a more complicated training environment to that used previously. More specifically, the central focus of this thesis is to explore Visnet performance when tasked with learning about individual objects in a multi-object environment. The majority of previous research with VisNet has presented only one object at a time during training. However, it is shown here that VisNet is able to learn about individual objects even when more than one object is presented to the network during training. Different mechanisms for this ability are explored and their application within the primate visual system is discussed. The final chapters present results obtained with a scaled-up version of the VisNet model. This version has a significantly larger retina and more neurons within the four processing layers. The different types of neural representations that form during learning, and the information that the neurons convey about the view or location of a stimulus, are reviewed.

Chapter 2

Learning transform invariant representations with multiple stimuli present during training

This chapter presents original research investigating VisNet's ability to learn individual invariant representations when pairs of objects are presented during training. In particular, how much training is necessary for the successful self-organisation of the model, as defined above? This research builds on previous simulations by Stringer et al. (2007) and Stringer and Rolls (2008). The results reveal that the network requires relatively little training when compared to what was suggested in the previous research. The particular training paradigm and visual environment are important aspects of this research. In particular, the statistics of how visual features and objects occur together.

2.1 Introduction

Neurophysiological research reveals that neurons towards the end of the ventral visual pathway of the primate visual system learn to respond to increasingly complex stimuli, regardless of translation (Tovee et al., 1994; Op De Beeck and Vogels, 2000), view (Hasselmo et al., 1989b; Booth and Rolls, 1998) or size (Rolls and Baylis, 1986; Ito et al., 1995). The resultant effect is that the primate visual system is able to recognise objects in any position or orientation within a visual scene.

A key ability of the the primate visual system is that it is able to learn about novel objects in a cluttered environment without exposure to the novel object in isolation. Early computational modelling research with VisNet addressed the simpler situation of presenting one object at a time (Wallis and Rolls, 1997; Rolls and Milward, 2000; Elliffe et al., 2002; Perry et al., 2006). This is an important and necessary first step required to develop our understanding of how the ventral visual system may learn about objects. However, more recent research by Stringer et al. (2007) and Stringer and Rolls (2008) investigated what types of neuronal representations formed when a one-layer competitive network was trained with pairs of objects. This research investigated performance in both a simple one-layer competitive network, and also in VisNet.

In the case of the one layer competitive network, pairs of abstract stimuli were presented to the network, and these pairings were created from a pool of parent stimuli. In all experiments, regardless of the size of the parent training pool of stimuli, the network was trained with all possible $N(N-1)/2$ paired-stimulus input patterns. For example, when $N = 6$ there were $M = 15$ pairings presented to the network during training. A translation training paradigm was investigated, and transform invariant output representations formed due to CT learning. When the pool of parent stimuli was less than or equal to five, the network failed to develop neurons that represented each stimuli individually and

instead output neurons responded most strongly to the pairs of stimuli with which the network was trained. However, when the parent pool of training stimuli increased to six and above, the output neurons of the one-layer competitive network learned to respond to the individual stimuli instead of the pairs.

In the second part of their research, Stringer and Rolls (2008) investigated the same principle using VisNet. Instead of abstract stimuli, VisNet was presented with pairs of objects rotating over 90°. The pairs of objects were created from a total of nine parent objects. Each one of these objects was presented within a 3x3 grid on the VisNet retina, where each object occupied one of the grid locations exclusively. In total, during training there were 36 object-pairs rotating over 90°. The results show that many of the VisNet output neurons learned to respond to one of the nine objects over all views, and did not respond to any of the other objects from any view.

The results obtained from the controlled one-layer competitive network study, combined with comparable results when using the VisNet model architecture, have important implications. More specifically, they may shed light on how the primate ventral visual pathway may process visual input and form separate representations of individual objects even though multiple objects are always presented together within a visual scene.

The number of object-pairings and the associated amount of training is an important limitation of the research conducted by Stringer and Rolls (2008). In their research, each object was presented with every other object. In the real world, where there are potentially hundreds of novel objects that comprise a scene, the amount of training required would be considerable. This is because the original authors used all possible $N(N-1)/2$ object pairings to train the network. The number of training patterns would increase quadratically with respect to N, and as N increases the amount of training

required becomes intractable. Furthermore, the primate visual system is not systematically exposed to every object in the environment in pairings with every other object.

The aim of this current research is to investigate the ecological validity of the previous research and ascertain whether it is possible for a biologically plausible model of the ventral visual pathway, VisNet, to develop exclusive and invariant representations for individual objects (rotating over 360 °) without the need for such an exhaustive training regime, as used by Stringer and Rolls (2008). More specifically, even as the number of objects that comprise the parent training pools, N, increases, the number of other objects with which each object is paired, M, does not need to also increase. Accordingly, unlike the previous research, the current study is concerned with the variation of M, and how the network performs as M is manipulated with respect to N.

It was hypothesised that the output neurons of the VisNet model would learn to respond to the pairs of stimuli that were used to train the network when the value of M is small. However, as the value of M increases, VisNet would learn to represent each of the individual objects despite the fact that they were always presented as pairs to the network during training. Most importantly, it was hypothesised that the value of N should not impact the value of M required for VisNet to develop separate invariant representations of each object. Supporting this hypothesis, the simulation results presented within this chapter show that where $M \geq 3$, output neurons within VisNet learn to represent individual objects invariantly. For smaller values of M, the network learns to represent the object-pairs with which it was actually presented during training. Furthermore, the effect of sparseness is also investigated. This implies that much less training is necessary than originally proposed by Stringer and Rolls (2008), thus substantially improving the ecological validity of this mechanism.

2.2 Method

All experiments within this chapter consist of presenting pairs of objects to the network during training, where the objects rotate together through 360° in lock-step. During testing, individual objects are presented to the network in order to establish whether the network has learned to develop invariant responses, and whether these invariant responses are exclusive to a preferred object. Since objects are always presented as pairs during training, rotating together in lock-step, CT learning theory indicates that the different views of an individual object will be bound together into one holistic representation. However, because objects are always presented as pairs during training, a certain amount of pair decoupling is necessary in order for the network to learn to separate the individual objects from one another.

2.2.1 Stimuli

Figure 2.1 shows the objects used to train the network. In total, there were up to N = 9 continuously rotating 3D objects on a grey background. The objects were designed and created using the 3D modelling tool Swift 3D [1]. Ambient lighting with a diffuse light source was added to allow different surfaces to be shown with different intensities. Each object rotated in depth around the vertical axis in 1° steps through 360°. The 360 views of each object were then exported as 2D JPG images and encapsulated as Adobe Shock Wave Files. The objects were then aligned and organised using Adobe Flash CS4. Specifically, objects were arranged into all possible pairings such that the left and right halves of the image canvas contained each of the two objects that together comprise the pair. For every degree that the objects rotated together in lock-step, a single PNG image was generated and subsequently exported. Accordingly, for every rotating pair there is a sequence of 360 images. Prior to

[1] Swift 3D can be obtained from publisher Electronic Rain, http://www.erain.com/

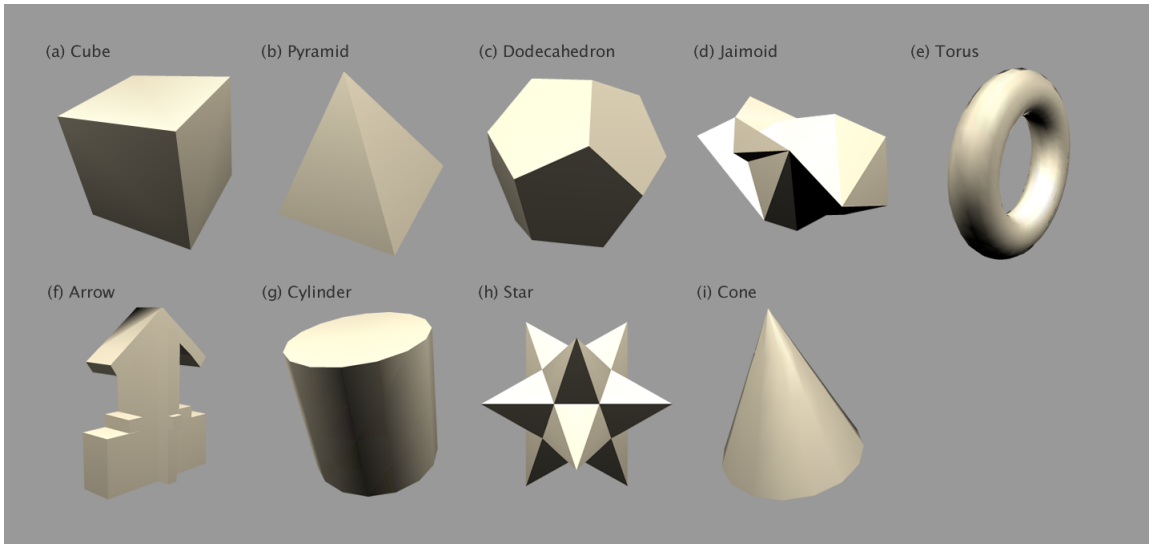

Figure 2.1: Nine training and testing stimuli. This figure shows an example frame of each of the nine objects used to train and test the model: (a) Cube; (b) Pyramid; (c) Dodecahedron; (d) Jaimoid; (e) Torus; (f) Arrow; (g) Cylinder; (h) Star; (i) Cone. The Jaimoid is an irregular multifaceted three dimensional object.

their presentation to the network, images were filtered in accordance with the Difference-of-Gaussian equation presented previously in Equation 1.1.

These $N = 9$ objects together comprise the total set of objects from which the training and testing stimuli were created. The canvas was divided into a 3x3 grid creating a total of nine possible locations. Each object was assigned to an orthogonal location and therefore the nine different objects did not overlap. The spatial arrangement of the objects is shown in Figure 2.2.

2.2.2 Model Architecture and Parameters

The experiments presented as part of this chapter were conducted using the VisNet model architecture as described in section 1.3.3. The specific sparseness percentile and slope values implemented are presented in Table 2.1. The lateral inhibition parameters are provided in Table 2.2.

Six different sparseness percentile values are used within the experiments and they are presented within Table 2.1. Past VisNet research typically refers to ‘sparseness percentile’ as the percentage of

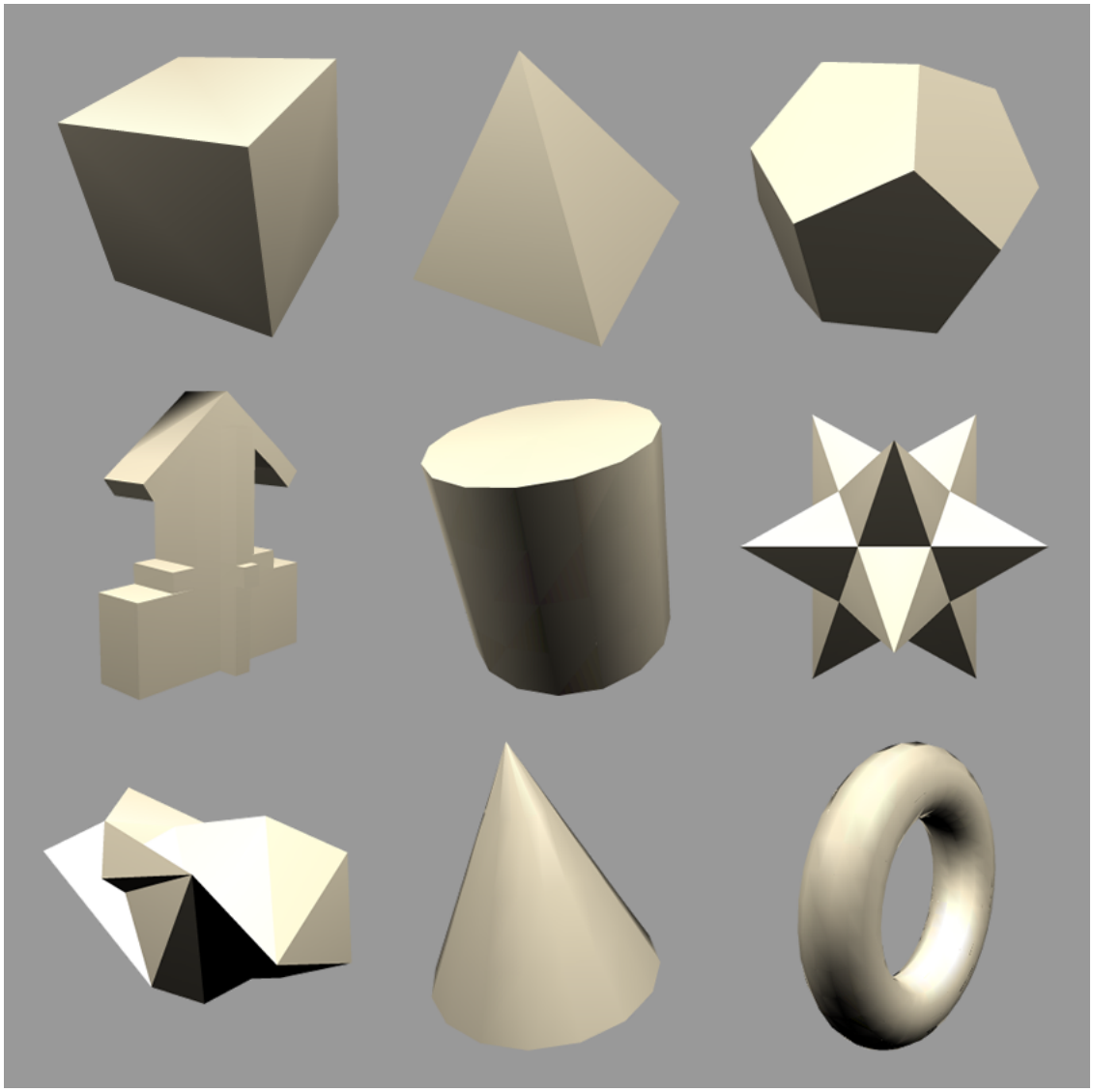

Figure 2.2: Stimuli arrangement. This figure shows the 3x3 arrangement of the nine stimuli used during training and testing. Each object occupies one of the nine possible locations: cube (top left); pyramid (top centre); dodecahedron (top right); arrow (left); cylinder (centre); star (right); Jaimoid (bottom left); cone (bottom centre); torus (bottom right).

cells that are not allowed to fire within a given layer. On the other hand, 'sparseness' is the opposite value; the percentage of cells that are allowed to fire. For example, when the sparseness percentile is 94.0, only 6% of the cells within the layer are allowed to fire, and this is reported as 0.06 and termed 'sparseness'. For consistency, these existing VisNet reporting conventions are maintained within this chapter and throughout the rest of the thesis. Accordingly, the term 'sparseness' is the preferred reporting value and will be used unless otherwise stated.

As with past VisNet research, typical sparseness values range between 0.12 and 0.02 depending on the number of stimuli used during training and the learning rule implemented during learning. However, a range of values are usually explored in order to ensure that the results obtained are robust across a number of different sparseness value. For strong effects, such as CT learning (Stringer et al., 2006), results are normally robust to a very wide range in sparseness values greater than the range outlined previously. With more sensitive effects, the sparseness value may have an interesting and qualitative impact on network performance. In this situation, the role that the sparseness plays during learning is systematically explored and the results are discussed accordingly.

Layer	1	2	3	4
Percentile A	88.0	88.0	88.0	88.0
Percentile B	90.0	90.0	90.0	90.0
Percentile C	92.0	92.0	92.0	92.0
Percentile D	94.0	94.0	94.0	94.0
Percentile E	96.0	96.0	96.0	96.0
Percentile F	98.0	98.0	98.0	98.0
Slope β	190	40	75	26

Table 2.1: Sigmoid parameters

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

Table 2.2: Lateral inhibition parameters

2.2.3 Training Procedure

Training consists of presenting pairs of rotating objects to the network where individual objects are never presented in isolation. During training, 360 images are presented to the network per pairing. This is because the objects rotate through 360° together in lock-step displayed in Figure 2.3. Each degree change is presented as a separate image. Upon presentation, the activation of individual neurons within a layer is calculated, then their firing rates are calculated, and the feedforward synaptic weights between layers w_{ij} are updated according to Equation 1.7. This process is repeated for each of the four layers in turn. A single training epoch comprises the presentation of all of the images from each of the pairings presented as part of the experiment. The network is trained one layer at a time starting with layer 1 and finishing with layer 4. The number of training epochs per layer and learning rates are presented in Table 2.3 and Table 2.4, respectively.

Layer	1	2	3	4
Number of epochs	50	100	100	75

Table 2.3: Number of training epochs per layer.

Layer	1	2	3	4
Learning Rate	0.01	0.01	0.01	0.01

Table 2.4: Learning rates per layer.

The total number of N objects and the number of other M objects with which each object is paired during training are both systematically varied. By varying N and M independently, it is possible to explore the effect each has on the network's ability to build separate representations of individual objects. For every possible value of N objects (up to a maximum of $N = 9$), every possible value of M pairings is explored, each as a separate experiment. For example, if $N = 9$ objects then there is a maximum of $M = 8$ possible pairings per object giving rise to a total of ${}_9C_2$ (36) possible stimuli

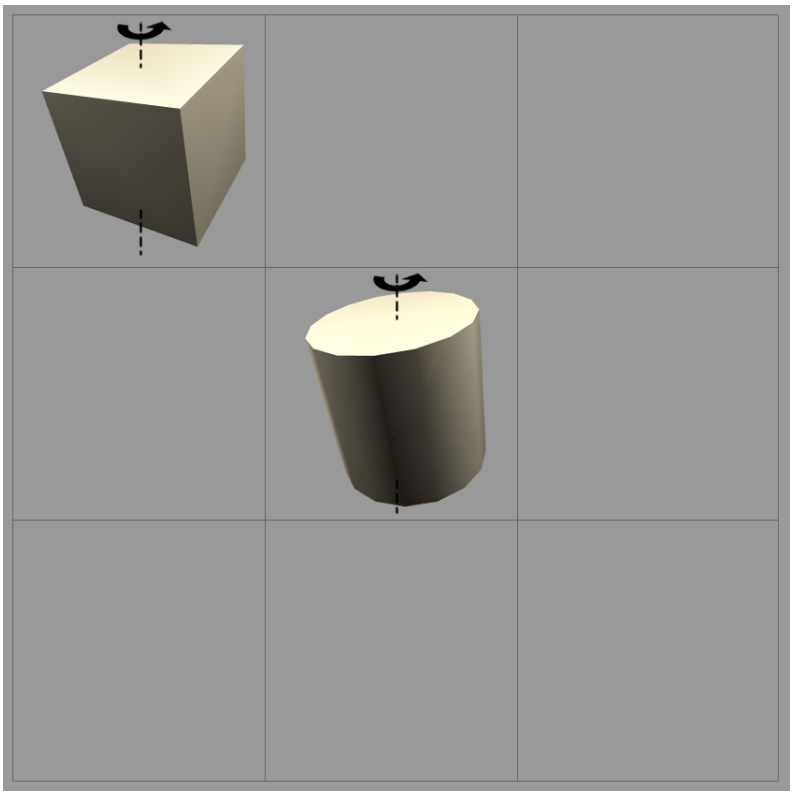

Figure 2.3: Example frame of the Cube and Cylinder rotating together in lock-step. This figure shows one example frame of the Cube and Cylinder in place on the VisNet retina. Presented on the figure for illustrative purposes only are the vertical axis for each object and the direction of rotation. Furthermore, the light grey grid lines are presented purely for illustration and are not presented to the VisNet model under any condition.

pairings. Similarly, when $N = 9$ there are a total of 8 sets of experiments. When $N = 8$ there are a total of 7 sets of experiments. The method of systematic variation is as follows. The value of N is decreased gradually from 9 to 2, and for every value of N, the value of M is decreased gradually from $N - 1$ to 1. Specifically, when a new N value is used, M is always reset to be $N - 1$. Each experiment is repeated 5 times and on each occasion a different random seed is used. The random seed specifies the network's initial connectivity profile, including which neurons are connected to one another between layers, and also the strength of these connections. A different random seed is used for each experiment ensuring that the results obtained are not due to a particular synaptic weight configuration established prior to learning.

Previous research has explored the effect of varying N when the maximum possible number of M pairings is presented to the network during training. The parameter M is of particular interest within this chapter. The purpose of this study is to investigate the effect of varying M with respect to N and careful measures are taken to ensure that, during a particular experiment, the objects used as part of the pairings are presented equally. Wherever possible, the frequency with which each object is presented is identical with respect to the other objects.

2.2.4 Testing Procedure

During testing, the synaptic weights between different neurons remain fixed as there is no learning. Individual objects are presented to the network rotating around the vertical axis through 360° and an example object is presented in Figure 2.4. For a given experiment, only those objects that were presented during training are used during testing. For example, when $N = 5$ then the network is only tested to see how it has learned to respond to each of the 5 objects.

Each object is presented individually so that the extent to which the network has learned to build

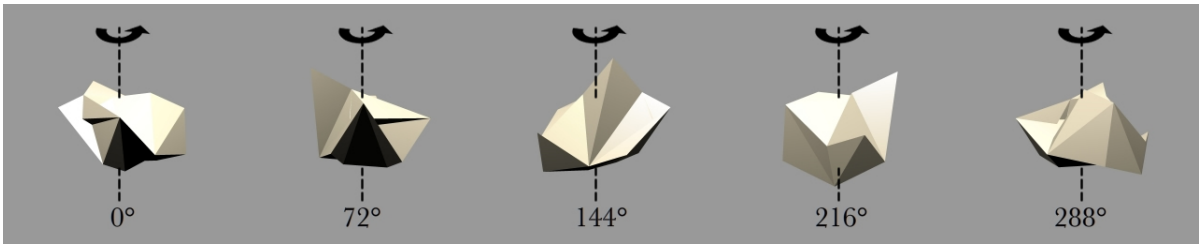

Figure 2.4: Example object rotating during testing. This figure presents five equidistant frames of the Jaimoid rotating over 360 °: 0 °; 72 °; 144 °; 216 °; and 288 °. Each of the nine objects rotates around its vertical axis in an identical manner, located within one of the cells that comprise the 3x3 grid on the VisNet retina.

separate invariant representations can be explored. The complete presentation of an individual object consists of presenting all 360 images as the object rotates around the vertical axis through 360 °. After image presentation, the response of every neuron within the network is calculated. By exploring the responses of neurons from all four layers of the network, it is possible to investigate how earlier layers may also represent invariance information. However, it is the output neurons of the network which receive the most comprehensive analysis as this is the layer where completely invariant responses usually form. That is, invariant responses are usually only possible in the output layer of the model due to the convergent connectivity between layers. For example, if a neuron in the output layer has learned to respond to all 360 views of an object, and does not respond to any views of any of the other objects, then it is a cell that has learned to respond invariantly and exclusively to just one object.

Using the testing methodology outlined above, the experiments presented within this study explore the types of representations that form when pairs of objects are presented to the network during training.

2.3 VisNet Simulation Results

It is shown that as the value of M varies for each value of N, the VisNet output layer shifts from representing object pairs to representing individual objects. In both cases, the network learns to respond invariantly to either the pairs or individual objects through all 360 $^\circ$ of rotation. In particular, it is shown that as the value of N varies, the value of M that is necessary to achieve separate individual object representations is constant for different N.

2.3.1 The effect of varying the number of pairings for each stimulus during training

As outlined in the Training Procedure in section 2.2.3, simulations were run with a possible total of $N = 9$ objects, where the value of N is gradually decreased throughout the different simulation sets. For each set of simulations with a particular value of N, the number of M objects with which each object is paired during training is systematically varied. In all experiments, the network was trained to completion, where a stable and steady state is observed.

Table 2.5 presents the results of all possible N and M combinations for sparseness percentile 94.0 across all four layers (sparseness of 0.06) for the first random seed only. Each value presented within the table represents the total number of neurons within the output layer of the VisNet model that learned to respond invariantly and exclusively to one of the N objects. It can be seen that when $M = 1$, there are no cells that have learned to respond to just one object invariantly, and exclusively. Instead, neurons learn to respond to object pairs rather than individual objects. When $M = 2$, it can be seen that for every value of N there are now cells that have learned to respond invariantly and exclusively to individual objects (e.g. $N = 6$ and $M = 2$ there are 34 'exclusively invariant' output cells), however, the majority of cells still respond to object pairs. For example, when $N = 9$ and

$M = 2$ there are 237 cells that respond invariantly to complete object pairs (not presented within the table). As M increases to $M \geq 3$, it can be seen that for every value of N there are now many more cells that have learned to respond invariantly and exclusively to individual objects. Dashes indicated that experiments could not be run with this particular combination of N and M.

	N = 2	N = 3	N = 4	N = 5	N = 6	N = 7	N = 8	N = 9
M = 1	0	0	0	0	0	0	0	0
M = 2	–	50	27	94	34	41	62	12
M = 3	–	–	390	414	422	390	465	400
M = 4	–	–	–	420	456	404	440	426
M = 5	–	–	–	–	460	427	473	417
M = 6	–	–	–	–	–	382	419	404
M = 7	–	–	–	–	–	–	404	405
M = 8	–	–	–	–	–	–	–	443

Table 2.5: Table of results for all possible N and M combinations. The values in the columns represent the number of exclusively invariant cells for each specific value of N that developed in the output layer of the model over the course of training with a sparseness value of 0.06, for the first random seed only. For all values of N, when there is only one pairing per object ($M = 1$) the network is unable to build separate invariant representations of individual objects. However, when the value increases to $M = 2$, for all values of $N > 2$, there are a number of exclusively invariant neurons that respond to each object. As the number of pairings per object increases to $M = 3$, the number of cells that learned to respond exclusively to just one object increases considerably. For values of $M > 3$ there are comparable numbers of exclusively invariant cells regardless of the increasing M value.

The summary results in table 2.5 show that as the number of pairings per object is increased past $M = 3$, the number of exclusive invariant cells does not continue to increase. Figure 2.5 plots a portion of the data displayed in the table. For $M = 1$ to $M = 5$ (x-axis), the number of exclusive invariant cells (y-axis) are plotted as line plots for $N = 6$ to $N = 9$. It can be seen that as the value of M increases past $M = 3$, the line plots representing the number of exclusive invariant cells level out. The number of exclusively invariant cells remains approximately constant for the additional values when $M > 5$, independent of N (not presented in Figure 2.5).

Figure 2.6 and Figure 2.7 show the cell response plots for prototypical output Cell 24, 26 (row 24, column 26) when $N = 9$ and $M = 3$ before and after training, respectively. Prior to training, cells fail

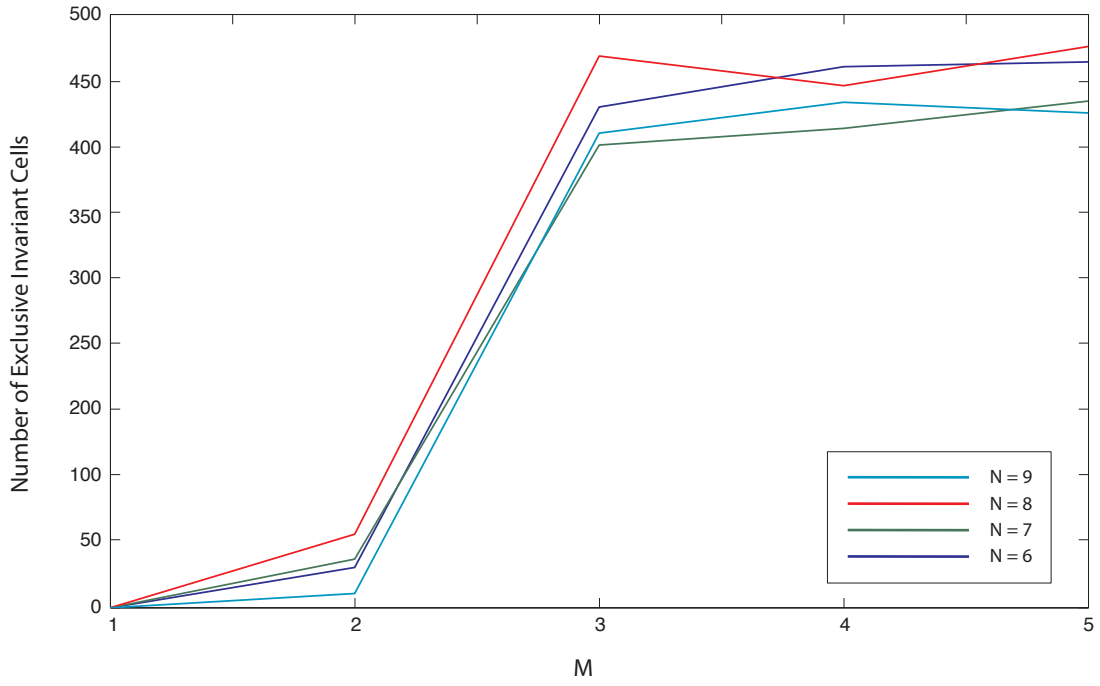

Figure 2.5: The effect of varying M with respect to N. This figure clearly shows the effect of reduced training. For the different N values (6, 7, 8 and 9) presented within the figure, the value at which cells start to respond exclusively and invariantly to their preferred object is the same, when $M = 3$. Importantly, this value is far less than that used by Stringer and Rolls (2008).

to respond exclusively or invariantly to any object and either respond to sporadic individual views, or to no views at all. After training, Cell 24, 26 has learned to respond invariantly to a preferred object, the Pyramid. This output cell only responds to the Pyramid and does not respond to any of the views of any of the other objects.

Figure 2.8 presents the cell response plot after training for prototypical Cell 17, 18 for when $M = 2$ and $N = 9$. This cell has learned to respond invariantly to both the Cone and the Cube which together comprise one of the pairs presented during training. Each of the different pairs presented during training is represented by a cell that has learned to respond to the whole pair in a similar manner to Cell 17, 18.

2.3.2 The effect of varying the level of competition within the network

The sparseness parameter governs the global inhibition within a layer and as such, it controls how many cells may become active when an image is presented to the network. When the sparseness value is high, more cells remain active and their connections with the firing pre-synaptic neurons will become strengthened during training. The results presented so far are obtained from simulations with a sparseness value set to 0.06 for all four layers of the VisNet model. A range of different sparseness values was explored. In particular, the values 0.12, 0.10, 0.08, 0.06 and 0.04 were investigated in detail and the variations in the results obtained are presented next.

For the different sparseness values, each simulation set was conducted five times, each set with a different random seed. It was found that varying the sparseness percentile influences the distribution of cells between objects. For example, when the sparseness was 0.02, the number of cells that responded invariantly to each object were not evenly distributed. Upon changing the random seed, the specific object to which cells learned to respond invariantly also changed. For example, for simulation set

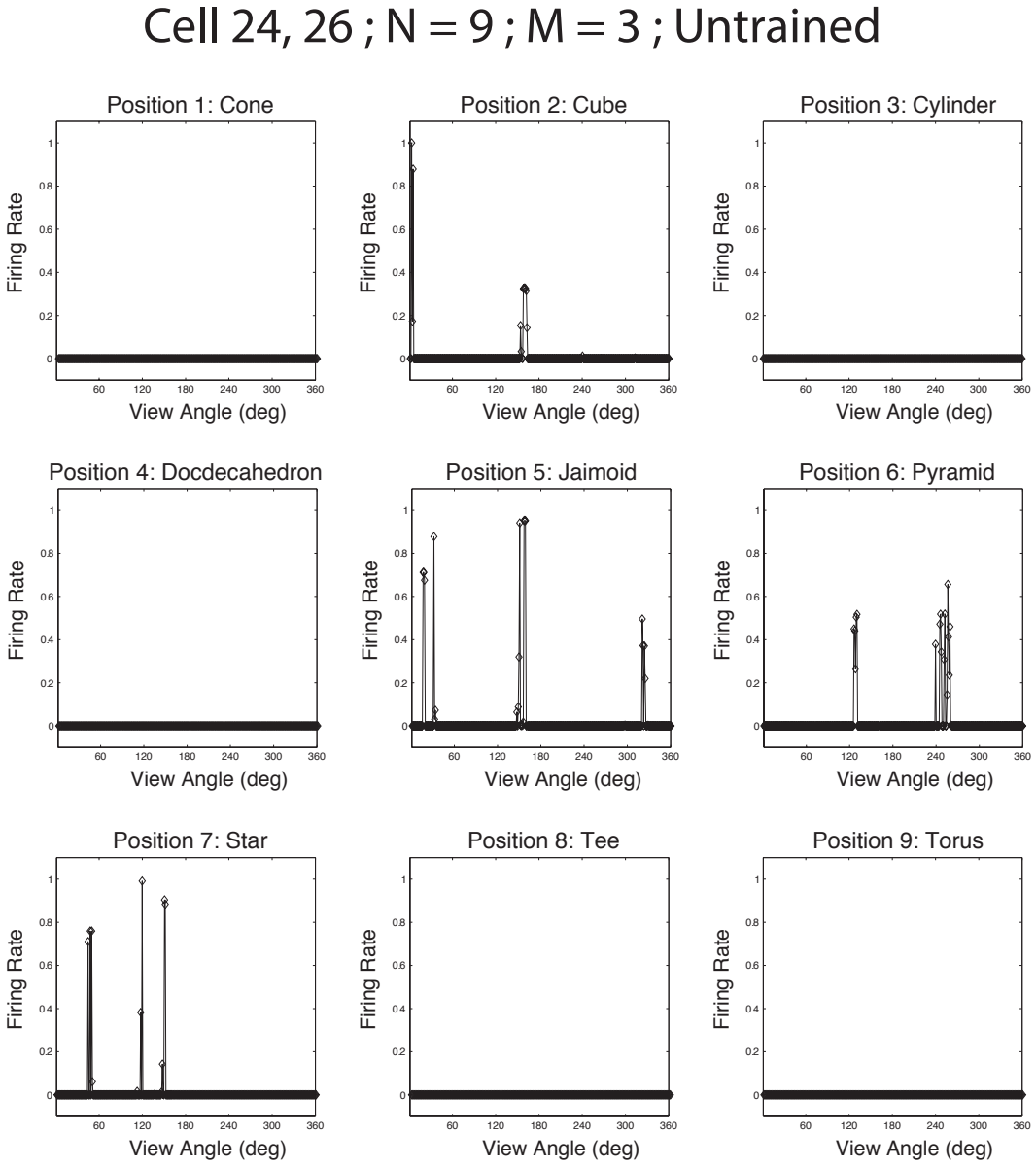

Figure 2.6: Untrained cell response plots for Cell 24, 26; N=9; M=3. It can be seen that before training this cell does not respond invariantly or with preference to any object used during training. This cell responds sporadically and randomly to certain views of the nine different objects.

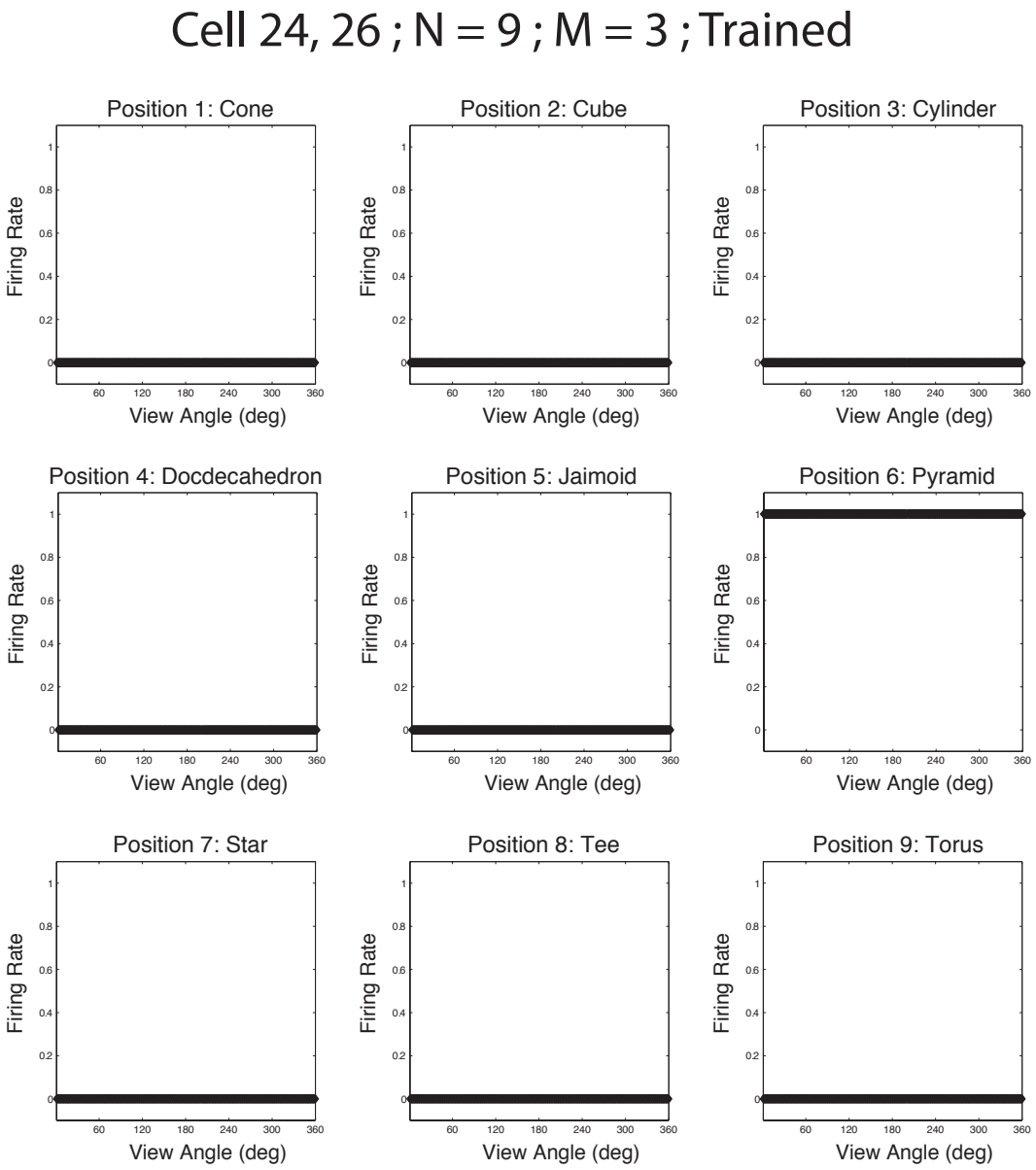

Figure 2.7: Trained cell response plots for Cell 24, 26; N=9; M=3. It can be seen that after training this cell has learned to respond invariantly to a preferred object, the Pyramid. This cell does not respond to any of the other views of any of the other objects. Every other object is also represented in this way by other cells within the output layer.

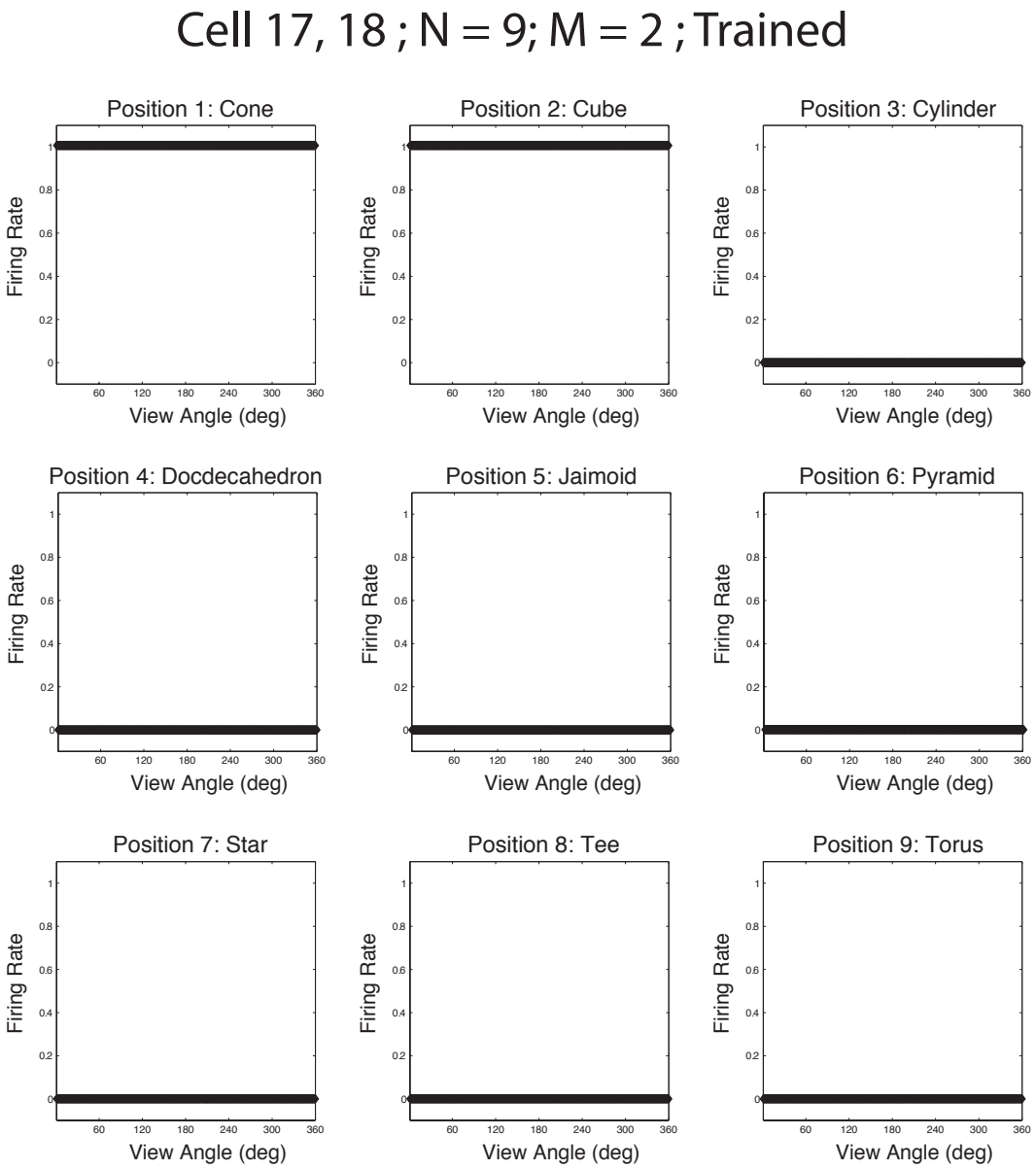

Figure 2.8: Trained cell response plots for Cell 17, 18; N=9; M=2. It can be seen that after training this cell has learned to respond to a pair of objects invariantly, the Cone and the Cube. It does not respond to any of the other views of any of the other objects. All of the other pairs that were presented to the network during training are also represented in this way.

$N = 6$ and $M = 3$ using the first random seed, only 5 cells learned to respond exclusively and invariantly to the Cube while for the second random seed, 82 cells responded to the Cube.

Table 2.6 presents the standard deviation of the number of exclusive invariant cells that respond to each object when $N = 6$ and $M = 3$. Each standard deviation is obtained when using a different random seed, and the pooled standard deviation is presented as a summary statistic. As the sparseness value increased from 0.02 to 0.12 (sparseness percentile 98.0 to 88.0, respectively), the standard deviation of the number of exclusive invariant cells also decreases. That is, the cells that represent each of the six objects become more evenly distributed between the objects. The pooled standard deviation values presented in Table 2.6 are plotted in Figure 2.9.

	RS = 1	RS = 2	RS = 3	RS = 4	RS = 5	Pooled SD
Percentile F = 98.0	53.5	43.6	45.0	51.8	42.3	47.46
Percentile E = 96.0	39.4	37.1	41.5	37.7	36.3	38.44
Percentile D = 94.0	34.2	32.8	29.3	33.1	32.2	32.36
Percentile C = 92.0	24.0	27.5	26.3	27.0	27.7	26.53
Percentile B = 90.0	18.5	22.5	17.3	24.6	24.2	21.63
Percentile A = 88.0	11.2	19.6	14.2	13.0	14.9	14.85

Table 2.6: Table of standard deviations. This table presents the standard deviations of the number of exclusive invariant cells that respond to each object when $N = 6$ and $M = 3$ when using five different random seeds (RS) and seven different sparseness percentile values. Specifically, each standard deviation is calculated using the mean number of cells that have learned to respond to each of the objects presented during testing. To provide a different example, when $N = 3$ and $M = 2$, if 80 output neurons learned to respond to the Cube, 5 output neurons learn to respond to the Pyramid, and 8 output neurons learned to respond to the Cone, then the mean is 31 and the standard deviation would be 42.5.

Figure 2.9 clearly shows a steady decrease in the pooled standard deviation value across the different random seeds when the sparseness percentile value is lowered and more cells are able to fire accordingly. At the lower sparseness percentiles, the individual objects receive a more equally distributed number of cells that learn to respond to each object.

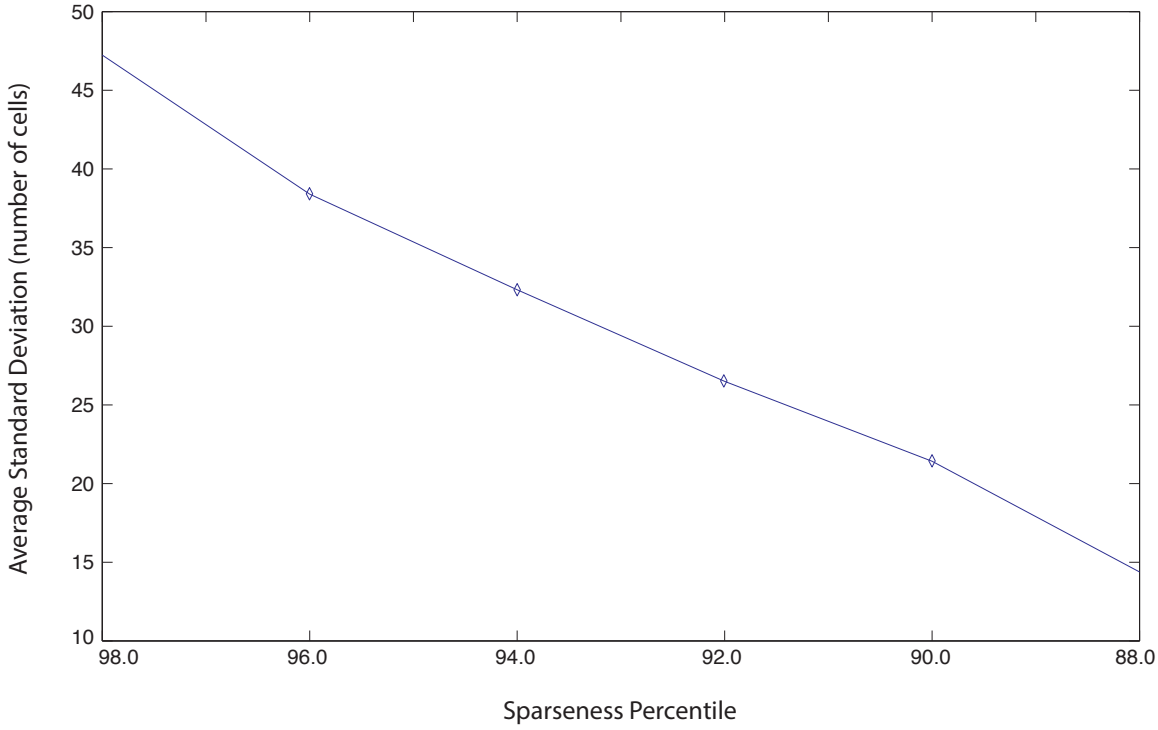

Figure 2.9: Change in pooled standard deviation when sparseness percentile is lowered. This figure shows that as the sparseness percentile is decreased (sparseness is increased) from 98.0 to 88.0, the pooled standard deviation of the number of cells also falls.

2.3.3 Information Analysis

For simulation sets: $N = 9$, $M = 8$; $N = 9$, $M = 3$; and $N = 9$, $M = 1$ information analysis was conducted to confirm whether the network had developed cells that responded invariantly to a single preferred object. After training, the network was tested with each object presented during training and the information analysis results are presented below accordingly.

Single cell information

Single cell information measures for the 4th layer neurons ranked in order of their invariance to the objects are shown in Figure 2.10. For this test case, the maximum amount of information that may be conveyed may be calculated as $log_2 N = 3.17$ bits where $N = 9$. The dashed line represents the results obtained after testing the network with the individual objects, having previously trained the network when $N = 9$, $M = 8$. The unbroken line represents the results obtained when $N = 9$, $M = 3$. The dash-dotted line represents the results obtained when $N = 9$, $M = 1$. Finally, the dotted line represents the results obtained after testing the network without any training. In all cases each of the nine objects is presented individually rotating through 360 °.

Figure 2.10 reveals that after training the network with $N = 9$ and $M = 1$, there were no cells that reached the maximum level of information and therefore confirmed that there are no cells that have learned to respond invariantly to an individual object. Instead, under this condition cells reached 2.0 bits of information. However, after training the network under conditions $N = 9$ and $M = 3$, approximately 400 of the 4th layer neurons reached the maximal level of single cell information of 3.17 bits for the trained objects. Similarly, when $N = 9$ and $M = 8$ approximately 500 cells convey maximal information. These neurons have learned to respond to all of the views of their preferred

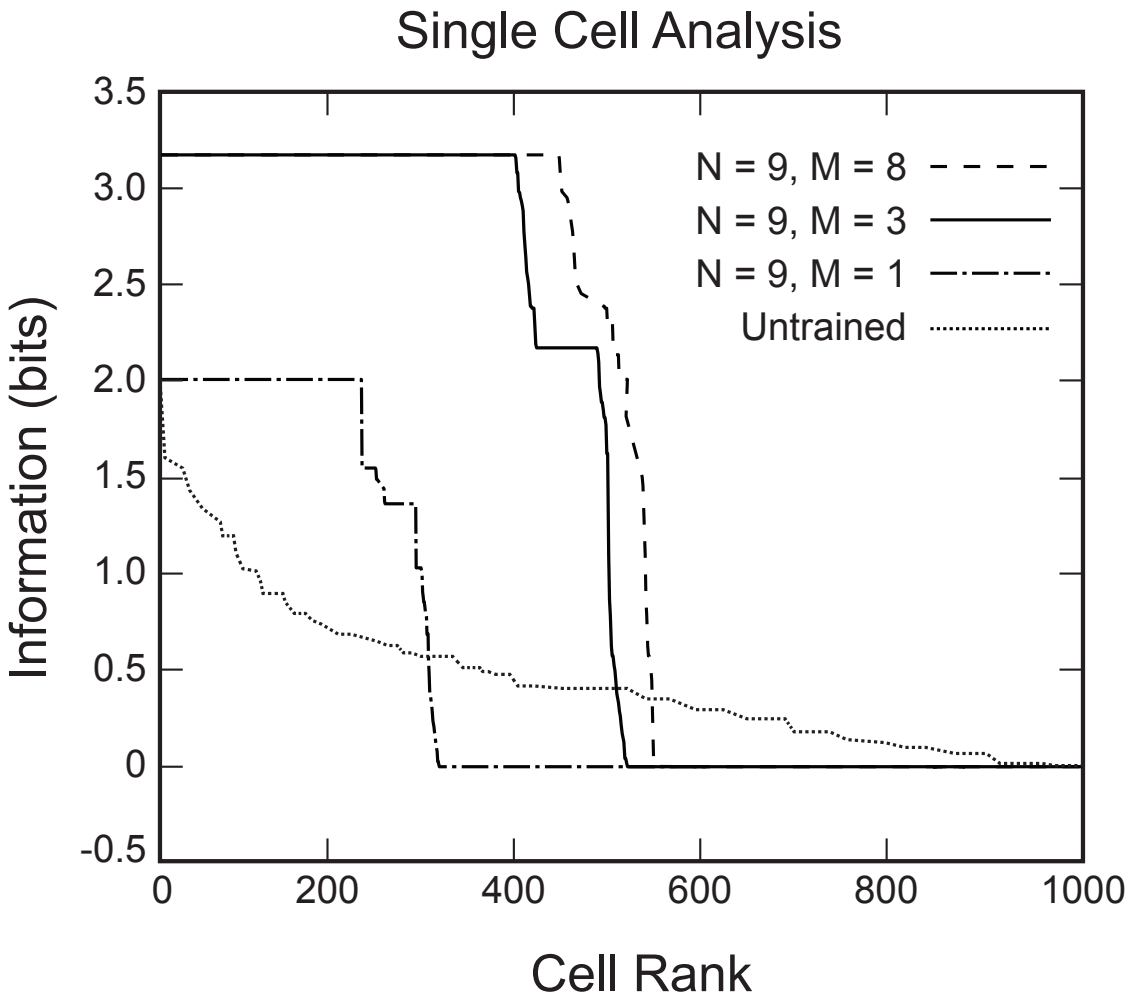

Figure 2.10: Single cell information analysis. This figure compares the amount of information available in bits for when $N = 9$, as M varies. For comparison, results obtained from an untrained network are presented. When $M = 8$ approximately 500 cells convey maximal information of 3.17 bits. Importantly, when $M = 3$, there are 400 cells that convey maximal information. However, when $M = 1$ no cells obtain maximal information and instead they reach approximately 2.0 bits of information. For comparison, the results obtained with an untrained network show that no cells reached maximal information.

object. The random untrained network provides a comparison.

Single cell information analysis alone is not sufficient to establish whether all objects presented during training are represented invariantly by neurons in the output layer of the model. It is possible that these neurons learned to respond to only one of the objects presented during training because this would still produce maximal information. To ensure that there are cells that learn to respond to each and every object presented during training, multiple cell information analysis was performed.

Multiple cell information

Figure 2.11 shows the multiple cell information analysis obtained when VisNet was tested with the nine individual objects rotating through 360° in 1° steps. As with the single cell information analysis, the dashed line represents the results obtained after testing the network with the individual objects, having previously trained the network when $N = 9$, $M = 8$. The unbroken line represents the results obtained when $N = 9$, $M = 3$. The dash-dotted line represents the results obtained when $N = 9$, $M = 1$. Finally, the dotted line represents the results obtained after testing the network without any training.

The results show that in the case of $M = 8$ and $M = 3$, the network has developed neurons that produce maximal multiple cell information. This confirms that each of the nine objects presented during training is faithfully represented by neurons in the output layer of the model. These neurons have learned to respond exclusively and invariantly to their preferred object. In the case of $M = 1$, the single cell information analysis results are confirmed. Here, the multiple cell information analysis is not maximal since the network has failed to produce neurons that have learned to respond exclusively and invariantly to each of the nine objects presented during training. The untrained network results are presented for comparison. Together, the single cell information, multiple cell information and cell

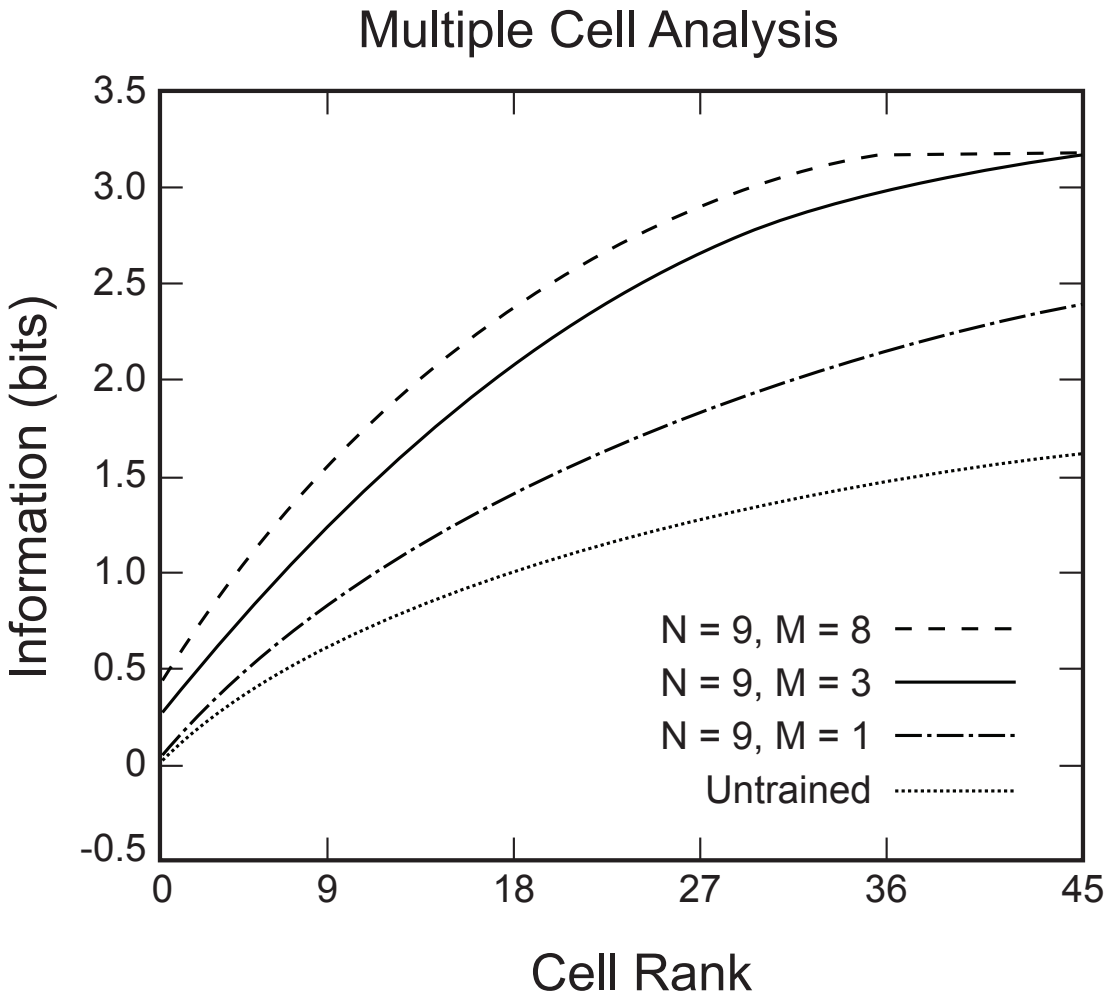

Figure 2.11: Multiple cell information analysis. This figure shows the multiple cell information analysis results after testing with all nine objects rotating through 360° in 1° steps. Results are presented for when $N = 9$, $M = 8$ (dashed line), $N = 9$, $M = 3$ (unbroken line), $N = 9$, $M = 1$ (dash-dotted line) and untrained (dotted line). It can be seen that in the case of when $M = 8$ and $M = 3$, maximal information is obtained (3.17 bits) where each of the nine objects is uniquely and invariantly represented by cells in the output layer of the model. When $N = 9$, $M = 1$, the network fails to produce cells that reach maximal information, confirming the results of the single cell information analysis, whereby neurons did not learn to respond exclusively to a preferred object. The untrained network is given as a comparison.

response plots show that the network has developed output neurons that respond exclusively to their preferred object across all views for when $M \geq 3$.

2.4 Discussion

Previous research had shown that it was possible for a multilayer feedforward model of the ventral visual pathway, VisNet, to build invariant representations of N rotating objects when each object is paired with all possible other objects within the training set, that is, ${}_{N}C_{N-1}$ total pairings (Stringer et al., 2007; Stringer and Rolls, 2008).

These results are crucial for highlighting some of the robust mechanisms with which a neural network such as VisNet may learn to form independent invariant representations of individuals objects when training within a multiple object environment.

However, the original authors (Stringer et al., 2007; Stringer and Rolls, 2008) highlighted a number of important limitations to their results. The most striking of which is the unrealistic amount of training that would be necessary in order to build separate representations of individual objects. More specifically, in a cluttered scene where many objects comprise the environment, if all objects must be paired with all other objects, then the addition of a new object to the scene will require an additional $N - 1$ pairings.

VisNet is able to build independent separate representations of each object within the training set so long as the features that comprise an object are seen more frequently together than they are with the features of any of the other objects. The purpose of this chapter has been to investigate how much statistical decoupling is necessary for the development of individual object representations when more than one object is always present during training. In particular, it would be highly beneficial for such

representations to form with much less training than that proposed by Stringer and Rolls (2008).

The results presented as part of this chapter reveal that, not only are fewer pairings required than originally used by Stringer and Rolls (2008), the number of M pairings per object required does not increase as N increases. It has been shown that the number of pairings necessary for each value of N is constant, at approximately $M = 3$. In terms of validating the integrity of the original findings, the results presented here are crucial; the robust nature of the statistical decoupling presented as part of this chapter would suggest that any neural network that operates in a similar manner to a competitive network would also achieve similar results. Accordingly, it is possible that such a mechanism exists in the primate visual system, whereby the statistics in the environment play a significant role in helping primates develop separate representations of independent objects.

Although the results presented within this chapter help explore the major limitations of the original study, there are limitations with the current research. For example, this research does not investigate an environment with more than nine objects. Nine locations were used due to the convenience of splitting the retina into a 3x3 grid. A larger grid would have required the objects that are presented within each part of the grid to become smaller since they must remain orthogonal. Orthogonal objects are required because the VisNet retina is relatively small, but also to maintain the necessary level of control required in order to systematically study statistical decoupling between stimuli and the effect it has on learning. As the resolution of the objects decreases, so does the amount of object specific visual information that is available to the network from which it learns. Increasing the size of the retina would help alleviate this problem and allow for a larger grid, and therefore more objects. However, this highlights another consideration.

The network was able to develop separate representations of each object at approximately $M \geq 3$.

This value is specific to the VisNet model, although it is believed that this general finding will persist in any similar competitive network. The value may also be specific to other parameters. For example, the size of the retina and the resolution of the filtered images; as the resolution of the filtered images increases in accordance with the increase in the size of the retina, so does the amount of visual information which the network is able to use for learning. The increase in visual information available could allow the network to benefit during learning. Furthermore, with a large enough retina it would be possible for objects to be presented in overlapping locations because the difference between the filtered outputs would be significantly different from one another.

It should be noted that the network was only trained using two objects at a time. This is a significant step forward from previous VisNet research, where only one object was presented at a time during learning, or when objects were presented on noisy backgrounds (Stringer and Rolls, 2000). However, a real-world environment contains more than just two objects. It would be worthwhile investigating the performance of VisNet when learning with three or more objects present simultaneously. New questions now arise, such as how much statistical decoupling is required in this scenario and how does the value of M vary with respect to the number of objects presented simultaneously during learning?

As previously outlined, invariant output representations form due to CT learning. However, given that different views of the same object tend to occur close together in time, consideration should be given to how a trace learning rule could be incorporated into the training paradigm. It is not obvious how this may be implemented, since in order to isolate the trace learning effect from the CT learning effect the training paradigm would need revision so that different transforms of the same objects did not overlap. Overlapping transforms can give rise to the CT effect which would confound the results. Accordingly, a translation training paradigm would probably be the best option, but further study must

be performed in order to establish where it will be possible to construct a training paradigm that will satisfy all the necessary constraints.

In this doctoral thesis chapter, the amount of training required for the development of individual object representations was investigated. A view invariance study was performed whereby pairs of objects were presented to the network during training similar to previous research by Stringer and Rolls (2008). It has been shown that the significant amount of training assumed by Stringer and Rolls is not necessary, and instead reduced training is possible. Furthermore, the amount of training does not appear to depend on the number of N objects that together make up the parent training pool of objects. An important aspect of this research is the consideration given to the particular training paradigm and visual environment, in particular, the statistics of how visual features and objects occur together.

Chapter 3

Learning View Invariant Recognition with Partially Occluded Objects

The previous chapter investigated how much training is necessary for VisNet to learn individual, transform invariant object representations when presented with pairs of objects during training. This is a very important consideration because previous research has suggested that a significant amount of training may be required (Stringer and Rolls, 2008). A high level of training is ecologically implausible, and therefore unlikely to be representative of how the primate visual system learns. However, the results in the previous chapter clearly indicate that VisNet is capable of learning individual transform invariant object representation with much less training than originally suggested.

Rarely does the primate visual system have the opportunity to view all possible combinations of a set of objects in a convenient manner. Another ecological concern with previous VisNet research, is that objects have been presented in isolation and never with other objects partially occluding them during training. In the real world, the primate visual system is able to learn about new stimuli even

when they are partially occluded. This is another important challenge that has not yet been addressed by previous VisNet research. This chapter reports the results obtained when VisNet is trained with object pairs, where one object is always partially occluded by the other object during training.

3.1 Introduction

It is important to understand how invariant representations of individual objects are built in the primate visual system even when multiple objects are present in natural scenes during learning. Neurophysiological research has provided substantial evidence showing that over successive stages, the visual system develops neurons that respond with view, size and position (translation) invariance to objects or faces (Desimone, 1991; Perrett and Oram, 1993; Rolls et al., 1992; Rolls and Deco, 2002; Tanaka et al., 1991). For example, it has been shown that the inferior temporal visual cortex has neurons that respond to faces and objects with translation (Ito et al., 1995; Kobatake and Tanaka, 1994; Op De Beeck and Vogels, 2000; Tovee et al., 1994), and view (Booth and Rolls, 1998; Hasselmo et al., 1989b) invariance.

The "biased competition hypothesis" of attention suggested that feedback connections are necessary to build separate representations of individual objects in a complex scene by providing the mechanism for attentional selection (Rolls and Deco, 2002). However, it has been shown that this separation can be achieved without the need for an attentional mechanism by using purely feedforward connectivity in a hierarchical neural network model of the ventral visual pathway, VisNet (Stringer et al., 2007; Stringer and Rolls, 2008). The statistical properties of the input stimuli play a crucial role, whereby the features within individual objects occur more frequently together than the features between different objects. Therefore, although the role of feedback connections is an important area

for future research, they will not be implemented in the present study.

Stringer and Rolls (2000) showed that a hierarchical neural network model of the ventral visual pathway, VisNet, could recognise objects presented against natural cluttered scenes, providing the model had been previously trained with each object presented individually transforming against a blank background. However, the network failed to learn to recognise individual objects if the objects were presented against a natural cluttered background during training.

Recent studies by Stringer and Rolls (2008) and Stringer et al. (2007) have shown how VisNet may cope with complex scenes during training, and learn invariant representations of individual objects even when no single object is seen in isolation. These modelling studies used the statistics of the natural environment where features within an object occur together more frequently than features between different objects. Specifically, VisNet could learn invariant representations of individual objects if different combinations of transforming objects were seen at different times.

However, a further major challenge is to explain how invariant representations can be learned when the objects are partially occluded by one another during learning. Stringer et al. (2007) proposed that Continuous Transformation (CT) learning (Stringer et al., 2006) combined with the statistical independence of objects presented in different combinations might allow the network to solve this problem. Specifically, consider presenting a number of objects to the network in different subset combinations, but where one of the objects is always partially occluded by some other displayed object. The hypothesis is that the network will simultaneously form separate representations of all of the different objects, where an invariant representation of the partially occluded object is formed by linking together the different partial views through CT learning. However, Stringer et al. (2007) provided no simulation evidence that this could work. In this chapter this process is demonstrated for

the first time, operating with simulated 3-dimensional rotating objects. It is important to investigate this issue because objects in the natural environment will often overlap. The task is more difficult than simply forming separate representations of different objects because, in order for the network to build a complete invariant representation of the partially occluded object, the network has to link together the different partial views of the object as well as separate these partial views from the other objects present.

The simulations described below show how VisNet can form separate view invariant representations of individual objects seen rotating together, where one of the rotating objects is always partially occluded by the other objects present during training. The network develops cells in the output layer which have learned to respond invariantly to particular objects over most or all views, with each cell responding to only one object. All objects, including the partially occluded object, are individually represented in this way by a unique subset of output cells. This learning process relies on the statistical independence of the objects that are shown in different combinations, as well as an invariance learning mechanism known as Continuous Transformation (CT) learning that is described next.

3.2 Method

3.2.1 Learned Object Selectivity

Continuous Transformation (CT) learning develops transform invariant representations that are object-specific. That is, as long as each object is not always presented together with another particular object transforming in lock-step during training, individual neurons typically learn to respond to one object only (Stringer et al., 2006). Consider an object rotating at a particular retinal location during training. Successive views of the object are represented by the outputs of the oriented input filters representing

V1 simple cells as described in Equation 1.1. Continuous Transformation learning utilises Hebbian competitive learning. At each presentation of a view of an object during training, activity is propagated up through successive neuronal layers in the network. Within each layer, a small subset of neurons wins the competition.

The feedforward synaptic connections from the input filters to the first layer of neurons are modified according to a Hebbian learning rule (Equation 1.7). This learning rule strengthens only the synaptic connections from those V1 filters that are activated by the particular visual form of the object view currently presented. The weight vector of each of the first layer neurons gradually shifts during learning to point in the same direction as the V1 input pattern(s) to which it is learning to respond. Since each first layer neuron computes its activation (Equation 1.2) according to the dot product of its weight vector and the current input pattern from the V1 input layer, after training each neuron will respond in proportion to the similarity between the current input pattern and the input pattern(s) the neuron learned to respond to during training. That is, each neuron will respond maximally to the input pattern that it has learned to respond to during training, and generalise to other input patterns depending on their similarity (Hertz et al., 1991).

The competitive Hebbian learning rule operates in a similar manner for the feedforward connections between all of the later layers of the network. This ensures that a subset of neurons in layer 1 and all successive layers learn to respond to the pattern of visual features present in the current view of the trained object. This means that the subset of output neurons in the higher layers of VisNet learn to respond to the visual form of the current view of the trained object and not its retinal location.

If output neurons simply learned to respond to retinal location, then the feedforward connections would need to be strengthened from *all* of the V1 filter inputs in a particular location regardless of the

visual form of the objects. But this cannot occur because the Hebbian learning rule ensures that only the synaptic connections coming from those V1 input filters actually activated by the particular visual form of the object can be strengthened.

As previously described, the Hebbian learning rule is able to associate different views of the object onto the same active output neurons as long as the different object images presented during training cover a space of smoothly changing views. Again, only those V1 input filters that were activated by the different object views can become associated with the active subset of neurons in the higher layers. So, even after many stimulus views have been presented, the neurons in the later layers of the network will not normally learn to respond to all of the V1 filters in a particular retinal location. Thus, after training, the output neurons become object-specific. The output neurons will respond maximally to different views of the particular object that has been learned.

If another different untrained object is presented in the same retinal location as the first trained object, then there will be a rather different pattern of V1 input filters activated. However, as discussed above, the output of each of the neurons in the network reflects the similarity between the input pattern in the previous layer to which it learned to respond during training and the currently tested input pattern. This means that the neurons in the higher layers that have been previously trained to respond to the first object will respond to the second untrained object in proportion to the degree of visual similarity between the two objects. Therefore, due to the properties of Hebbian competitive learning, the neurons through the higher layers of the network must learn to respond to the visual forms of objects rather than locations.

There is a potential conflict between the need to develop representations that are object-specific and the need to develop transform invariant representations within each object. In principle, if two

different objects have similar transforms, then the CT learning mechanism may encourage output neurons to learn to respond invariantly across both objects. This is a fundamental issue with CT learning, which requires ongoing investigation. In simulation studies, this author finds that increasing the size of the VisNet architecture (increasing the number of cells within each layer) improves the ability of the model to learn separate representations of, for example, similar faces.

It is also possible that combining CT learning with a trace learning rule (Foldiak, 1991) could improve the ability of the network to form separate invariant representations of different objects. However, this has not been implemented in the simulations reported here, which use only a standard Hebbian learning rule.

3.2.2 Multiple Objects

How the brain can build invariant representations of individual objects even when multiple objects are present in a scene during learning is a very important question in natural vision. How the visual system learns about individual objects rather than the combination of objects that make up the scene has only recently been investigated successfully in a biologically realistic model (Stringer et al., 2007; Stringer and Rolls, 2008).

In Chapter 2, the work by Stringer et al. (2007) was extended to show that VisNet can develop individual invariant object representations during learning with relatively little training compared to what was previously thought. The following research makes use of these findings, whereby the Visnet model can exploit the fact that the features that make up a given object occur together more frequently when presented during training compared to the features that make up different objects. Depending on how often a given object is presented during training with another object, the frequency of how often features between these two different objects occur together will vary. However, the features that

make up any individual object are always presented with one another and are therefore completely correlated.

It has been shown that a competitive network will operate usefully in this situation. The network will learn primarily to form representations that reflect the high probability of co-occurrence of features from one object, and do not reflect the features of other objects presented simultaneously during training, if the object being trained is seen far more frequently than when it is presented with any other object.

In order for a competitive network to build representations of individual objects, there must be a statistical decoupling between the features that comprise each of the objects presented during training. Providing that there are a sufficient number of objects present during training, each object will be presented with many other objects and therefore the features within a single object will appear together significantly more often than when they are coupled with features of any other object. It has been previously demonstrated that this allows a competitive network to form transform invariant representations of individual objects, rather than the combinations of objects seen during training, by a mechanism such as Continuous Transformation learning (Stringer and Rolls, 2008).

3.2.3 Learning to Recognise Partially Occluded Transforming Objects

The major new challenge addressed in this chapter is how VisNet can form separate view invariant representations of individual objects seen rotating together, where one of the rotating objects is always partially occluded by the other objects present during training. To create view invariant representations of the occluded object, the network will have to separate it from the occluding objects and link together different partial views to create a representation of the whole object. The potential solution described by Stringer and Rolls (2008) for separating out individual objects in a scene with multiple objects

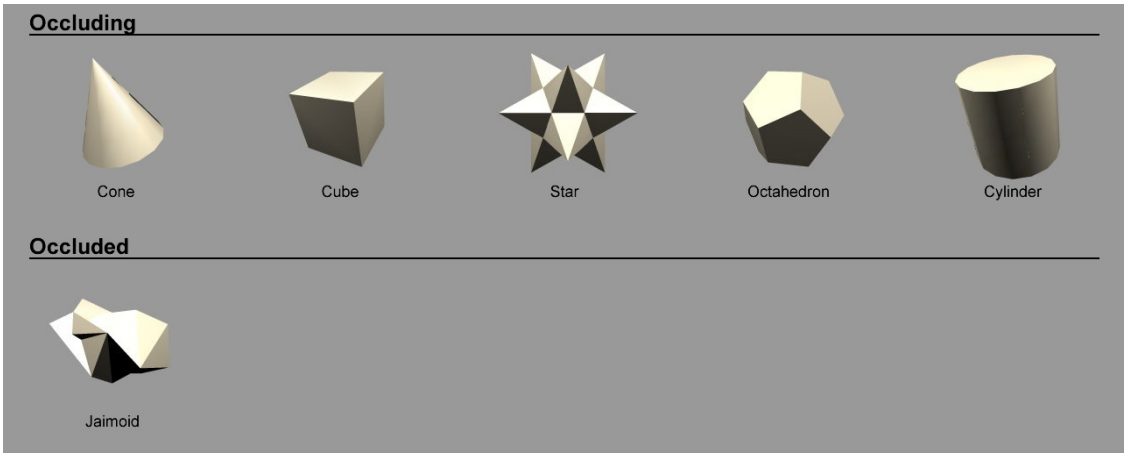

Figure 3.1: The six objects used to train the network. The objects are 3D objects each shown from 360 different views. The effect of the ambient lighting and single diffuse light source is illustrated. This allows different surfaces to be shown with different intensities. Objects are split into two groups; occluding objects are presented in the top row and the occluded object is presented in the bottom row.

present will be used to separate the occluded and the occluding objects. CT learning will be used to link together the different transforms of each individual object, including associating together the occluded and unoccluded views. By training VisNet with multiple objects that partially occlude one another, it is shown that this model of the ventral visual stream is able to learn reliably in increasingly realistic visual environments.

3.2.4 Stimuli

Figure 3.1 shows the objects used to train the network. There were $N = 6$ continuously rotating 3D objects on a grey background. Previous research (Stringer and Rolls, 2008) has shown that $N = 6$ objects is sufficient to allow VisNet to develop representations of individual objects when the network was trained on object-pairs. The objects were designed and created using the 3D modelling tool Swift 3D. Ambient lighting with a diffuse light source was added to allow different surfaces to be shown with different intensities. Each object rotated in depth around the vertical axis in 1° steps through

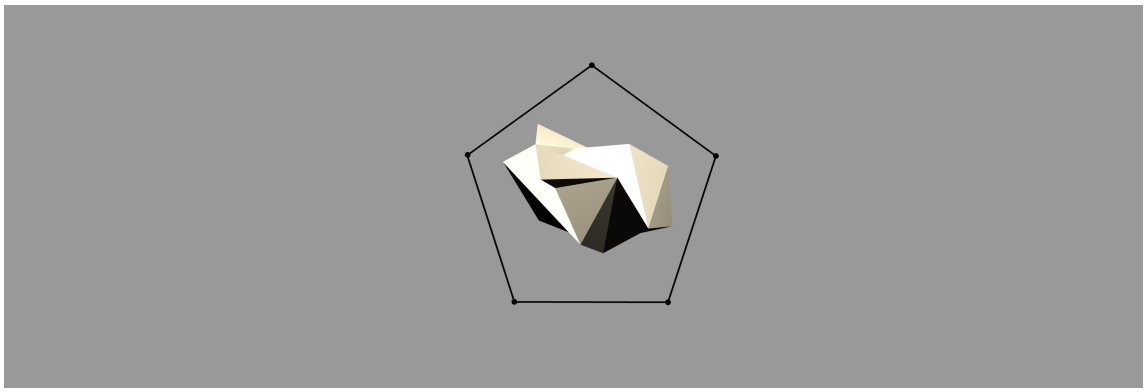

Figure 3.2: The pentagon formation specifying the location of the occluding and occluded objects. The occluded object, the Jaimoid, is always presented in the centre of the pentagon. The Jaimoid is an irregular multifaceted three dimensional object. Each of the occluding objects is placed at one of the five points of the pentagon, partially overlapping the Jaimoid at the centre. This ensures that the occluding objects are equidistant from the centre of the occluded object, helping to maintain a comparable level of partial occlusion for the Jaimoid.

360 °. This step size was chosen because past research (Stringer et al., 2006) has revealed that it was sufficiently small for CT learning to operate. The 360 views of each object were then exported as 2D JPG images and encapsulated as Adobe Shock Wave Files. The objects were then aligned and organised using Adobe Flash CS4.

The stimulus set was comprised of five occluding objects and one occluded object. During each training sequence, the occluded object was shown rotating with one of the occluding objects. In all cases, each object would rotate about its own vertical axis and therefore all axes were in parallel with one another. The spatial arrangement of the objects is shown in Figure 3.2. The occluding objects were presented in a pentagon formation. The occluded object, the Jaimoid (irregular multifaceted three dimensional object, Figure 3.3), was always presented in the centre of the pentagon.

Each of the occluding objects was placed at one of the five points of the pentagon, partially overlapping the Jaimoid at the centre. The occluding objects were equidistant from the centre of the occluded object, therefore occluding it to the same extent. The occluded object was always behind

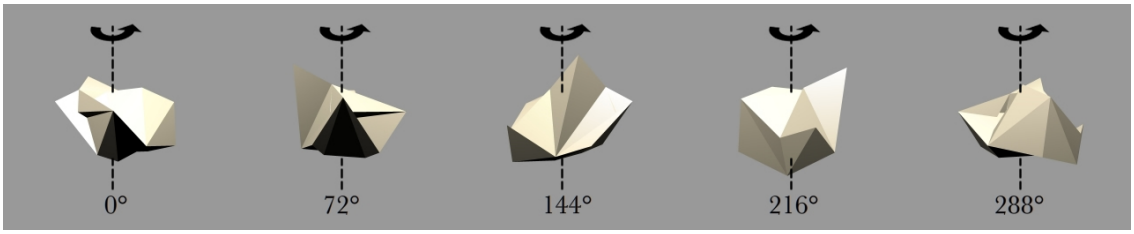

Figure 3.3: Five example frames selected from the 360 frame testing image sequence of the Jaimoid rotating in depth around the vertical axis through 360 ° in 1 ° steps. The selected frames shown are for 0 °, 72 °, 144 °, 216 ° and 288 °. All 5 of the occluding objects were also presented to the network in the same manner.

the occluding objects and in the middle of the pentagon formation. This spatial formation was chosen because it was necessary to ensure that different parts of the occluded object were covered by the five occluding objects. In all simulations the objects were rotating at the same speeds.

3.2.5 Model Architecture and Parameters

The experiments presented as part of this chapter were conducted using the VisNet model architecture as described in Section 1.3.3. Images were filtered in accordance with the Difference-of-Gaussian Equation 1.1 previously presented. The specific sparseness percentile and slope values are presented in Table 3.1. The lateral inhibition parameters are provided in Table 3.2.

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

Table 3.1: Sigmoid parameters

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

Table 3.2: Lateral inhibition parameters

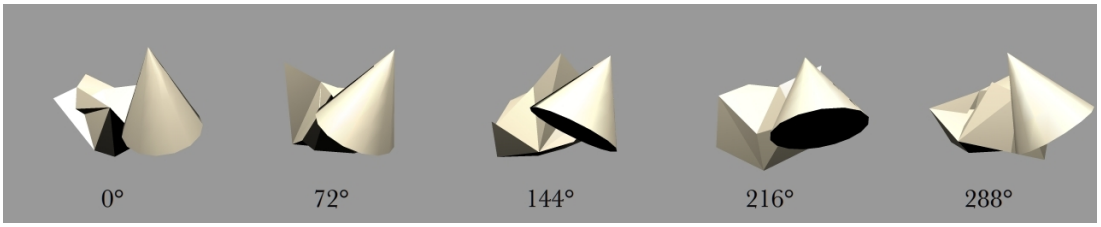

Figure 3.4: Five example frames selected from the 360 frame training image sequence of the Jaimoid and the Cone rotating through 360 ° in 1 ° steps. The selected frames shown are for 0 °, 72 °, 144 °, 216 ° and 288 °. The Jaimoid is also occluded by the four other occluding objects in separate image sequences. Therefore, in total, there are 5 image sequences used during training, each containing 360 frames.

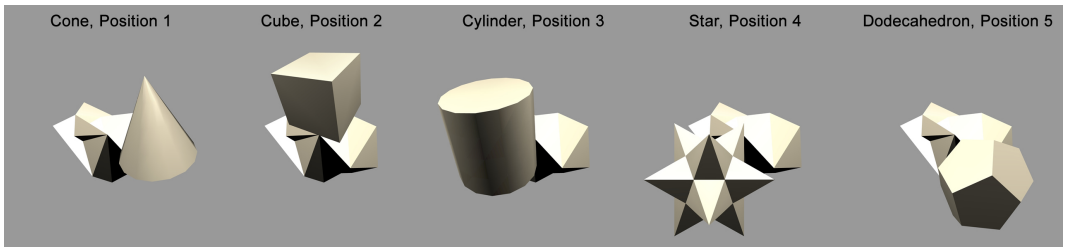

Figure 3.5: Five example frames of the Jaimoid occluded by all five occluding objects in their five corresponding positions. The occluding objects are arranged around the pentagon formation so that they are equidistant from the centre of the Jaimoid. The rotating objects used during training are presented in the same locations: Cone, position 1; Cube, position 2; Cylinder, position 3; Star, position 4; Dodecahedron, position 5.

3.2.6 Training Procedure

The occluded object, the Jaimoid, paired with each of the five surrounding occluding objects, is presented to VisNet with both objects rotating through 360 ° (Figure 3.4). Each full revolution through 360 ° of the pair is followed by the occluded object paired with a different occluding object in a different location around the pentagon formation. This process is repeated until the occluded object is paired with all five occluding objects. The rotating object pairs are presented as follows: Cone, position 1; Cube, position 2; Cylinder, position 3; Star, position 4; Dodecahedron, position 5 (Figure 3.5).

In addition, all possible pairings of the five occluding objects are then presented in a similar fashion rotating through 360 °. This helped VisNet to learn separate representations of the objects by using the statistics of the natural environment where the features within an object occur together

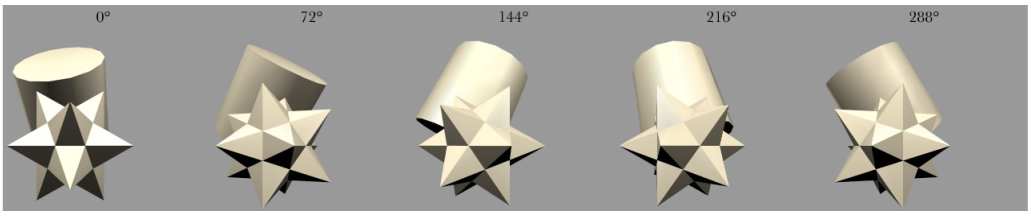

Figure 3.6: Five example frames of two occluding stimuli (Cylinder and Star) rotating together, demonstrating the typical overlap between the occluding objects.

more frequently than features between different objects (Stringer and Rolls, 2008). It should be noted that adjacent pairs of occluding objects would also sometimes overlap during training, leading to one occluding object being partially occluded by another occluding object (Figure 3.6).

At each image presentation, the activation of individual neurons within a layer is calculated, then their firing rates are calculated, and the feedforward synaptic weights between layers w_{ij} are updated according to Equation 1.7. This process is repeated for each layer in turn for all four layers of the VisNet model. One training epoch consists of the occluded object paired with all five occluding objects across all 360 transforms followed by all possible pairings of the occluding objects rotating through all 360 °

In this manner, the network is trained one layer at a time starting with layer 1 and finishing with layer 4. Fifty training epochs were used for layers 1-4. The learning rate for layers 1-4 were 0.109, 0.1, 0.1 and 0.1, respectively.

Due to high computational expense, the Jaimoid was the only object that was partially occluded by its neighbours in the simulations described below. However, the underlying theory described above predicts that similar effects would be found if the simulations were repeated with more objects partially occluded by each other during training.

3.2.7 Testing Procedure

During testing, the synaptic weights within the model are fixed and cannot be altered. Firstly, in order to test whether VisNet had built an invariant representation of the central partially occluded object, the occluded object was presented individually, rotating around the vertical axis through 360 ° (Figure 3.3). The surrounding five occluding objects were also presented in isolation in a similar fashion to verify that VisNet had build invariant representations of these objects too. The neuronal outputs of the network were then recorded during the testing presentations of each view of each object. Although it is the responses of neurons in the output layer of the VisNet model that are of most interest, the output responses of the intermediary layers were also investigated.

Secondly, VisNet was also tested with six novel objects rotating through 360 ° in 1 ° steps. This test demonstrates whether Visnet has learned to respond to the specific objects presented during training or whether VisNet has learned to respond selectively to only the location where these objects were presented.

Finally, VisNet was also tested with the different partial views of the occluded object as presented in Figure 3.7 and exemplified in Figure 3.8. As the different object pairs rotate together through 360 °, parts of the partially occluded object are not visible. By testing VisNet with the partial views that were not visible during training it is possible to establish whether VisNet is able to bind together the partially occluded views of the occluded object into one holistic invariant representation. This test is important because it shows that VisNet does not need to rely on a key component of the training stimuli in order to recognise it.

In addition to the cell response plots produced from the tests outlined above, information analysis measures were also performed. The procedures for the information analysis are presented in Section

1.3.5.

3.3 VisNet Simulation Results

3.3.1 Analysis of Individually Rotating Objects

After the network had been trained on pairings of the occluded and five occluding objects, the network was tested to establish whether it had built transform invariant representations of the objects through a CT learning effect. By presenting the rotating objects individually to the network it was possible to record the cell response properties of the neurons in the 4th layer of VisNet for each of the objects. A large number of individual experiments were performed across different parameters and random seeds to ensure the consistency and validity of the results. However, the results presented below are taken from the same individual experiment.

Populations of cells that responded invariantly to the individual objects were found. These cells responded to only one object and to no views of any of the other objects. Figures 3.9 (a) & 3.9 (b) show the cell response plots for cell (4, 17), selected at random, as each object is rotated through 360 $^{\circ}$ in 1 $^{\circ}$ steps. Figure 3.9 (a) shows the responses of the cell before training and Figure 3.9 (b) shows the cell responses after training. The six response plots of cell (4, 17) before training show that the cell responds at random to the six objects. After training, the cell has learned to respond to the central occluded object, the Jaimoid, invariantly and does not respond to any view of any of the other objects.

Figures 3.10 (a) and (b) show the cell response plots for cell (19, 1) before and after training, respectively. Before training, the cell responds to the objects randomly. After training this cell has learned to respond invariantly to all 360 views of the Dodecahedron, which was one of the occluding objects, and to no views of any other objects. Furthermore, other output cells learned to respond in

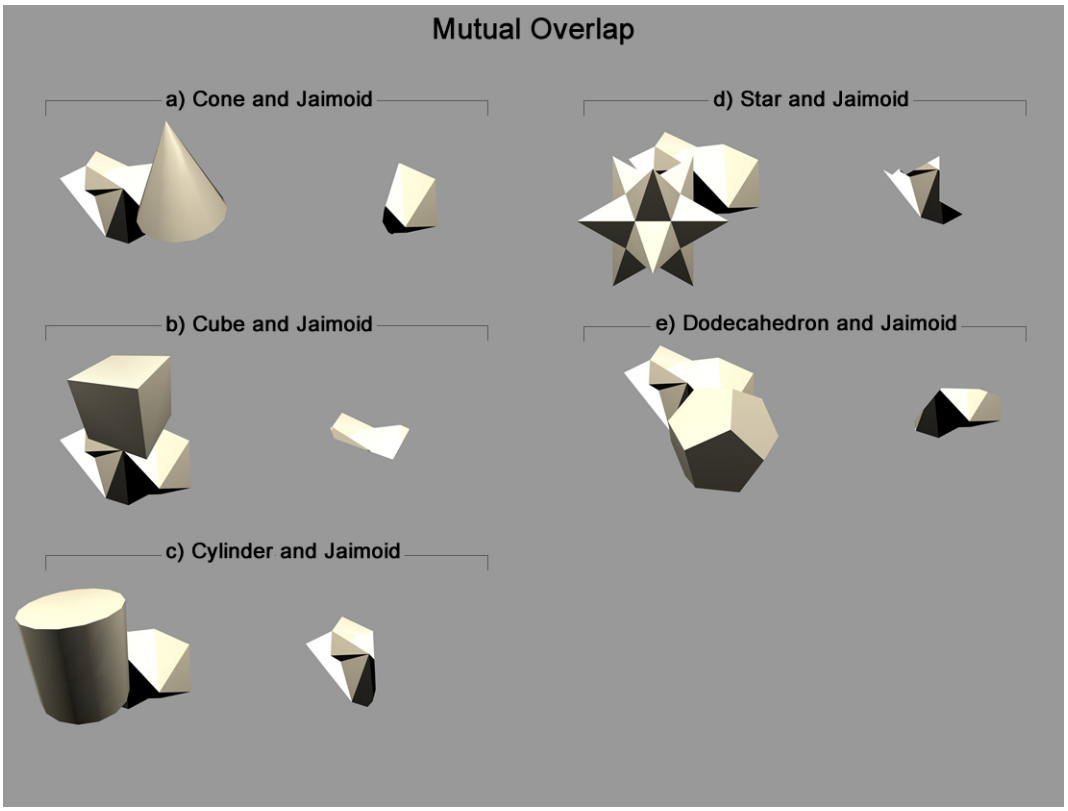

Figure 3.7: Mutual overlap: Areas of mutual overlap between the occluding and occluded objects during one example frame of rotation. As the two objects rotate together in lock-step, the area of mutual overlap creates a partial view of the occluded object: a) Shows the Cone and Jaimoid; b) Cube and Jaimoid; c) Cylinder and Jaimoid; d) Star and Jaimoid; e) Dodecahedron and Jaimoid. Each pair is presented alongside the partial view it creates. VisNet must learn to associate together all the different partial views of the occluded object to build an exclusively invariant representation.

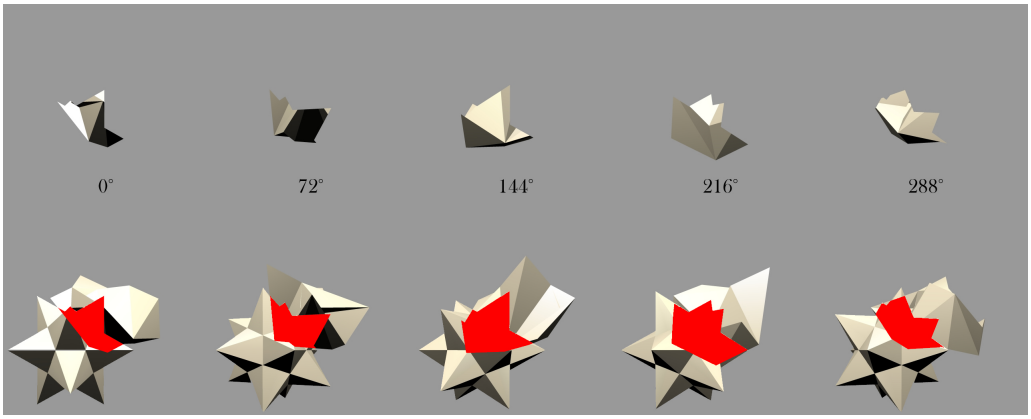

Figure 3.8: Star and Jaimoid mutual overlap: Five example frames of the Star and Jaimoid as they rotate together in lock-step. As in Figure 3.7, the area of mutual overlap creates a partial view of the Jaimoid. This is shown in this specific example where the Star and Jaimoid are presented over five equally spaced viewing angles.

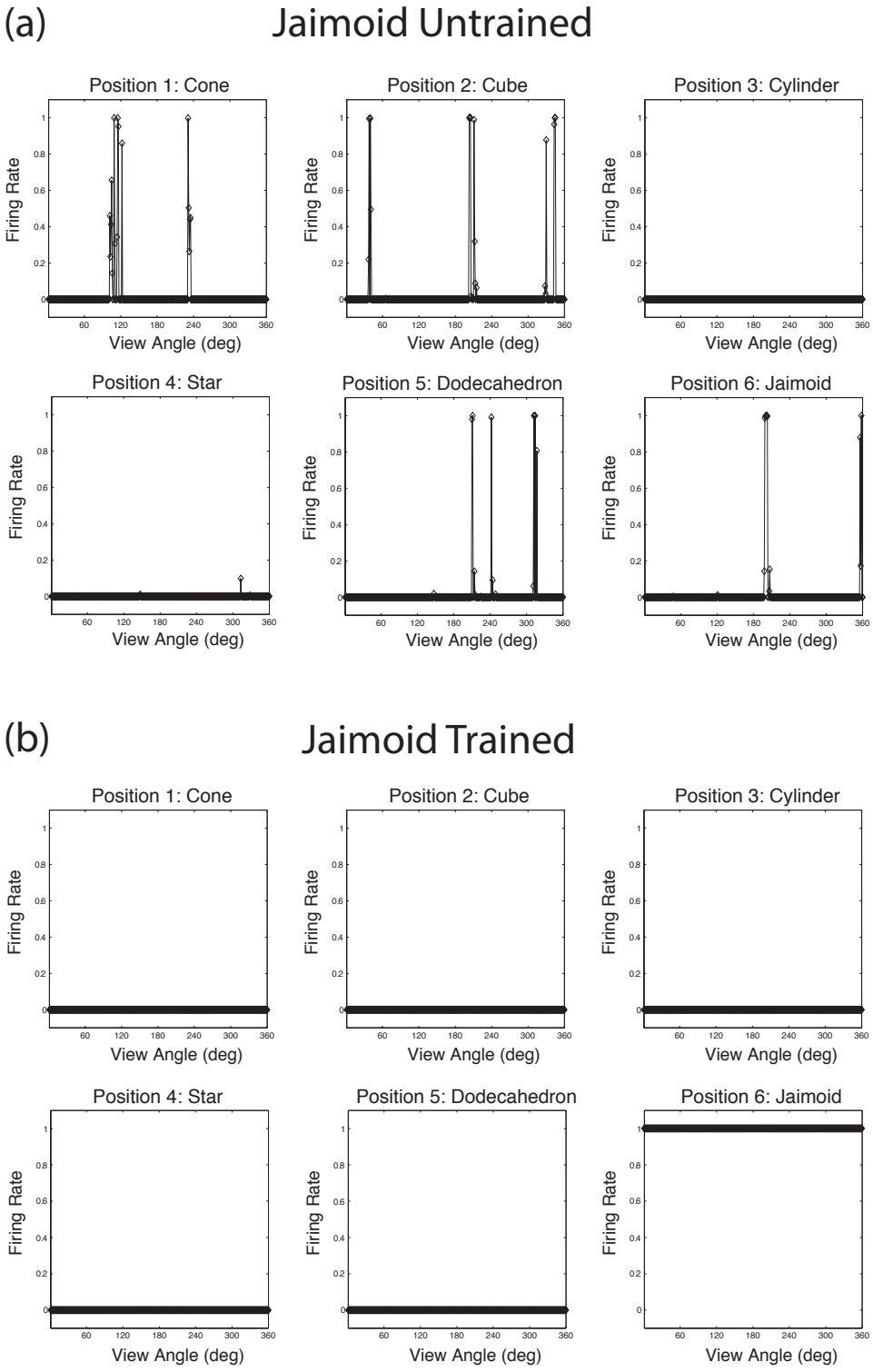

Figure 3.9: The firing rate responses of cell (4, 17) in the 4th (output) layer of VisNet to the central occluded object (Jaimoid) and the five surrounding occluding objects as they rotated through 360 ° in 1 ° steps before and after training. Before training, it can be seen that the cell responds randomly to different views of different objects. After training, it can be seen that the cell's response pattern has changed. This cell responds to all the views of Jaimoid, and to none of the views of the other objects.

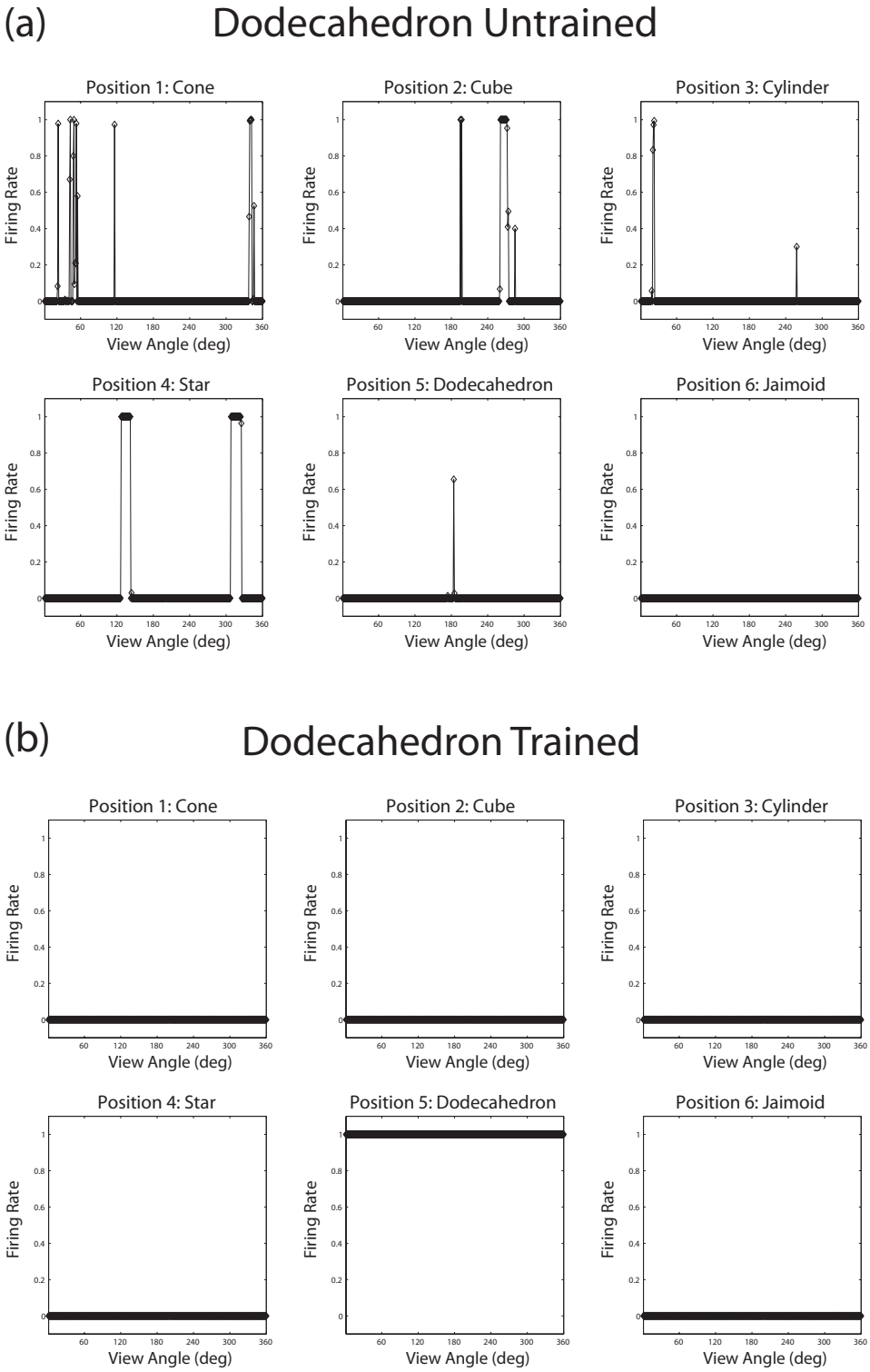

Figure 3.10: The firing rate responses of cell (19, 1) in the 4th (output) layer of VisNet to the central occluded object (Jaimoid) and the five surrounding occluding objects as they rotated through 360 ° in 1 ° steps before and after training. Before training it can be seen that the cell responds randomly to different views of different objects. After training it can be seen that the cell's response pattern has changed. This cell has become an exclusive invariant cell for the Dodecahedron. It responds invariantly to all 360 views of the Dodecahedron and to no views of any other object.

a selective and invariant manner to each of the other occluding objects (not shown). Thus, all of the objects were represented individually. When different sparseness values throughout the layers were investigated, the results were found to be robust. As the sparseness was gradually increased, a more distributed representation began to form with fewer exclusive cells whereby each cell began to respond to more than one object. In this situation, object identity is still encoded but over a population of cells.

The output layer not only produced cells that respond invariantly to a preferred object. Some cells in the output layer learned to respond to a contiguous range of views, while others learned to respond to seemingly more disjointed clusters of views. However, in both cases these types of cells still learned to respond exclusively to a preferred object. More specifically, these cells did not respond to more than one object.

Figure 3.11 shows the firing rate response plots of six different cells from the model's output layer after training. Each of the six cells responds selectively to only one preferred object and does not respond to any views of the other objects. Although all objects are represented by cells that respond invariantly, this figure presents three different cell response types. Cell (19, 1) responds exclusively and invariantly to all views of the Dodecahedron and cell (26, 1) responds in a similar manner for the Star. A different cell response type is exemplified by cell (10, 1) and cell (14, 14), which respond to the Cone and Cylinder, respectively. Specifically, these cells respond to contiguous views of their preferred object, but are not completely invariant. Finally, cells (26, 27) and cell (11, 8) respond to the Cube and Jaimoid, respectively, and are not fully invariant. There are a number of sporadic views to which the cell has not yet learned to respond, exemplifying the noisy responses that output cells can generate. In all cases, prior to training, each of the six cells did not respond exclusively to any of the six objects. That is, each cell responded randomly to a number of transforms from the six different

objects.

Table 3.3 shows in more detail how each of the six cells plotted in Figure 3.11 respond to each of the six objects before training. Each row of Table 3.3 shows the number of transforms of each object that a particular cell responds to before training. It can be seen that cells respond to the objects indiscriminately and show no prior bias towards the object to which they learn to respond preferentially after training.

	Cone	Cube	Cylinder	Star	Dodec	Jaimoid
Cell 10, 1	33	11	50	15	61	4
Cell 26, 27	11	17	17	6	10	30
Cell 14, 14	3	17	16	3	24	14
Cell 26, 1	5	24	7	36	0	1
Cell 19, 1	38	2	10	22	53	126
Cell 11, 8	63	60	56	59	37	10

Table 3.3: Untrained cell responses to all six objects. Cells exhibit no object selectivity and respond to sporadic transforms of the six objects.

Previous studies with VisNet have focussed on the development of perfectly invariant cells that respond to every transform of a preferred stimulus. In the experiments presented within this chapter, some neurons in the output layer of the model learned to respond invariantly to large parts but not all of their 360° view spaces. For example, Figure 3.12 shows that a complete 360° representation of the occluded object was encoded by a small population of cells. This figure shows the firing rate response profile of three 4th layer neurons which have learned to respond to the Jaimoid. Plot (a) shows the response profile of cell (5, 12) which has learned to respond to the Jaimoid over most of the views. Plot (b) shows the response profile of cell (32, 2), which has learned to respond to the Jaimoid over an overlapping region of the view space including some points to which the previous cell did not respond. Similarly, Plot (c) shows the profile of cell (6, 1), which has learned to respond to the Jaimoid over the remainder of the view space to which the two previous cells had not responded. Since all three

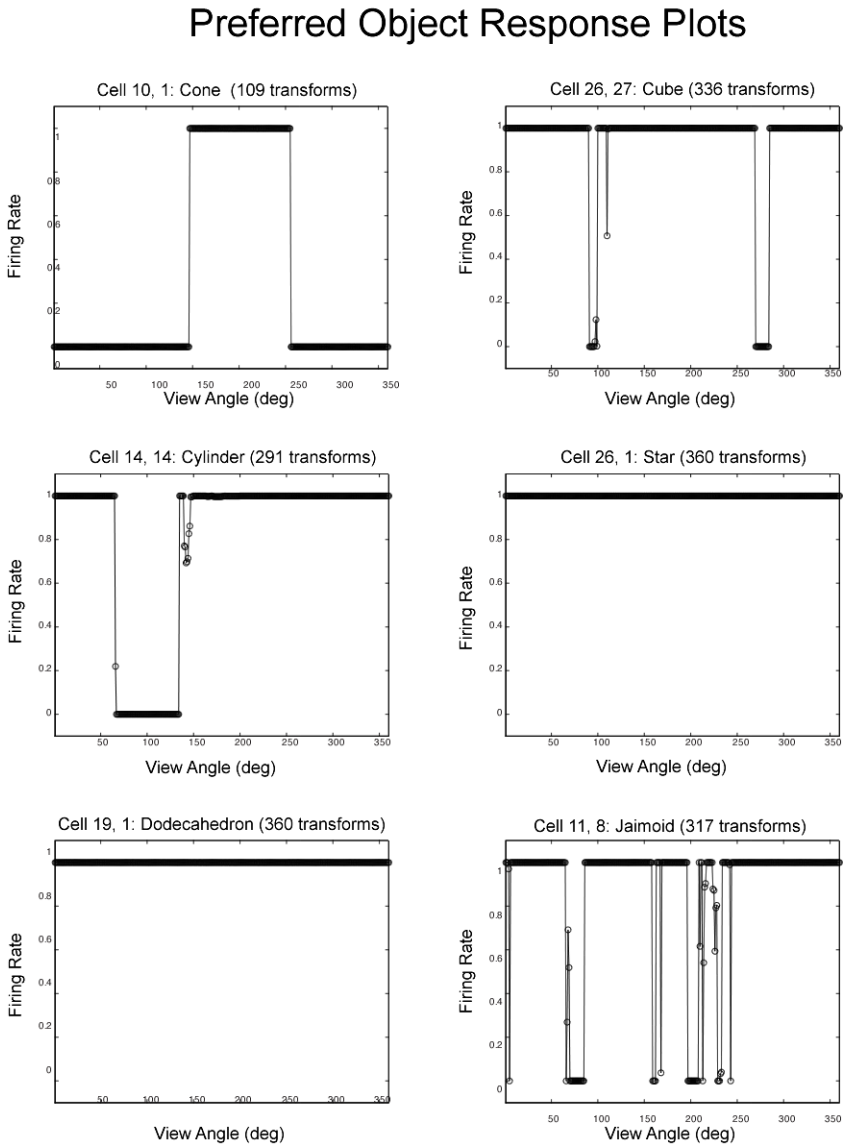

Figure 3.11: The firing rate response plots of six different cells from the model's output layer after training. Each of the six plots shows the responses of one of the six cells to its preferred object. Broadly speaking, there are three different cell response types: fully invariant response; contiguous responses; noisy responses. In most cases, each cell responds over large portions of the view space to its preferred object: Cone, Cube, Cylinder, Star, Dodecahedron and Jaimoid. Although not presented, each cell does not respond to any views of any of the other objects.

cells presented responded to the Jaimoid exclusively, the firing of any of these cells would allow a subsequent brain area to recognise that the Jaimoid is being viewed. In addition, these cells would convey information about the current orientation of the object.

3.3.2 Analysis of Cell Firing Properties in Earlier Layers

The analyses described above were applied to the output (4th) layer of the network. Cells in the output layer receive information through the feedforward synaptic connections from across the entire input retina. However, cells in the earlier layers receive more localised input from the retina due to the topographical feedforward connectivity present within the model. Therefore, additional analyses of the cell response properties in the earlier layers after training were performed.

Response plots for cells that have learned to respond to the Jaimoid are presented for each of the four layers in Figure 3.13. It was found that the responses of cells in layer 3 of the network were similar to those in the output (4th) layer. That is, cells in layer 3 were both object selective (responding exclusively to their preferred object) and highly transform (view) invariant. In layer 2 of the network, cells were object selective but showed more modest levels of transform invariance due to the limited convergence of feedforward connections from the retina. Individual layer 1 cells receive projections from a very limited region of the retina. These cells were object selective but provided very low levels of transform invariance.

3.3.3 Analysis of Partial View Response

To better understand how the network has learned to represent the partially occluded object, the different fragmented partial views of the occluded object that were obscured at different times during training were presented separately to the model during testing for all 360 views. This is a fundamen-

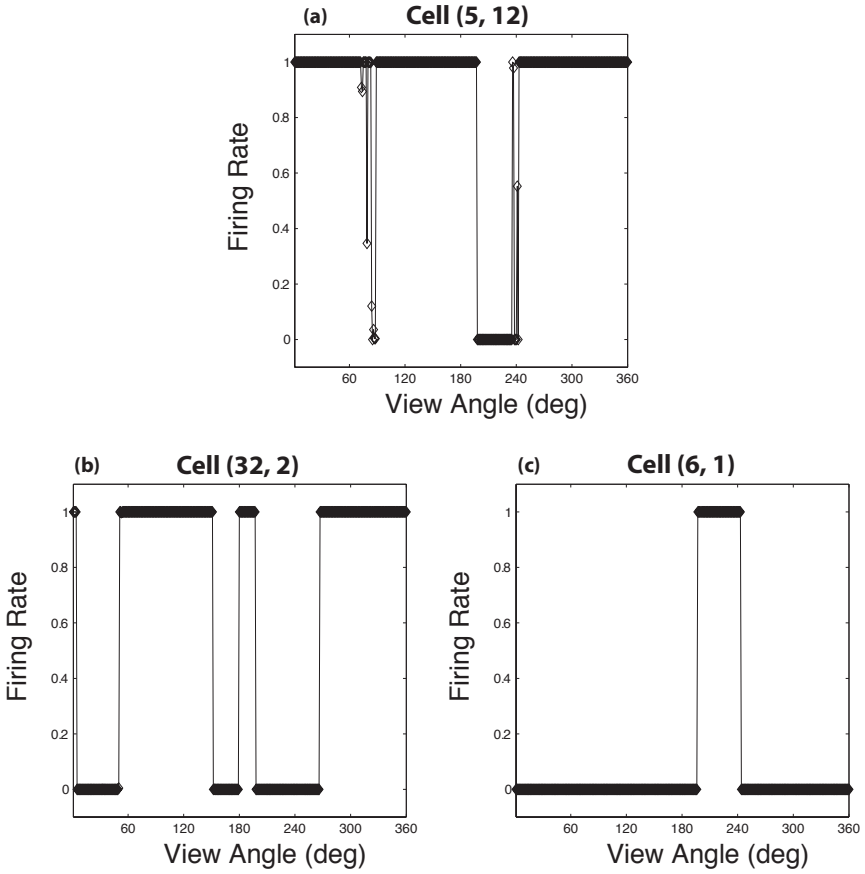

Figure 3.12: The firing rate responses of three cells in the 4th (output) layer of VisNet. These cells were found to respond exclusively to the Jaimoid, where they fire to certain views of the Jaimoid but to no views of any other objects. Plot (a) shows the response profile of cell (5, 12) which has learned to respond to the Jaimoid over most of the views. Plot (b) shows the response profile of cell (32, 2), which has learned to respond to the Jaimoid over an overlapping region of the view space including some points that the previous cell did not respond to. Plot (c) shows the profile of cell (6, 1), which has learned to respond to the Jaimoid over the remainder of the view space to which the two previous cells did not respond.

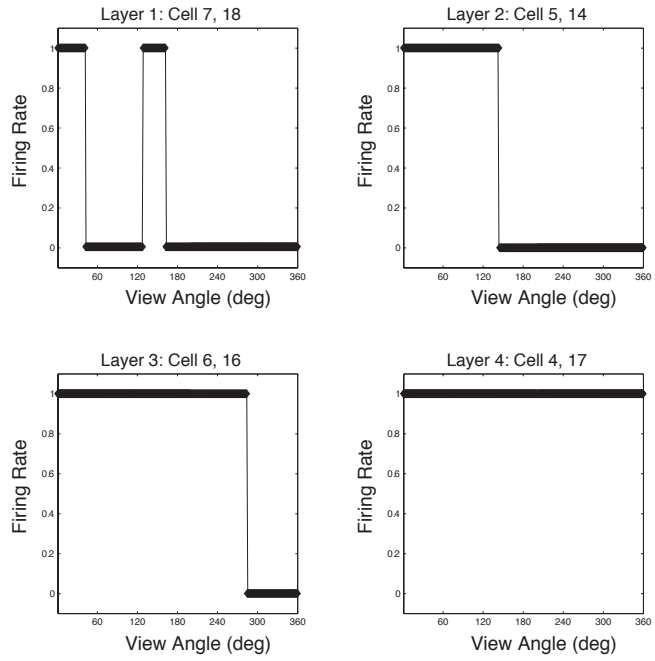

Figure 3.13: Example Jaimoid response plots for all four layers. A prototypical example cell is presented for each of the four layers within the VisNet model. For each cell, its response plot is presented with respect to the example object, the Jaimoid. It can be seen that a typical layer 3 cell is highly transform invariant while a layer 2 cell shows more modest levels of view invariance. Layer 1 cells demonstrate very little view invariance and responded to a very narrow set of views. In all cases, these cells responded exclusively to their preferred object, in this case the Jaimoid, and did not learn to respond to any views of any of the other objects. Comparable responses exist for all six of the objects presented during training.

tal test to ensure that output neurons have learned to respond to the fragmented parts of the partially occluded object. A similar but easier test would have been to only present the two different halves of the Jaimoid to the network for testing. By presenting the smaller fragmented views, it may be shown that VisNet has successfully learned to bind these partial views together to form a complete invariant representation of the Jaimoid. This test is necessary to show that output neurons do not rely on a key-marker or partial view of the Jaimoid in order to recognise it.

Figure 3.14 shows that neurons in the output layer of the VisNet model were able to successfully bind together all of the different partial views into a holistic invariant representation. Specifically, exactly the same output neurons (e.g. cell 4, 17) that responded invariantly when presented with the complete Jaimoid (e.g. Figure 3.9) were also activated in an identical manner when presented with the various partial views. Each and every partial view caused the same output neurons to respond invariantly as if the whole object had been presented in its entirety.

This important novel result shows that a feedforward hierarchical model of the ventral visual system such as VisNet does not need to rely on particular parts, or key-markers, of an object in order to recognise it. Furthermore, it shows that such a biologically inspired network is able to not only build invariant representations of individual objects despite the fact that pairs of objects were presented during training, but it also shows that such a network can solve a far more complex problem, that is, building invariant representations in a multi-object environment even when the different objects are partially occluding one another, as is often the case in the real world.

3.3.4 Location vs. Object Selectivity

An important question is whether the output cells learned to respond to the visual form of the objects or merely to the retinal locations. In order to minimise the possibility that the output cells had learned

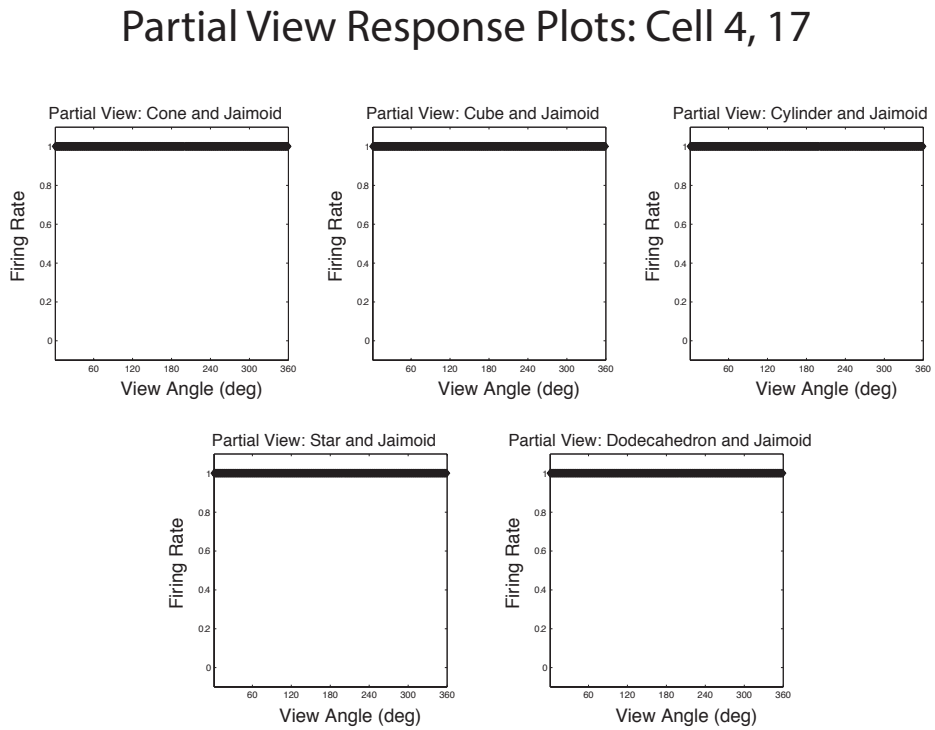

Figure 3.14: Partial view response plots for cell 4, 17. Cell response plots are presented after testing the network with the fragmented partial views of the Jaimoid, as exemplified in Figure 3.7. It can be seen that the example cell 4, 17 that responded invariantly to the complete view of the Jaimoid also responds in an identical manner to the different fragmented partial views of the Jaimoid. Cell 4, 17 has learned to bind together these different partial views into a holistic representation and responds equally well to all of them, thus proving that this cell does not rely on a specific partial view or key-marker in order to recognise the Jaimoid.

to respond to the locations, the objects were presented to the network in overlapping locations as shown in Figure 3.5. Even with the objects presented in highly overlapping locations, the output cells learned to respond to the objects themselves, and to no views of any of the other partially overlapping objects.

The network was also tested with six novel objects, such as a pyramid, rotating through 360° in 1° steps. These objects are novel in the sense that the network was not trained with them and was only exposed to them for testing. The novel objects were presented in the same locations as the original occluding and occluded objects. If the network had learned to respond to the individual trained objects rather than the locations, then the responses of the output cells to the novel objects should be less clearly tuned than to the trained objects. That is, the cells should not respond so uniformly (invariantly) over the different views of any particular novel object, and the cell responses should not be selective to individual novel objects. Figures 3.15 and 3.16 show cell response plots for cell (4, 17) and (19, 1) after testing the network with the novel objects. To reiterate, when tested with the original set of objects, cell (4, 17) responds invariantly to the Jaimoid, and to none of the views of the other objects (Figure 3.9 (b)). However, when the network is tested on six novel objects presented in the same locations as the trained set of objects, the cell responds very poorly to small portions of view of a number of objects. Similarly, when tested with the original set of objects, cell (19, 1) responds invariantly to all 360 views of the Dodecahedron and to no views of any other object (Figure 3.10 (b)). When tested on the six novel objects presented in the same locations as the trained set of objects, the cell responds very poorly to small portions of view of a number of objects. These results demonstrate that the network has learned to respond to the trained objects in particular, and not just to their locations.

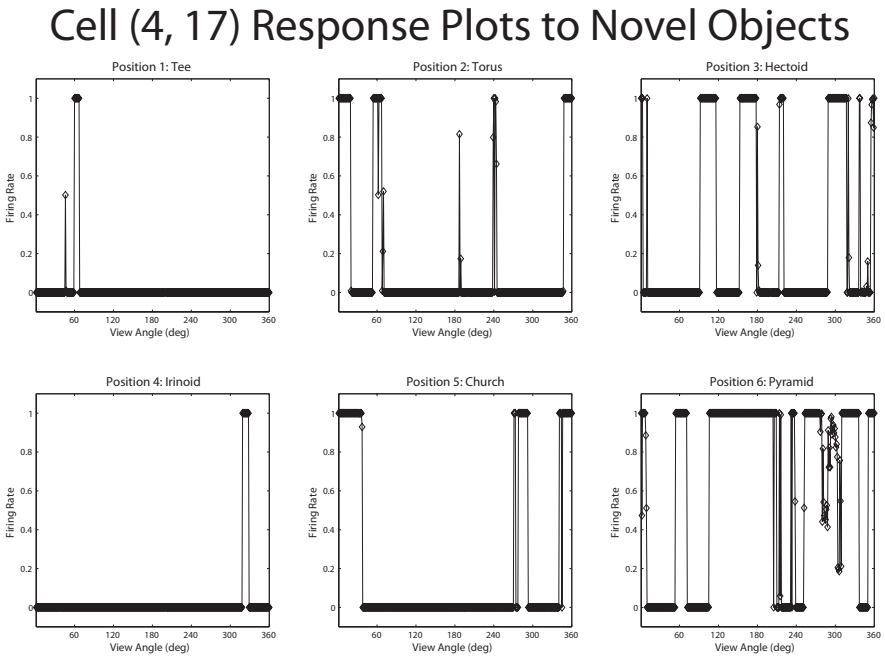

Figure 3.15: The firing rate responses of cell (4, 17) in the 4th (output) layer of VisNet to six new objects on which the network was not trained, as they rotated through 360 ° in 1 ° steps after training. It can be seen that the cell's response pattern has changed compared to its response pattern to the six objects on which the network was trained: the occluded and the five occluding objects (Figure 3.9 (b)). When tested with the set of objects with which the network was trained, the cell responds to all of the views of Jaimoid, and to none of the views of the other objects. When the network is tested on six novel objects presented in the same locations as the trained set of objects, the cell responds very poorly to small portions of view of a number of objects. This shows that the network has learned to respond to the trained objects in particular, and not just to their locations.

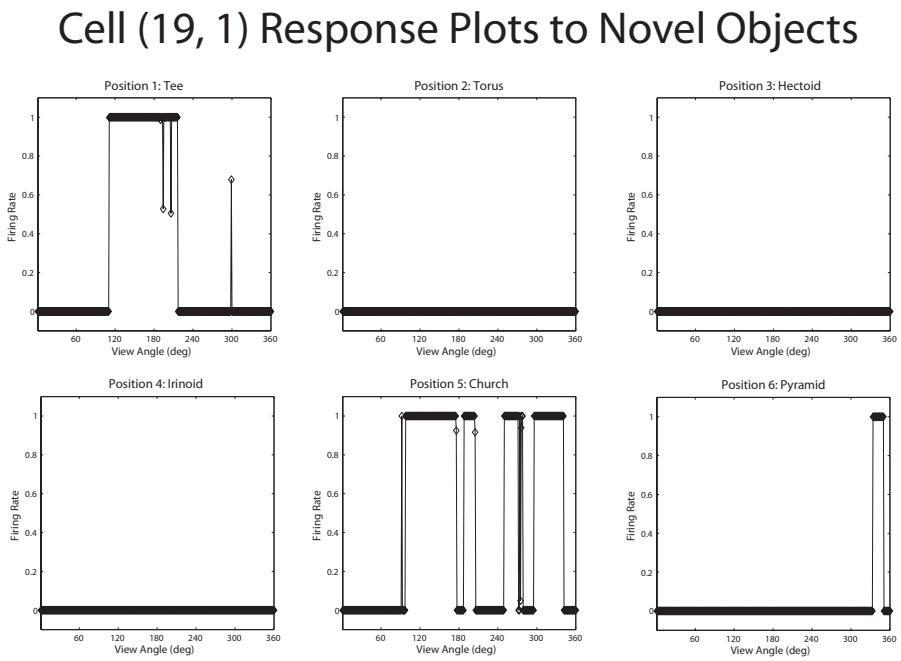

Figure 3.16: The firing rate responses of cell (19, 1) in the 4th (output) layer of VisNet to six new objects with which the network was not trained, as they rotated through 360 ° in 1 ° steps after training. It can be seen that the cell's response pattern has changed compared to its response pattern to the six objects with which the network was trained: the occluded and the five occluding objects (Figure 3.10 (b)). When tested with the set of objects with which the network was trained, the cell responds invariantly to all 360 views of the Dodecahedron and to no views of any other object. When tested on the six novel objects presented in the same locations as the trained set of objects, the cell responds very poorly to small portions of view of a number of objects. This shows that the network has learned to respond to the trained objects in particular and not just to their locations.

3.3.5 Information Analysis

Single cell information analysis was conducted to confirm whether the network had developed cells that responded invariantly to their preferred object (Figure 3.17). The unbroken line represents the results obtained after presenting the six original trained objects to a network after training on the 360 views of all possible object pairs. The dashed line represents the results obtained after presenting six novel objects rotating in the same positions as the six original trained objects. The dotted line represents the results after presenting the six original objects to a random untrained network. Single cell information measures are shown for the 4th layer neurons, ranked in order of the amount of information they convey.

It can be seen that training the network on the object pairs has lead to many of the 4th layer neurons attaining the maximal level of single cell information of 2.58 bits for the trained objects calculated using Equation 1.14. These neurons have learned to respond to all of the views of their preferred object. However, when the network, which had been trained on the six original objects, was tested with novel objects, no cells reached the maximum level of information. This reflected the fact that the output cells of the trained network were not able to respond to the novel objects in a view-invariant or object-selective manner, as shown in Figures 3.15 and 3.16. These results thus further demonstrate that when the network was trained on the six original objects, the output cells had learned to respond selectively to the trained objects and not the untrained objects. This in turn confirms that the network learned to respond to the visual forms of the trained objects rather than their retinal locations.

However, it is unclear from the single cell information analysis whether all of the six objects are individually represented by a unique subset of invariant output cells. Indeed, it is possible that these cells are responding to the same object and are therefore unable to provide information regarding

which object is present. To ensure that there are cells that respond preferentially to each of the six objects multiple cell information analysis was performed.

Figure 3.18 shows the multiple cell information analysis obtained when VisNet was tested with the six individual trained objects rotating through 360 ° in 1 ° steps. Multiple cell information analysis results are also plotted for six novel objects on which the network was not previously trained. These novel objects were rotating in the same positions as the six objects on which the network was originally trained. Results are presented having tested the trained network with the original set of trained objects (unbroken line), after testing the trained network with the novel set of objects (dashed line), and after presenting the six original objects to a random untrained network (dotted line). After the network was trained and tested with the original set of objects, over 2.5 bits of information was reached (substantially higher than 1.1 bits reached by the untrained network or 1.7 bits reached by the trained network that was tested on the novel set of objects) suggesting that the single cell information results included cells that preferentially responded to all 6 objects. These plots show that the network did not learn to respond to the locations of the objects, and instead bound together different views of the occluding and occluded object to form object specific representations. This was also confirmed by inspection of the cell response plots as shown in Figs. 3.9 (b) and 3.15, as well as Figures 3.10 (b) and 3.16.

3.4 Discussion

An important question in natural vision is how the brain forms invariant representations of objects that are always partially occluded by other objects during learning. In a real world visual environment, this will often be the case. Stringer and Rolls (2008) have shown that a biologically plausible

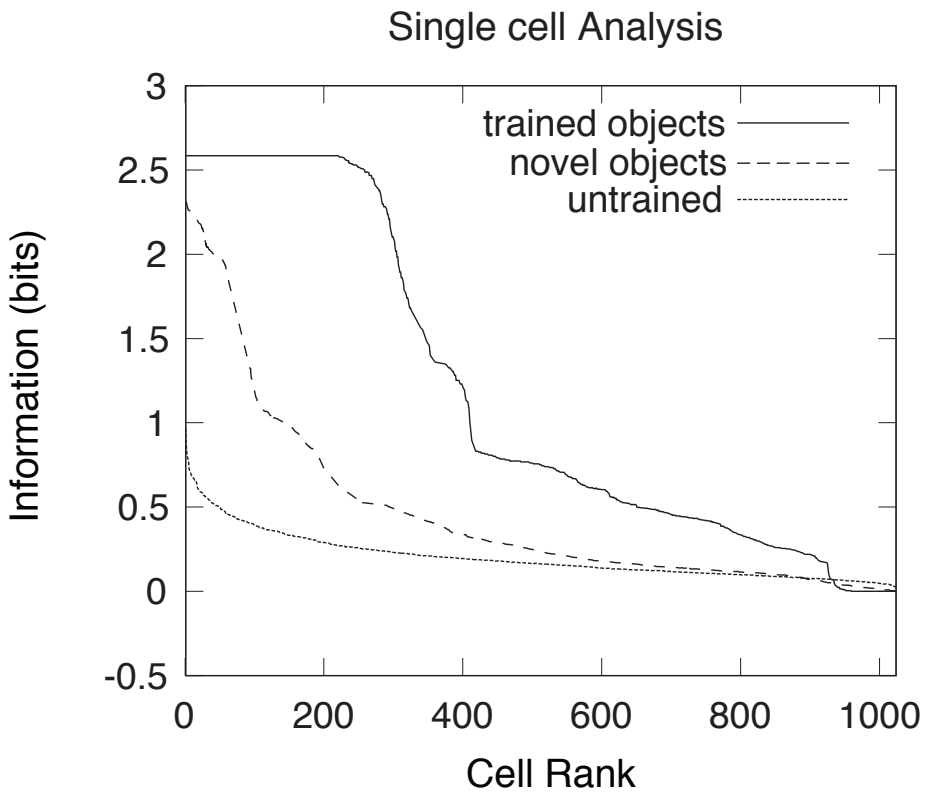

Figure 3.17: Single cell information results obtained when VisNet was tested with the occluded object and five occluding objects rotating through 360 ° in 1 ° steps. Single cell information analysis results are also plotted for six novel objects on which the network was not previously trained. These novel objects were rotating in the same positions as the six objects with which the network was originally trained. Results are presented having tested the trained network with the original set of trained objects (unbroken line), after testing the trained network with the novel set of objects (dashed line), and after testing a random untrained network with the original six objects (dotted line). The single cell information measure is shown for all 4th layer neurons ranked in order of the amount of information they convey. It can be seen that training the network on the object pairs has led to many 4th layer neurons attaining the maximum level of single cell information of 2.58 bits for these trained objects. These cells have learned to respond selectively to individual trained objects invariantly over all views. It can also be seen that the novel objects produce less information and no cells reached the maximal information. The random untrained plot provides a baseline comparison.

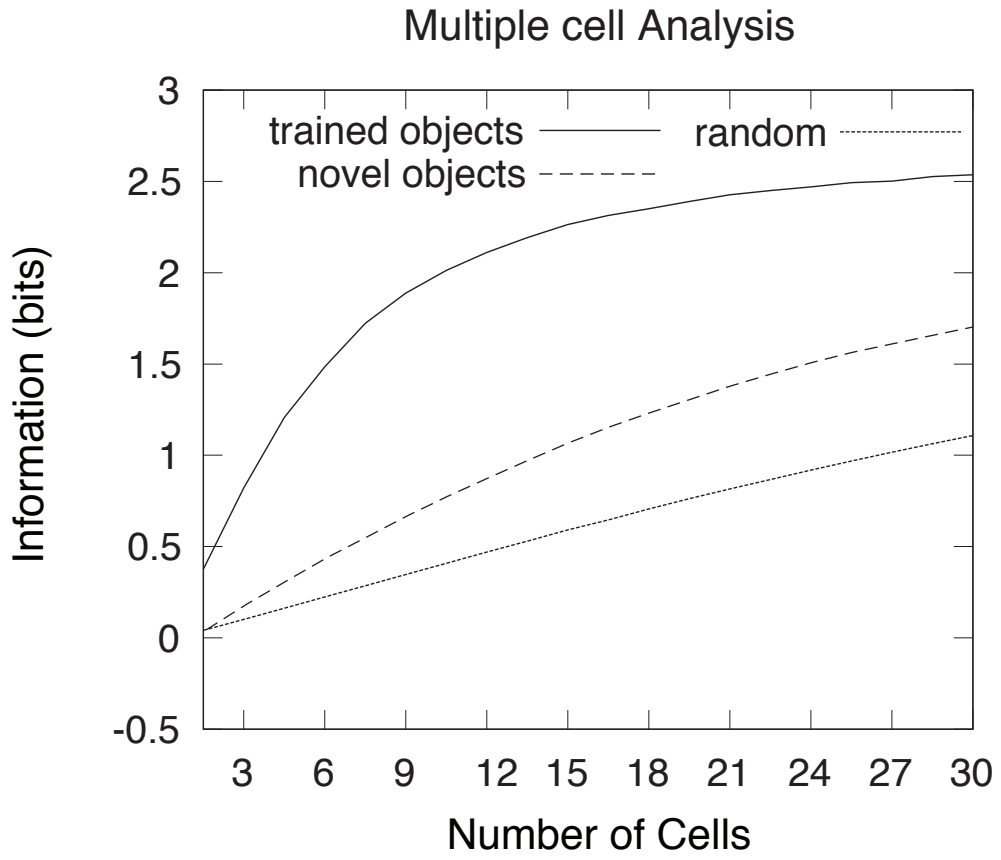

Figure 3.18: Multiple cell information results obtained when VisNet was tested with the original trained occluded object and five occluding objects rotating through 360 ° in 1 ° steps. Multiple cell information analysis results are also plotted for six novel objects that the network was not previously trained on. These novel objects were rotating in the same positions as the six objects on which the network was originally trained. Results are presented after training the network (unbroken line), after testing the trained network with the novel set of objects (dashed line) and after presenting the trained objects to a random untrained network (dotted line). After the network was trained, over 2.5 bits of information was reached when the original trained objects were presented. This was substantially higher than 1.10 bits reached by the untrained network or 1.7 bits reached by the trained network that was tested on the novel set of objects. This confirmed that, after training the network, there were cells that form object specific representations to each one of the six trained objects and do not respond to the object locations.

model, VisNet, comprised of a hierarchical series of competitive neural networks, can develop invariant representations of individual objects when no single object is seen in isolation. In this chapter it is demonstrated how such a network might form an invariant representation of an object that is always partially occluded by other objects. The mechanism employed for invariance learning is Continuous Transformation (CT) learning. CT learning uses the spatial continuity between the views of individual objects as they transform in the real world, combined with associative learning of feedforward connection weights.

It was found that, after training the network with a rotating object that is always partially occluded, the network is able to form view invariant representation of the partially occluded object. In addition, by testing the network with the fragmented partial views of the occluded object in isolation, it was also shown that the same output neurons that learned to respond to the Jaimoid when presented in its entirety also responded in an identical manner to each of the fragmented partial view sequences. This shows that the network has learned to bind together the different partial views of the occluded object presented during training into a holistic invariant representation despite always seeing the Jaimoid partially occluded and therefore never in isolation. It was also found that view invariant representations are also formed for all five occluding objects. This is a challenging task since the occluding objects were always overlapping the occluded object and therefore VisNet had to learn to separate the objects. This is the first time such learning has been shown to happen in a biologically inspired model of the ventral visual system.

Despite the fact that the objects were presented in the same location during training and testing, the network was able to form representations of the objects' identities instead of just learning to respond to particular locations. During training there was significant overlap between the objects (Figures 3.5

& 3.6), which would have precluded VisNet from learning about each object just because it was presented in the same location. Instead, VisNet built separate representations of each of the individual objects. This is confirmed by invariant cells that responded maximally to all views of only one of the stimuli and not to any views of any other stimuli. After training, all of the stimuli were represented in this way. Given that the objects were highly overlapping during training, if the network had learned to respond to location instead of stimulus identity, then the network would not have developed cells which responded specifically and invariantly to individual objects, with all of the objects represented uniquely in this way.

This conclusion is also confirmed by additional results obtained with a novel set of six objects presented during testing. These novel objects were presented in the exact same locations as the six original objects that were presented during training. The output neurons that learned to respond preferentially and invariantly to the original trained objects were then inspected and an example response plot was presented. It was shown that these output neurons did not respond in a exclusive or invariant manner to any of these novel objects thus confirming that the network had learned object selectivity. The residual firing that was present within the response plots can be explained by the fact that after the network has been trained, neurons will respond in proportion to how similar the novel input pattern is to the previous learned patterns. By virtue of the fact that the input spaces are not orthogonal with respect to the input filters that they activate when objects occupy the same input space on the retina, they will cause a degree of activity in the network based on previous learning. This is a fundamental property of competitive networks, which will try to generalise to novel input patterns depending on their similarity to the previous learned input patterns.

In addition to the invariant cell responses, individual objects were also represented by populations

of cells that responded invariantly to different parts of the view space and together represented the object through all 360 views. These cells encode the viewing angle of the object and because they learned to respond exclusively to one preferred object, they also encoded the identity.

Most artificial computer vision systems do not seek to mimic processing exactly as it is carried out in the brain. Also, the challenges addressed by artificial computer vision systems are often more focussed, for example, on the problem of object or face recognition after training the system with individual segmented objects or faces. For such a task, non-biologically inspired artificial visual systems often rely on either template matching or searching for the presence of a subset of key features in order to recognise a partially occluded object (Ullmann, 1992; Ying and Castaon, 2002; Do et al., 2005). However, computational neuroscientists are interested in the more ecological problem of understanding how the primate visual system learns in an unsupervised manner to make sense of complex natural visual scenes containing multiple objects. This is a valuable long term goal, which may ultimately offer engineers powerful new approaches to intelligent visual scene analysis and object recognition. Understanding how the primate brain learns to process visual input from scenes will involve the step-by-step uncovering of many key neurodynamic mechanisms, which ultimately blend and work together in the brain. This current study examines the problem of how the primate visual system might develop separate transform-invariant representations of individual objects even if these objects are always seen partially occluding each other during unsupervised learning. The simulations reported within this chapter have shown for the first time how a biologically plausible model of the ventral visual pathway, VisNet, with a Hebbian associative synaptic learning rule, is able to solve this particular problem.

3.4.1 Future work

The results described above have shown how a biologically plausible neural network model of the ventral visual pathway, VisNet, is able to develop object-selective and rotation-invariant representations of objects that were partially occluding each other during training. However, more work needs to be done to explore the limits of this mechanism.

For example, what proportion of an object needs to be visible? Future research could investigate the exact degree to which objects can be occluded before learning of the partially occluded object breaks down. This avenue of research could address the effect of occlusion by two or more objects at a time, or where there is more than one occluded object. The results of these experiments presented within this study suggest that the extent to which the partially occluded object is covered could increase quite considerably so long as the different parts of the occluded object have all become visible at some point during training.

The use of simple geometric shapes is a limitation of the current study that should be addressed as part of future research. The choice to use simple geometric shapes was not to help VisNet solve the task at hand. These shapes allowed for a level of control necessary to answer the question, "how has VisNet solved this problem"? This level of control is harder to achieve with more complex objects, but their use is a sensible next step to explore their effect on the self-organisation of the network. The types of objects used will not have a qualitative impact on the results presented within this chapter, but this should be confirmed. So long as the resolution of the retina is high enough to convey the necessary detail of the more natural objects, then the VisNet model will make use of the same principles discussed within this study to solve the problem.

In the simulations described above, individual objects rotated on the same part of the retina. Per-

haps a more challenging problem is the *translation* of objects across the retina. This would happen naturally as an observer moves past objects in a scene, such as when a person walks past a chair, or behind a picket fence. In this case, all of the objects would be seen moving over the retina. The input representations of the objects would overlap over different locations on the retina. Yet the network must still form separate output representations of the objects, which are also translation invariant. It is hypothesised that the network described in this chapter should still be able to solve this problem using similar learning principles.

Another limitation of the current study is that it explores only one type of invariance learning mechanism, Continuous Transformation (CT) learning (Stringer et al., 2006). This binds different transforms of a particular object together by exploiting the spatial similarity that exists between the different transforms of that object. As discussed, CT learning relies on a simple Hebbian learning rule. It would be very interesting to investigate if similar results can be achieved with a different type of biologically plausible learning rule such as Trace learning (Foldiak, 1991; Wallis et al., 1993). This alternative learning rule exploits temporal continuity of successive transforms of an object in order to build a transform-invariant representation of that object by utilising a memory trace of the recent firing of the post-synaptic cell.

Chapter 4

Separation of Object Representations by Independent Motion

The previous two chapters have demonstrated that VisNet is able to develop separate view invariant representations of individual objects when pairs of objects are always presented to the model during training. In both chapters, pairs comprising different objects were required in order for VisNet to successfully self-organise in this manner. However, what happens when different object pairs are not available during training, and the training is limited to just one pair?

This chapter reports the results obtained when VisNet is only trained with one pair of objects that are always presented together during training, but where the objects rotate independently of each other. In particular, this chapter explores whether independent motion is able to help VisNet learn to develop separate representations of individual objects.

4.1 Introduction

As previously outlined, the ventral visual system plays a significant role in the processing of visual form and ultimately object identification (Goodale and Milner, 1992). Neurophysiological studies suggest that these regions are organised in a hierarchical fashion, where neuronal responses become more complex and receptive field sizes increase towards the latter stages of the ventral stream (Gattass et al., 1988; Levitt et al., 1994; Gegenfurtner et al., 1997). Over successive stages, the visual system develops neurons that respond with view, size and position (translation) invariance to objects or faces (Hasselmo et al., 1989b; Desimone, 1991; Rolls et al., 1992; Perrett and Oram, 1993; Kobatake and Tanaka, 1994; Tovee et al., 1994; Ito et al., 1995; Booth and Rolls, 1998; Op De Beeck and Vogels, 2000).

Given these findings, a major computational challenge is to understand the fundamental mechanisms of how the ventral stream may actually form separate neuronal representations for individual objects, despite the fact that multiple objects are always presented together in complex natural scenes. The research presented in this chapter investigates how the primate visual system may build invariant representations of individual objects when multiple objects are always present in a scene. How the visual system learns representations of individual objects instead of just the combined scene is an important question for natural vision.

Recent research has successfully shown that it is possible for the VisNet model of the ventral visual stream to use the statistics in the training input to build separate invariant representations of objects used during training, despite the fact that pairs of objects were always presented (Stringer et al., 2007; Stringer and Rolls, 2008). Many different object pairings were required to produce a training set where each object was combined with the other objects to create a selection of object pairs. Crucially,

during presentation the features that make up an object occur together more frequently compared to the features that make up different objects. The frequency with which the objects are presented together during training denotes the level of correlation between features from different objects. However, the features that comprise individual objects are always seen together. It has been shown that a competitive network will operate usefully in this situation forming invariant representations of individual objects, rather than the combinations of objects seen during training. Invariance may be built by a mechanism such as Continuous Transformation learning (Stringer and Rolls, 2008) or trace learning (Foldiak, 1991). The number of different object pairs that must be presented to the network during training may make this training method implausible. However, the previous two chapters of this doctoral thesis have progressed this research and have shown that the VisNet model is able to use similar learning mechanisms to develop separate invariant representations of objects even when there is significantly reduced training, or objects partially occlude one another during training.

There may be other underlying mechanisms available to assist this type of object selective learning. In this chapter, for the first time, an investigation is undertaken to determine whether a multi-layer hierarchical model of the ventral visual pathway is able to use independent rotation to form separate object selective representations when trained with just two realistic 3D rotating objects always co-present on the retina. In this case, the mechanisms described by Stringer and Rolls (2008), which rely on the presentation of many object pairings, cannot be exploited.

The performance of a competitive network is investigated, as well as the performance of a self-organising map (SOM). The SOM is produced by introducing short-range excitatory connections and long-range inhibitory connections (von der Malsburg, 1973; Kohonen, 1982). This lateral SOM connectivity uses a "Mexican-hat" profile and encourages spatially proximal neurons to develop similar

response properties due to their mutual excitation while neurons that are relatively far apart will experience mutual inhibition. This drives competition to create separate pools of functionally distinct neurons.

It was found that statistical decoupling of the learned output representations can occur between two objects through independent rotation, even when the objects are always presented together concurrently during training. Crucially, the features that comprise a view of an object are always presented together and these features are not regularly presented with a specific view of the alternative object. Independent rotation ensures that features between objects remain statistically decoupled. After learning, separate output representations for the two objects develop.

It was observed that a SOM network architecture is able to form output neurons that respond with complete view invariance to their preferred object. Although the competitive network architecture was also able to form neurons with object selective representations, they were not fully invariant over all views. Instead, these neurons primarily responded to small regions of contiguous views of their preferred object and a population of neurons is required to encode all the views of each object presented during training.

4.2 Method

4.2.1 Model Architecture and Parameters

The experiments presented as part of this chapter were conducted using the VisNet model architecture as described in Section 1.3.3. However, for the first time the original competitive network architecture is replaced within each layer with a self-organising map (SOM). That is, some model simulations adjusted the local graded inhibition within a layer to incorporate short-range excitation and longer-range

inhibition (von der Malsburg, 1973; Kohonen, 1982). Accordingly, in these simulations the VisNet model consists of a hierarchical series of four layers of SOMs, corresponding to V2, V4, the posterior inferior temporal cortex, and the anterior inferior temporal cortex. Images were filtered in accordance with the difference of Gaussian Equation 1.1 previously presented. The sparseness percentile and slope values are presented in Table 4.1. The lateral inhibition parameters for the competitive network are presented in Table 4.2 and the lateral inhibition and excitation parameters for the SOM are provided in Table 4.3, respectively.

Layer	1	2	3	4
Percentile	96	96	96	96
Slope β	190	40	75	26

Table 4.1: Sigmoid parameters

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

Table 4.2: Typical competitive network lateral inhibition parameters

Layer	1	2	3	4
Excitatory Radius, σ_E	2.1	1.65	1.2	1.8
Excitatory Contrast, δ_E	5.35	33.15	117.57	120.12
Inhibitory Radius, σ_I	4.14	8.1	12.0	18.0
Inhibitory Contrast, δ_I	1.5	1.5	1.6	1.4

Table 4.3: Typical self-organising map lateral inhibition and excitation parameters

4.2.2 2D Block Stimuli

As part of a pilot study, a pair of 2D blocks was presented to the network during training. This pilot study replaced independent rotation with independent translation in order to provide a greater level of control over the statistical decoupling between two different stimuli, as well as the spatial overlap between the different transforms of the same stimuli. This receives further comment in the

Discussion section of this chapter. In summary, the use of these carefully controlled blocks was necessary to first establish whether VisNet was able to use independent movement in the most ideal of training environments, as a method for separating two stimuli from one another during training with independent motion.

These blocks were presented on different halves (left and right) of the VisNet retina and gradually translated over orthogonal retinal regions. They were filtered in accord with the normal filtering routines outlined in Chapter 1, Section 1.3.3. The statistical decoupling mechanism involved in this translation paradigm forms the theoretical basis for the more complex 3D rotating objects paradigm.

4.2.3 2D Blocks Training Procedure

Figure 4.1 shows an annotated example frame of these two blocks translating during training. The block presented on the left half of the VisNet retina was paired with the block from the right half of the VisNet retina. In each half of the VisNet retina, there were 40 possible locations within which a block could be situated. Each block from either half was paired with the other, also over 40 possible locations. As such, in total there were 1600 transforms presented to the network during training. For example, when the left block was in position one (1 of 40), it was separately presented with the block from the right half of the retina over all 40 possible locations, creating 40 pairings in total. Then, the left block was shifted into position two (2 of 40), and was subsequently presented with the block from the right half of the retina over all 40 possible locations once again. This creates an additional 40 pairings, producing a running total of 80 pairings. This process was repeated such that the block occupying the left half of the VisNet retina was presented in all 40 possible locations, on each occasion with the block on the right half of the retina across all 40 possible locations. In total, this created 1600 pairings to present to the network during training.

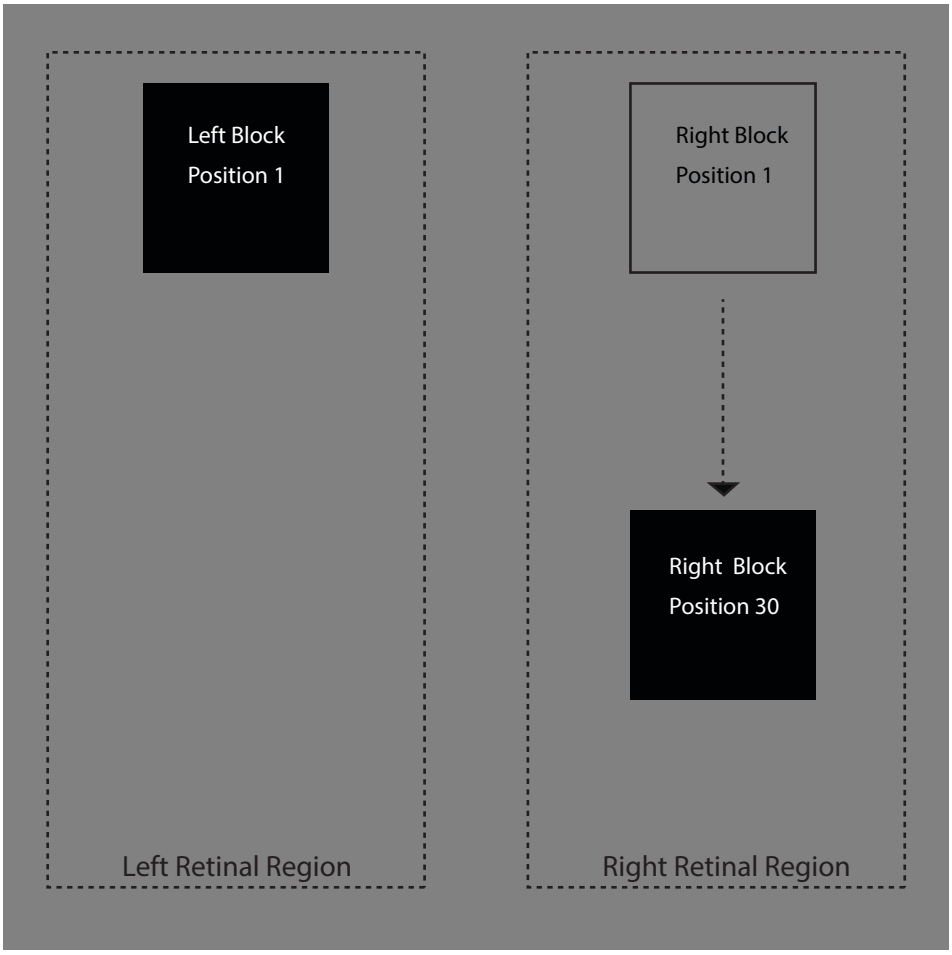

Figure 4.1: Pairings of two block stimuli. This annotated figure depicts how the left and right retinal regions are configured, and how the left block and right block translate over their retinal regions to create all 1600 transforms used during training. Within the figure, it can be seen that the left block is currently in position 1 and the right block is currently in position 30, having originally started from position 1.

4.2.4 2D Objects Testing Procedure

To test the network's performance when trained with the 2D block stimuli, the left block and the right block were presented to the network individually. In each case, the block was presented shifting over the assigned region of the retina over all 40 locations. The results are presented as cell response plots in Section 4.3 below, and form the basis of the experiments performed with 3D rotating objects.

4.2.5 3D Object Stimuli

In the main research experiments, three 3D rotating visual stimuli were used to create three separate pairings. These pairings were then used to train the VisNet model. Each pairing was run as a separate experiment and the objects used to create the pairings comprise a Cube, Dodecahedron and custom designed Hectoid. The Hectoid is an irregular multifaceted three dimensional object. To ensure that these results were robust, these three separate pairings with three very different objects were used.

Figure 4.2a shows the Cube and the Hectoid, Figure 4.2b shows the Cube and Dodecahedron and Figure 4.2c shows the Dodecahedron and the Hectoid. In each experiment, the same stimuli were also used during testing and were continuously rotating over 360°. Each object was designed and created in Electric Rain's vector based 3D modelling software, Swift 3D. Extra measures were taken to ensure that image outlines and borders were clearly defined to avoid any artefacts that may occur due to the size of the VisNet retina. Furthermore, ambient lighting with diffused light sources were used to ensure that different surfaces were shown with different illuminations.

4.2.6 3D Objects Training Procedure

The three pairings were presented to the network in separate experiments in exactly the same manner. The following section describes the process using the Cube and the Hectoid, but the process is the

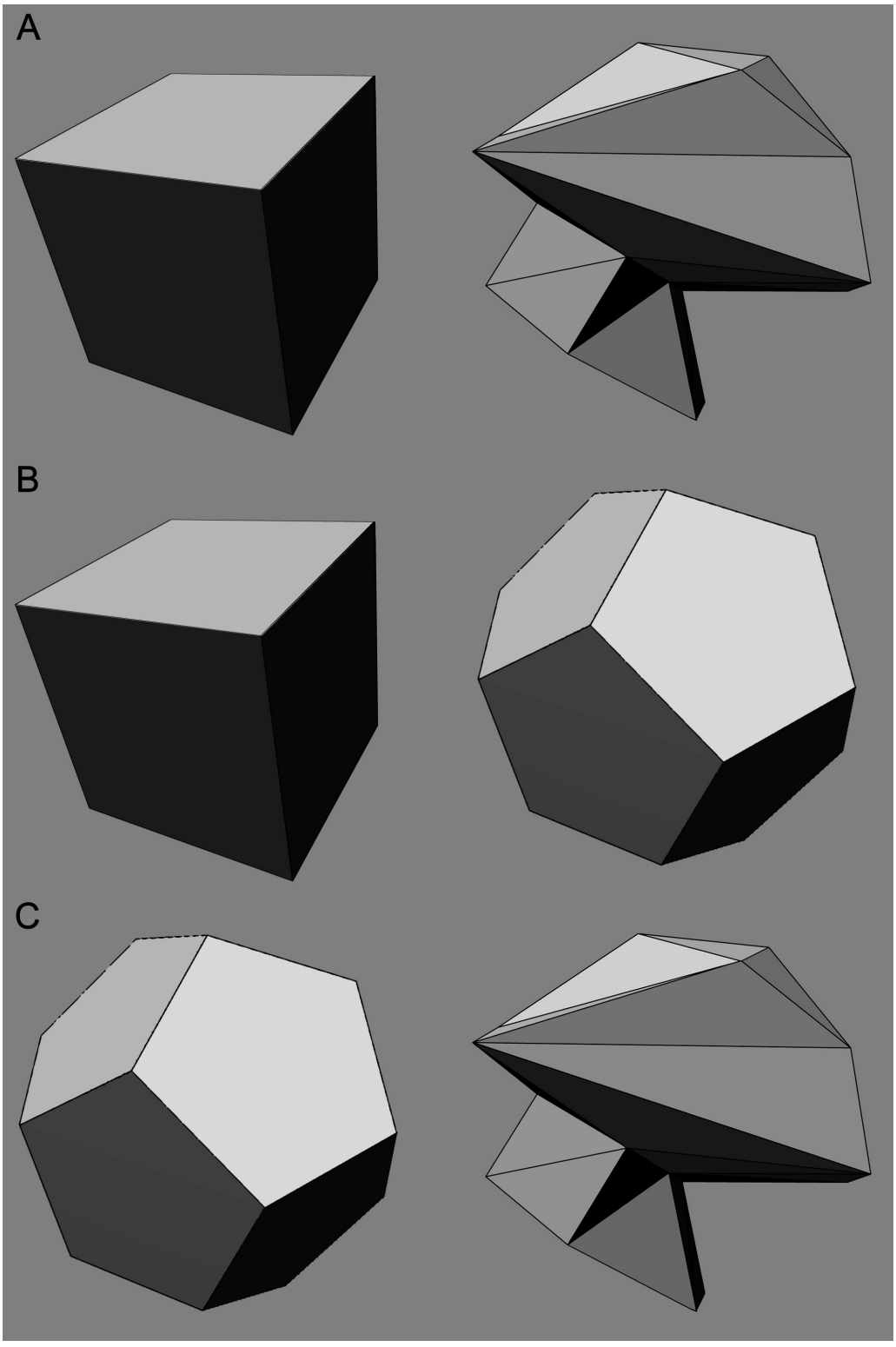

Figure 4.2: Stimuli pairings. Example images of the three pairs of objects used to train the model. A: Cube (left) and Hectoid (right); B: Cube (left) and Dodecahedron (right); C: Dodecahedron (left) and Hectoid (right). Each pair was presented as a separate experiment and the same objects were used during training and testing

same for the Cube and the Dodecahedron as well as the Dodecahedron and the Hectoid.

In order to simulate independent rotation, different views of object one, the Cube, were presented with different views of object two, the Hectoid. Specifically, each view of the Cube was presented with ten equally spaced views of the Hectoid. This process was repeated for the Hectoid, where every view was presented with ten equally spaced different views of the Cube. Each view of any given object was presented as frequently as any other view.

For example, in 'block 1' of training, a specific view of the Cube at an angle of 1 ° was presented with ten views of the Hectoid at 1 °, 37 °, 73 °, 109 °, 145 °, 181 °, 217 °, 252 °, 289 °, 325 ° respectively. In 'block 2' of training, a specific view of the Cube at an angle of 2 ° was presented with ten views of the Hectoid at 2 °, 38 °, 74 °, 110 °, 146 °, 182 °, 218 °, 253 °, 290 °, 326 ° respectively. This was repeated for every view of the Cube, therefore creating 360 Cube training 'blocks'. Similarly, 360 Hectoid training 'blocks' were also created. In total, 7200 object pairings were presented to the network, and together these comprise one epoch. The performance of the network was explored with a SOM and competitive network architecture within each layer. In all experiments, the learning rate of the model was set to 0.01 and as already outlined, the sparseness was set to 0.04. Other values were explored (not presented) and receive comment in the Discussion, Section 4.4.

4.2.7 3D Objects Testing Procedure

To test the network's performance within each experiment, each of the relevant objects was presented to the model in isolation. For example, all 360 views of either the Cube or the Hectoid were presented separately, and the firing rates from the output (fourth) layer were recorded. The results are presented as polar plots. This same procedure was used for the experiments with the Cube and the Dodecahedron pairing as well as the Dodecahedron and the Hectoid pairing.

In addition to the polar plots produced from the tests outlined above, information analysis measures were also performed. The procedures for the information analysis are presented in Section 1.3.5.

4.3 VisNet Simulation Results

The results obtained when presenting VisNet with the translating 2D blocks are presented in Figure 4.3 and Figure 4.4. During training, pairs of 2D blocks are presented to the network. Each block could occupy one of two retinal regions, where each region comprised the 40 possible locations. Accordingly, there were 1600 block pairings presented to the network during training. The figures show two prototypical example cells prior to learning (Figure 4.3), and after learning (Figure 4.4), having received testing with each block translating separately over the designated retinal region as outlined above. Prior to learning, the cells do not respond preferentially to either block and furthermore respond sporadically to different views of the left block and no views of the right block. However, after training the output neurons have learned to respond exclusively to their preferred block over all 40 possible transforms.

These results show that the network has successfully developed separate representations for each 2D block by using independent motion as a mechanism for statistically decoupling the blocks. There is a strong CT learning effect due to the carefully controlled overlap present between the different transforms of each block. Accordingly, the output representations are also translation invariant. These results are important since they show that, in principle, VisNet is able to use the statistics inherent within independent motion to form separate invariant object representations. These results form the basis for the 3D rotating objects results, which are presented next.

In each 3D object experiment the VisNet model was trained with one of the three different pair-

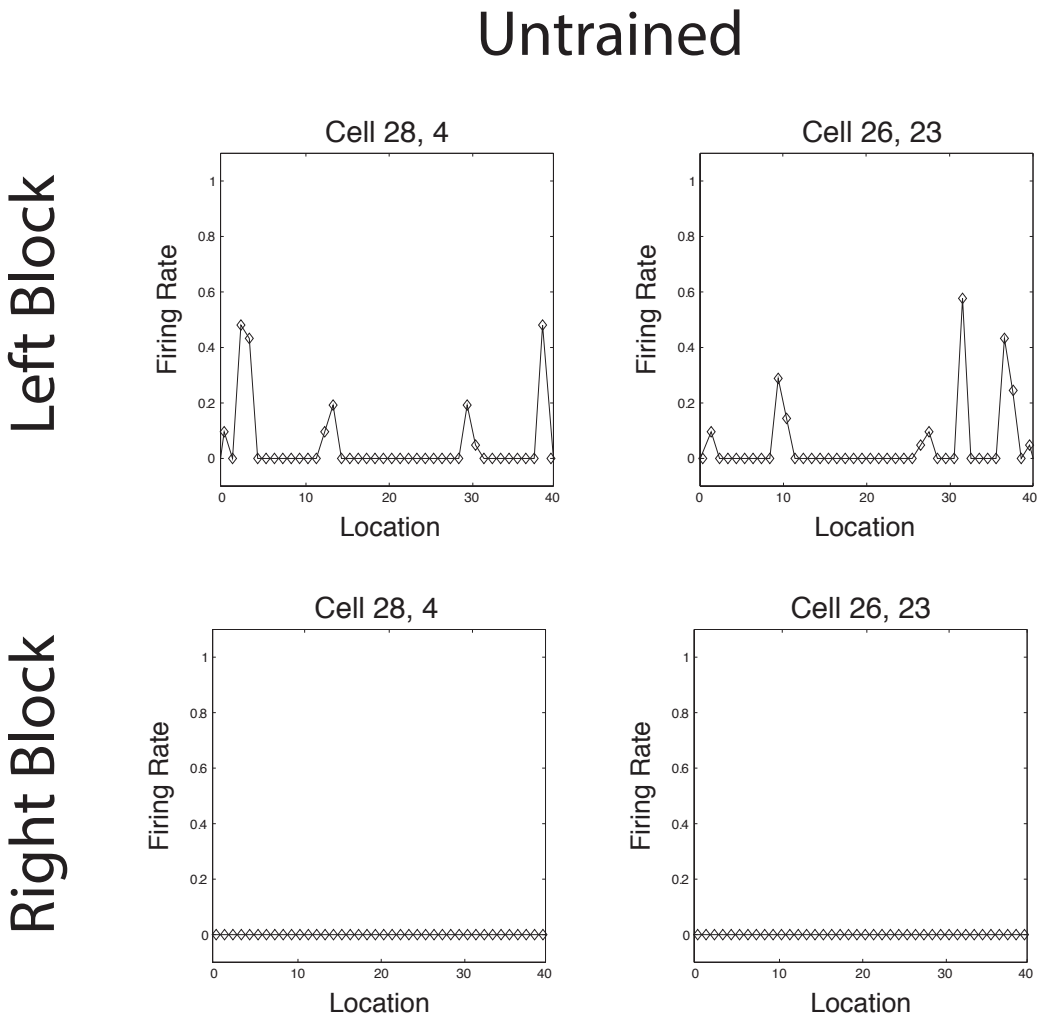

Figure 4.3: Untrained cell response plots from the 2D block pilot study. This figure shows the cell response plots before training for prototypical Cell 28, 4 and Cell 26, 23. Each column presents the same specific prototypical cell, while each row presents responses to either the left block (first row) or the right block (second row). It can be seen that both cells respond sporadically to different views of the left block, without preference. Both cells fail to respond to any of the views of the right block.

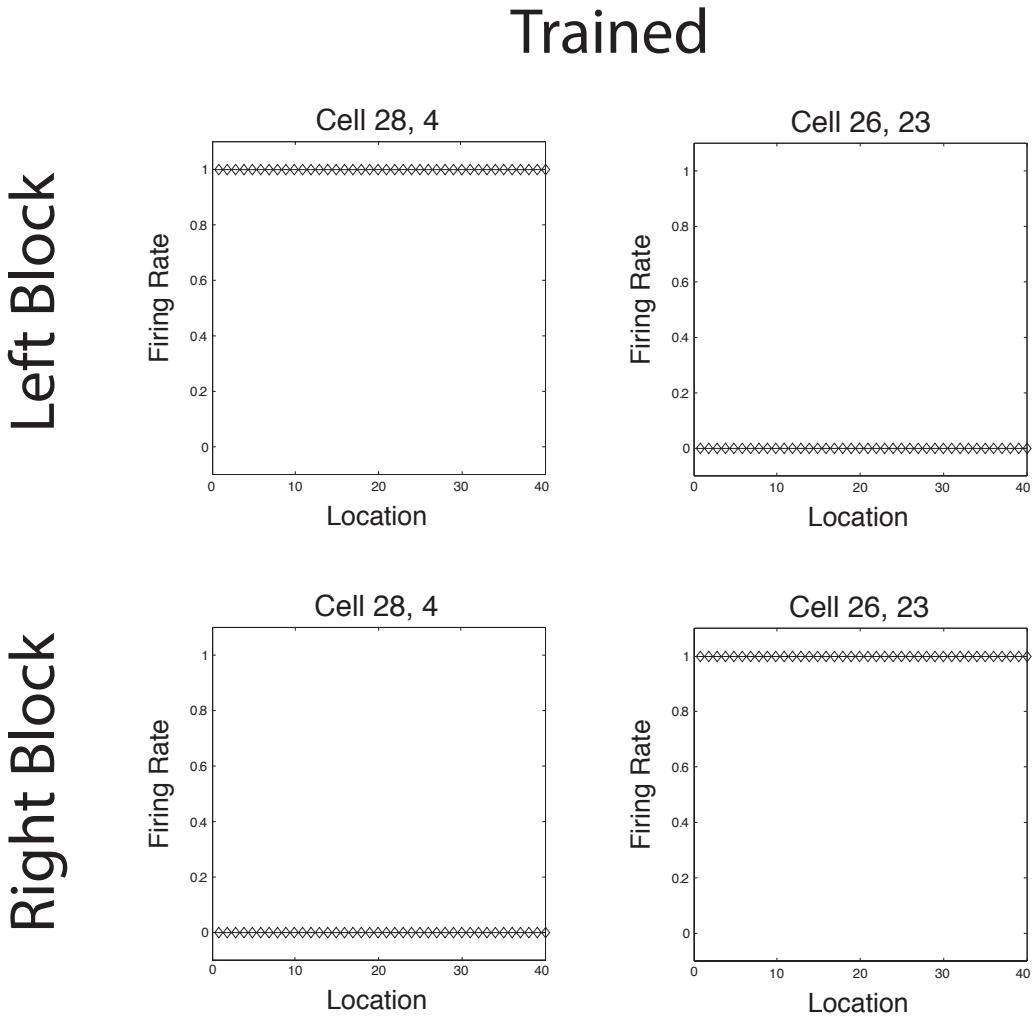

Figure 4.4: Trained cell response plots from the 2D block pilot study. This figure shows the cell response plots after training for prototypical Cell 28, 4 and Cell 26, 23. Cell 28, 4 has learned to respond to the left 2D block while Cell 26, 23 has learned to respond to the right 2D block. In both cases, each cell only responds to the preferred block stimuli and does not respond to any of the other views of the other stimuli.

ings: Cube and Hectoid; Cube and Dodecahedron; Dodecahedron and Hectoid. Each pairing was presented as a separate experiment. In each experiment the objects within each pairing either rotated dependently together in lock-step during training or rotated independently during training. The Cube and the Hectoid experiment was performed using both a SOM architecture and a competitive network architecture. For the remaining two pairings, the network's performance was only explored using a SOM architecture. The three different object pairs were used to ensure that the results were consistently robust.

The following results show that the network was unable to form separate representations of the individual objects when trained together in the dependent rotation condition. That is, when the objects rotate together in lock-step and there is no independent rotation. This was the case for both the SOM and the competitive network architectures. However, in the independent rotation condition, the network was able to form object selective representations for each of the objects, despite the fact that the objects were always presented rotating together during training. The effect of the SOM compared to the competitive network is explored in the context of this novel result.

4.3.1 Main Results With SOM Architecture

Figure 4.5 shows cell response polar plots of an 8x8 subset of the cells in the output layer after training using a SOM network architecture with the Cube and the Hectoid. Both the dependent and independent rotation paradigms are presented. The figure is split into four quadrants forming two rows and two columns. Each quadrant comprises a subset of neurons from the model's output layer where 8x8 neurons are represented by 64 polar plots. Each degree in the polar plot represents an associated view of the test object. Firing rates are represented by the distance from the centre point. The left hand column comprises two plots that show the results of testing with the Cube, and the right hand column

comprises two plots that show results of testing with the Hectoid. The top row of plots comprises results from the dependent rotation training paradigm, and the bottom row comprises results from the independent rotation training condition. Within each paradigm (dependent or independent rotation), the same cells are plotted when testing on both the Cube and the Hectoid.

It is clear that in the dependent rotation condition, cell response profiles exhibit a large degree of overlap between the two objects. That is, cells that respond invariantly to the Hectoid also respond in kind to the Cube, as if they were the same object. The network has not formed object selective representations, failing to build separate representations for the Cube and the Hectoid. However, in the independent rotation paradigm, most individual output cells have learned to respond invariantly to either the Cube or the Hectoid (but not both), despite the fact that the objects were always presented together during training. Crucially, the network has formed object selective representations for each object where cells respond to their preferred object and not to the alternative object.

Figure 4.6 and Figure 4.7 show the same results but for the Cube and the Dodecahedron, and the Dodecahedron and the Hectoid, respectively. It can clearly be seen that the response plots reflect the results found with the Cube and the Hectoid. That is, in the dependent rotation condition VisNet fails to develop object selective representations, while in the case of the independent rotation, cells in the output layer learn to respond exclusively to their preferred object. In the case of the Cube and the Dodecahedron, cells that respond preferentially to the Cube did not develop complete invariance and instead there is another cluster of cells that respond to the remaining views (not shown).

Information analysis plots are provided for when the SOM network was trained with object pairs rotating independently and rotating dependently. Information results with the Cube-Hectoid, Cube-Dodecahedron and Dodecahedron-Hectoid pairings are presented in Figure 4.8, Figure 4.9 and Fig-

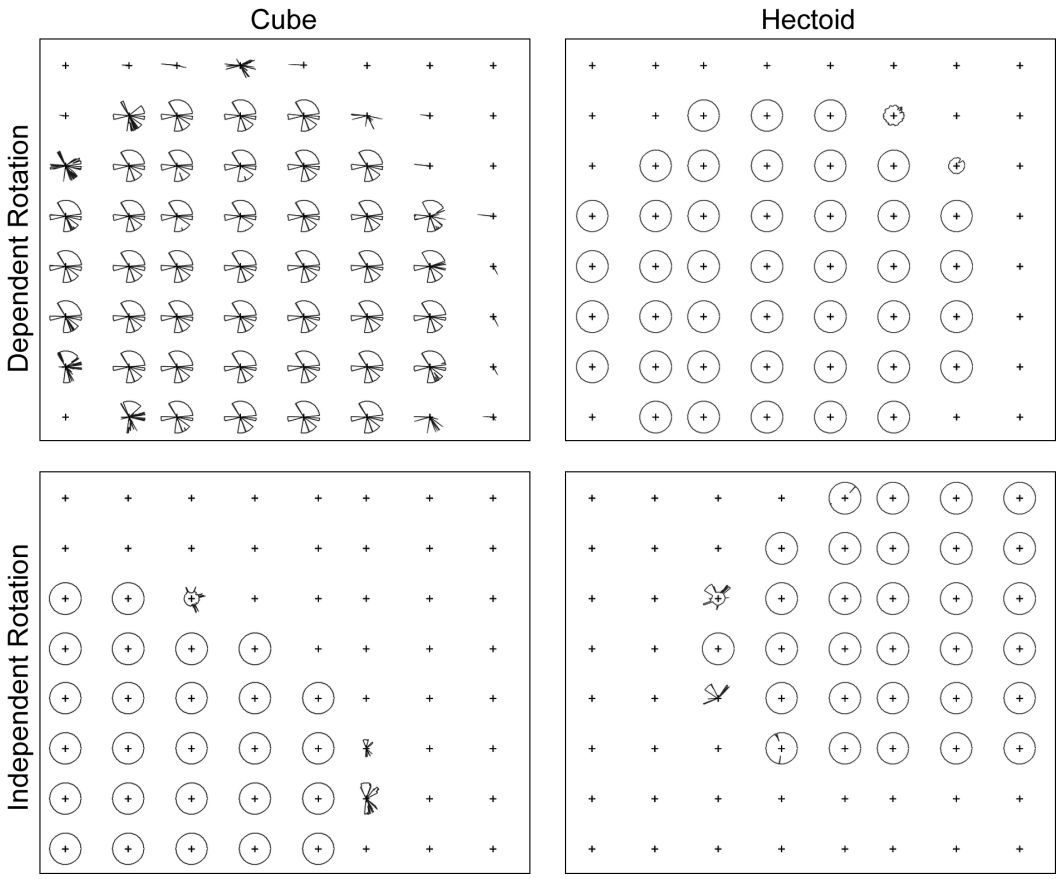

Figure 4.5: Cube-Hectoid pairing results with SOM network architecture. The figure shows the responses of a 8x8 subset of output layer cells after training with the dependent and independent rotation paradigms and testing with the Cube and Hectoid presented individually. The figure is divided into four quadrants. The left-hand column presents two plots comprising the cell responses when tested on the Cube and the right-hand column presents two plots comprising cell responses when tested with the Hectoid. The top row presents results after the dependent rotation training paradigm, and the bottom row presents results after the independent rotation training paradigm. Crucially, for the purposes of comparison, the same set of cells are shown within each training paradigm (rows) when tested with the Cube or the Hectoid. In the dependent rotation paradigm, cells learn to respond to both objects as if they were one with almost complete view invariance, and fail to build object specific responses. In the independent rotation paradigm, object selective responses form, that are also view invariant. Specifically, when the Cube and the Hectoid rotate independently during training, the cells in the lower left corner of the 8x8 array learn to respond to the Cube and the cells in the top right corner learn to respond to the Hectoid.

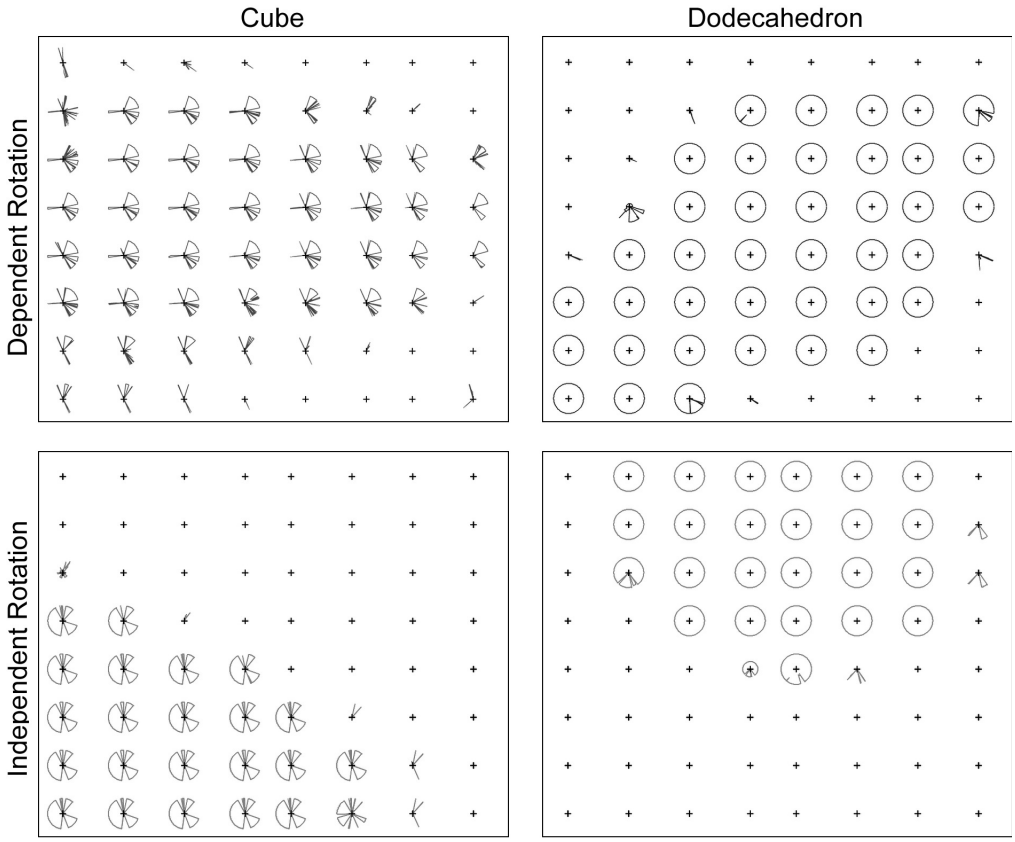

Figure 4.6: Cube-Dodecahedron pairing results with SOM network architecture. Conventions as in Figure 4.5. This figure shows that in the case of dependent rotation, the network has failed to build object selective responses while in the independent rotation case, the network has learned to develop object selective responses. Although the cells responding preferentially to the Cube do so for the majority of its views, they do not represent all 360 views. Instead, a separate pool of cells (not shown in this figure) respond to the remaining views.

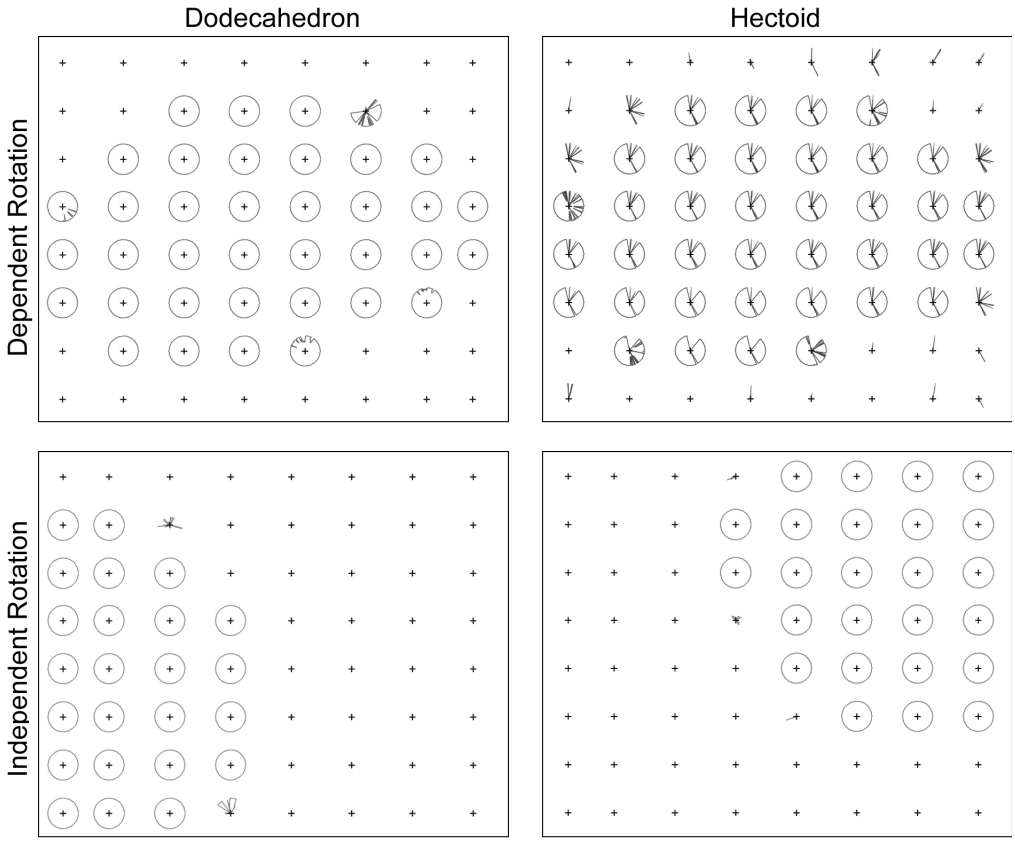

Figure 4.7: Dodecahedron-Hectoid pairing results with SOM network architecture. Conventions as in Figure 4.5. This figure shows that in the case of dependent rotation, the network has failed to build object selective responses while in the independent rotation case, the network has learned to develop object selective responses.

ure 4.10, respectively.

In all three figures, the left-hand plot shows the single cell information analysis. This was calculated to confirm whether the network had developed neurons that responded exclusively to their preferred object over all 360 views. In the case of the Cube-Hectoid pairing, it was observed that in the case of independent rotation, 72 neurons provided maximal information of 1 bit. Conversely, in the case of dependent rotation, the amount of information is much lower. Similar results are presented for the other two pairings, although in the case of the Cube-Dodecahedron pairing (Figure 4.9), fewer cells provide maximal information in the independent rotation condition because there were fewer completely invariant responses for the Cube. This can be observed in Figure 4.6.

In summary, the single cell information analysis plots confirm that independent rotation is sufficient to cause cells to develop object selective responses that are also invariant. Crucially, in the dependent rotation paradigm, cells are unable to build object selective representations, and fail to discriminate between the objects due to the statistical coupling between them present in the training input.

For all three Figures 4.8, 4.9 and 4.10, the right-hand plot shows multiple cell information analysis measures. In the independent rotation paradigm, a very small number of cells are required to reach the maximum information of 1 bit, confirming that the network learns to represent both objects used during training. In the dependent rotation condition more cells are required and maximum information is never achieved because cells in the network have failed to separate the two objects used during training.

In summary, the network produces markedly different responses in the independent rotation paradigm when compared to the dependent paradigm. In the dependent rotation condition, the network learns

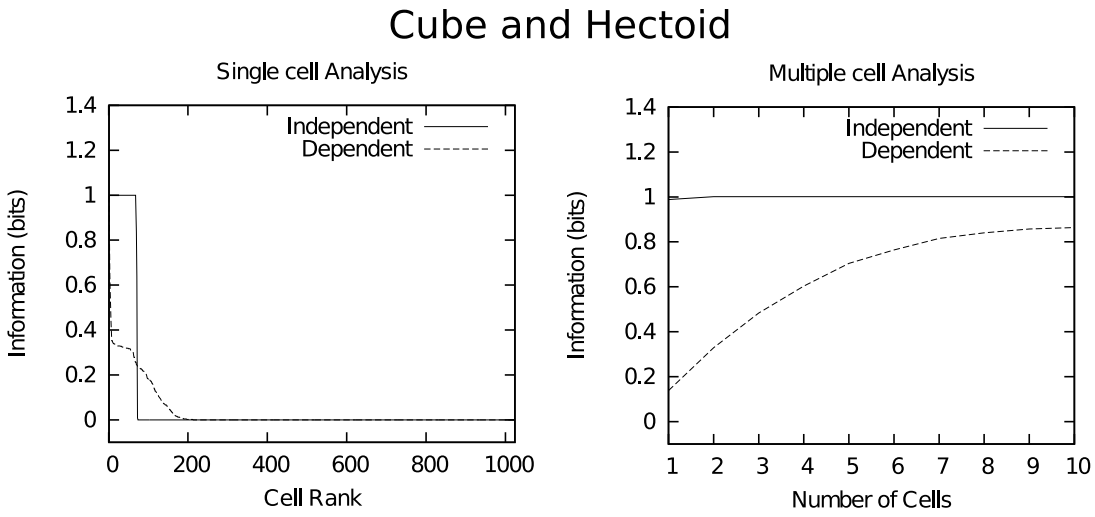

Figure 4.8: Cube-Hectoid pairing information analysis with the SOM network architecture. Single (left-hand plot) and multiple (right-hand plot) cell information results obtained when the trained SOM network was tested with the Cube and Hectoid rotating separately. Results are presented for both the independent (unbroken line) and dependent rotation (dashed line) paradigms. The single cell information measure for all 4th layer neurons ranked in order of the amount of information they convey about the objects is shown. In the independent rotation paradigm, the network develops many object selective and invariant neurons attaining the maximum level of single cell information of 1 bit. The dependent rotation condition produces less information and no cells reached the maximal information. The multiple cell information measure reaches the maximum level of 1 bit in the independent condition. This confirms that the network represents both trained objects and maximal information is obtained. The dependent condition provides a comparison, where the network has failed to separate the two objects presented during training.

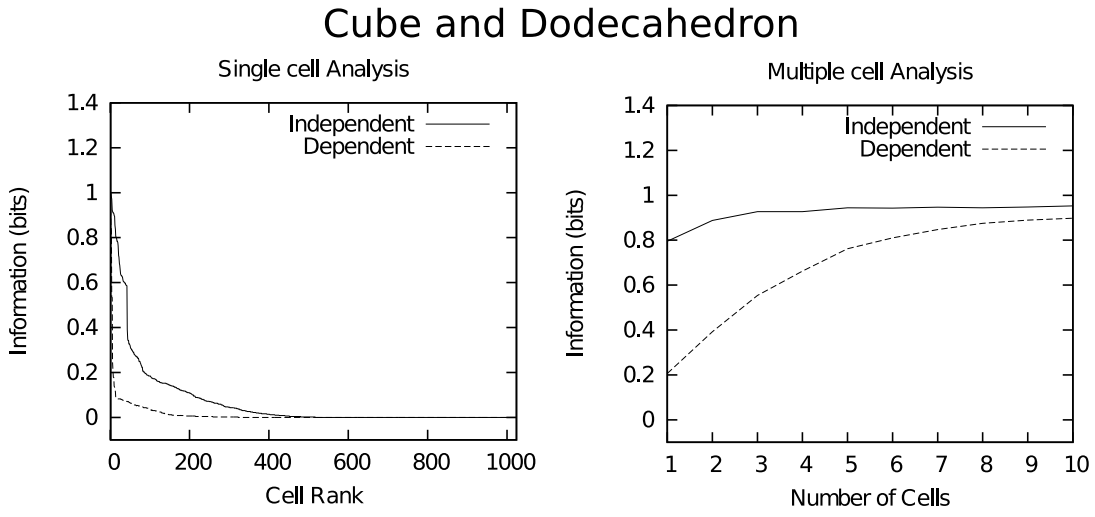

Figure 4.9: Cube-Dodecahedron pairing information analysis with the SOM network architecture. It may be observed that after independent rotation during training, the single cell information is relatively low. This is because cells did not learn to respond to all 360 views of the Cube and were less invariant.

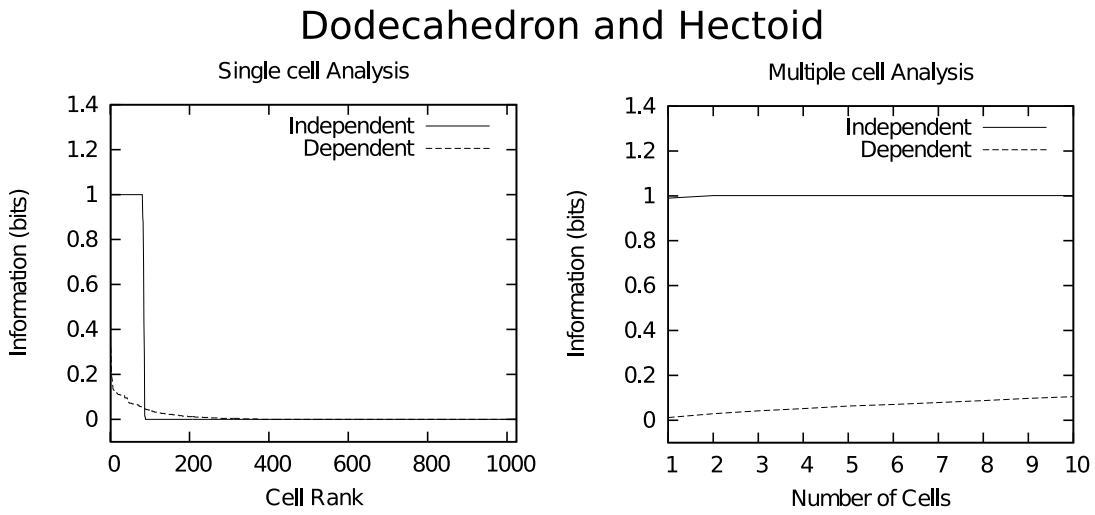

Figure 4.10: Dodecahedron-Hectoid pairing information analysis with the SOM network architecture. In the independent condition, maximal information is obtained in both the single cell and multiple cell analyses. In the dependent rotation condition, the network performs badly and is unable to build object selective representations. This results in very little information in both the single cell and multiple cell analyses.

to build combined representations for both objects after it had been trained. That is, cells learn to respond to both objects, such as the Cube and the Hectoid, with a high degree of invariance, but with no object selectivity. Crucially, in the independent rotation condition, the network forms invariant and object selective representations for both objects. This novel result is next explored in the context of a competitive network.

4.3.2 Competitive Network trained with Cube and Hectoid

Figure 4.11 shows four plots each consisting of an 8x8 sample of cells in the model's output layer after training using a competitive network architecture. Each set of polar plots shows the responses of the output layer neurons when tested with the Cube and the Hectoid rotating individually, for both the dependent and independent training paradigms. Similar to Figure 4.5, the left hand column shows results when tested with the Cube and the right hand column shows results when tested with the Hectoid. The top row presents results after training with the dependent rotation paradigm and the

bottom row presents results after training with the independent rotation paradigm.

In the case of the dependent rotation condition, the network was trained with both objects rotating together in lock-step and tested with each object separately. The results show that the cells learn to respond invariantly to both objects but fail to build object selective representations. That is, cells respond to both the Cube and the Hectoid, regardless of viewing angle. There are no cells that fire exclusively to just one of the objects.

Results from the independent rotation condition show a markedly different profile to that of the dependent rotation condition. In the independent rotation condition, the network has developed cells that exclusively respond to contiguous portions of the view-space of just their preferred object. That is, cells do not respond to more than one object. These cells do not respond completely invariantly to their preferred object and instead represent a portion of the total view-space. All views of the preferred object are encoded in this manner, such that a group of cells together are able to exclusively identify a specific object from any viewing angle. Both the Cube and the Hectoid are fully represented in this way.

Information analysis measures were calculated after training and testing using the competitive network architecture to help quantify the network's performance. Figure 4.12 shows the single and multiple cell information analysis plots when the network was trained with the Cube and Hectoid rotating independently and rotating dependently. In the case of independent rotation, the information analysis measures provide low levels of information, because the competitive network did not build invariant representations. The left-hand plot shows the single cell information analysis and it may be observed that although the independent rotation paradigm provides more information than the dependent condition, it does not reach the maximum information of 1 bit because cells do not respond

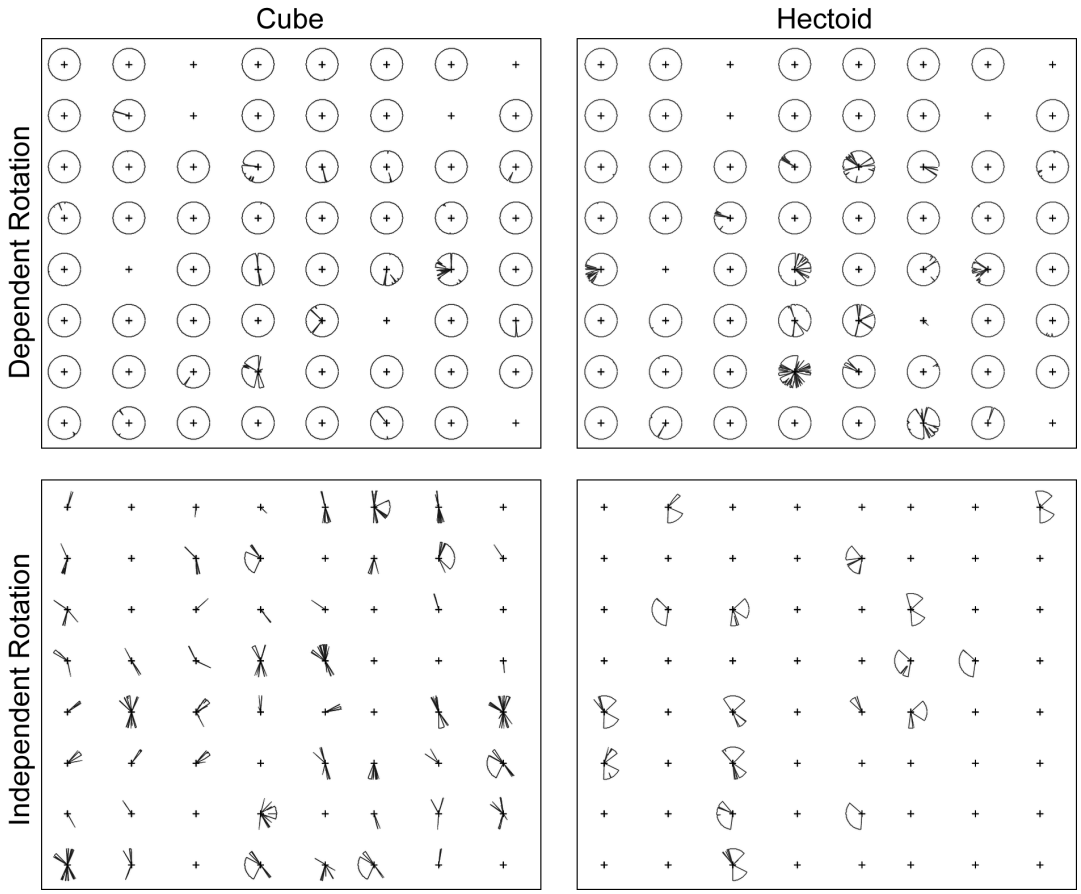

Figure 4.11: Results with competitive network architecture. The figure shows the responses of a 8x8 subset of output cells after training with the dependent and independent rotation paradigms and testing with the Cube and Hectoid individually. Conventions as for Figure 4.5. In the dependent rotation paradigm, cells learn to respond to both objects as if they were one, with complete view invariance to all 360 views, and fail to build object specific responses. In the independent rotation paradigm, object selective responses form. However, cells do not develop complete view invariance, primarily responding to contiguous views of their preferred object.

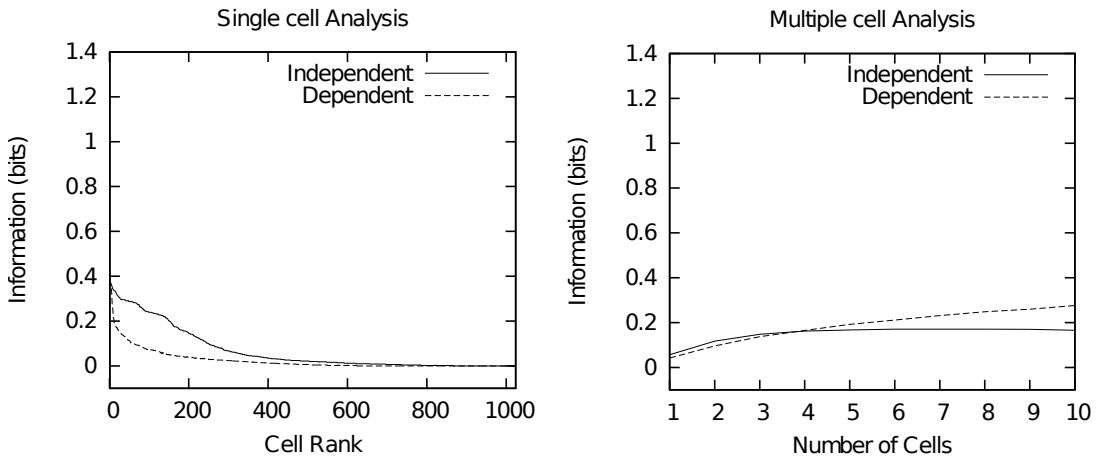

Figure 4.12: Information analysis with the competitive network architecture. Single (left-hand plot) and multiple (right-hand plot) cell information results obtained when the trained competitive network was tested with the Cube and Hectoid rotating separately. Results are presented for both the independent (unbroken line) and dependent rotation (dashed line) paradigms. The single cell information measure for all 4th layer neurons ranked in order of the amount of information they convey about the objects is shown. Although in the independent rotation paradigm the network was able to form object selective responses, both training conditions produced lower levels of information for both the single and multiple cell information analysis. This can be attributed to the lack of object selectivity in the dependent rotation paradigm, and the lack of invariance achieved in the independent rotation paradigm.

invariantly to their preferred object. The multiple cell information analysis also produces low levels of information. Similarly, low levels of single and multiple cell information were also recorded for the dependent motion condition. Maximum information is achieved when cells respond invariantly and exclusively to their preferred object. As such, low levels of information can be attributed to the lack of object selectivity in the dependent rotation paradigm, and the lack of invariance achieved in the independent rotation paradigm.

4.3.3 The Effect of the SOM Width

The effect of the SOM width was explored by varying the radii of the excitatory and inhibitory components of the SOM filter. To investigate the effect the SOM filter width has on the learning and self-organisation of the network, a constant ratio was maintained between the excitatory and inhibitory

components. A range of SOM widths was explored. Figure 4.13 presents the firing responses of output layer cells, after training the network with the Cube and Hectoid rotating independently during training. In particular, the figure shows results with narrow and wide SOM filters alongside the original, intermediate, SOM width results, initially presented in Figure 4.5. The values for the intermediate SOM width are presented in Table 4.3 and are three times as wide as the 'narrow' SOM width. The 'wide' SOM width is five times as wide as the narrow SOM width. Corresponding information results are shown in Figure 4.14.

These results show that when the SOM width is relatively narrow, the network builds invariant responses to one object, but not the other. However, as the width of the SOM increases so does the number of object selective invariant neurons for both objects. When the SOM width becomes too wide, output cells respond invariantly to all views of both objects at the same time. That is, they are not object selective. Possible causes for these results are mentioned in the Discussion, Section 4.4.

4.4 Discussion

Previous results from VisNet have shown that separate object representations can develop in the output layer when there is a sufficient level of decoupling between a large number of individual objects presented as random pairs during training. Crucially, there must be a large enough superset of objects from which to select when creating the different object pairings (Stringer et al., 2007; Stringer and Rolls, 2008). The aim of the research in this chapter was to establish whether independent rotation might play an important role for helping the primate visual system to build separate representations of just two objects always presented together. The differences in performance between a self-organising map architecture and competitive network were also investigated.

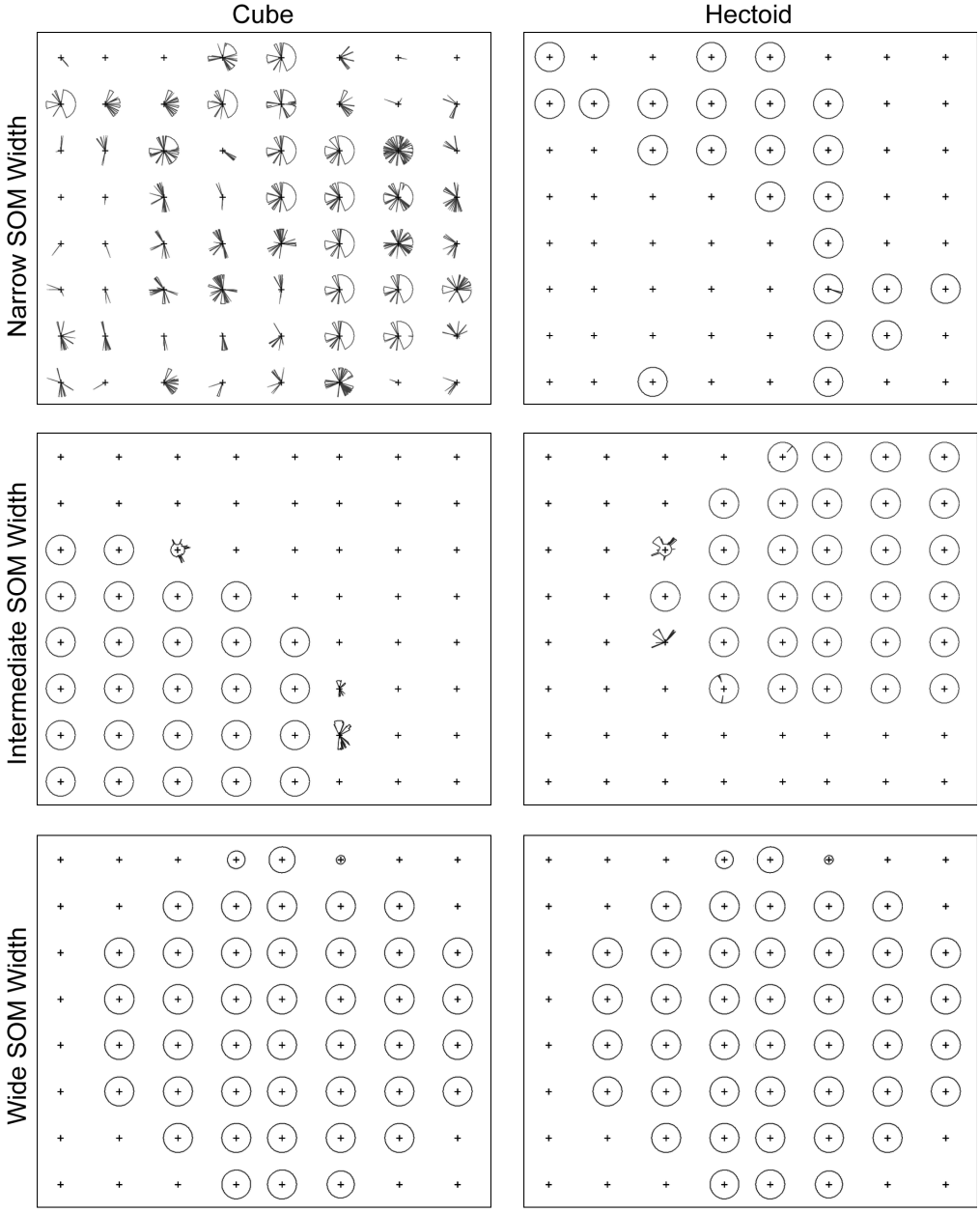

Figure 4.13: Polar plots results with three different SOM widths. Cube-Hectoid pairing results after independent rotation during training with three different SOM widths: narrow, intermediate and wide. Conventions as in Figure 4.5. When the SOM width is relatively narrow, fully invariant cells develop for the Hectoid but not for the Cube. However, as the SOM width is increased, the number of invariant cells also increases. With an intermediate SOM width, object selective and fully invariant cells form for both the Cube and the Hectoid. When the SOM width is relatively wide, the output cells learn to represent both objects invariantly but fail to produce object selective responses. The same set of cells are shown for each experiment.

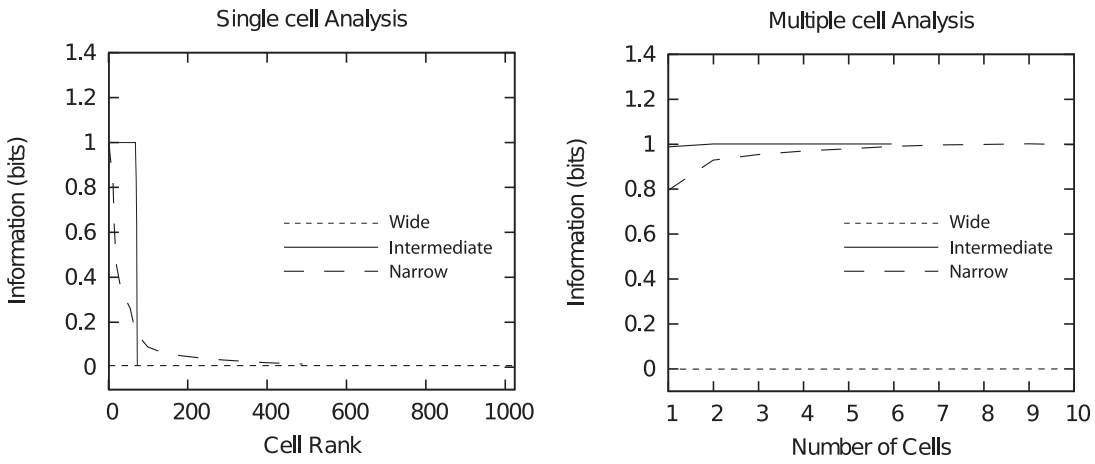

Figure 4.14: Information analysis results with three different SOM widths. Information analysis of Cube-Hectoid pairing results after independent rotation during training, with the SOM architecture with three different widths. Single (left-hand plot) and multiple (right-hand plot) cell information results are shown for the narrow (dashed line), intermediate (unbroken line) and wide (dotted line) SOM widths when tested with the Cube and Hectoid rotating separately after training. A small amount of information is available in the narrow SOM width condition, while maximal information is achieved in the intermediate SOM width condition. Zero information is obtained when the SOM is wide because cells learn to respond in an identical manner to both the Cube and the Hectoid.

A pilot study was performed to establish whether VisNet is able to use independent translation across the retina in order to build separate representations of two individual stimuli, despite the fact that they are always presented together on the retina in pairs. The results show that VisNet was successful at building translation invariant representations of the two individual block stimuli presented during training. For the pilot study, a translation paradigm was chosen over a rotation paradigm because it enabled more control over the training stimuli both in terms of overlap between transforms, but also with respect to the length of the stimuli contour boundaries. In principle, a translation training paradigm and a rotational training paradigm should involve similar learning mechanisms during training. The most significant difference is the precise control of the amount of overlap between the different transforms of each stimuli in the case of the translating block stimuli, which is important for CT learning. The pilot study thus provided a basis for the following results with rotating objects.

The results presented confirm that the network is able to successfully form object selective repre-

sentations when trained with objects that rotate independently. This applied to both a SOM architecture and a competitive network architecture.

In the independent rotation paradigm, different views of one object, for example, the Cube, were presented with different views of the Hectoid. Because the features that comprise a specific view of an object are always presented together 100% of the time, they are maximally correlated. Conversely, these features are not well correlated to any of the features of the alternative object. That is, during training, a particular view of an object was presented with ten different views of the alternative object. Particular pairs of views are seen together relatively infrequently. The decoupling between the different views of the two objects is sufficient for a particular output cell to strengthen its connection with a specific view of the Cube, without necessarily strengthening connections associated with a particular view of the Hectoid.

When the objects rotate together in lock-step, both network architectures failed to build representations specific to either object. Instead, VisNet built a single representation, as if both objects were in fact one object. This result is a consequence of the statistical coupling between the objects that occurs due to the dependent motion. With dependent motion, each view of the Cube is always presented with the same view of the Hectoid. If a particular view of the Cube happens to strongly activate a given output neuron due to the initially random feedforward synaptic weights, then not only will the synaptic weights from the Cube view strengthen, but the weights from the Hectoid view will too. After multiple training epochs, the output neuron will learn to respond equally well to views of both objects as if they are one.

The CT learning effect is able to bind together views that share a high degree of spatial overlap and helps form view invariant cells. View invariant cells were only observed in the case of the SOM

architecture. The competitive network architecture produced contiguous blocks of activation over a limited range of object views spanning about 45 °. A network exhibiting these types of cell responses requires a population of cells to encode the whole view-space. In the case of the SOM architecture paradigm, there were mutually excitatory lateral connections that cause cells with similar response properties to form close together. For example, assume that Cell A responds to a contiguous range of views of an object, from 90 ° to 135 °. Assume that Cell B responds preferentially to a range of views of the same object spanning 100 ° to 145 °. These two cells will form close together and as training progresses, cell A will cause cell B to become active when view 90 is presented during training, and similarly, cell B will cause cell A to become active when view 145 is present during training. The synaptic connections will strengthen according to Hebbian learning and this may have an effect of broadening the tuning profiles of both cells until the network reaches a trained state where both cells respond invariantly over the whole range of views.

The sparseness of the network defined in Equation 1.6 was identified as one of the key parameters for developing object selective invariant representations with the SOM network architecture. When the sparseness value was raised to high levels it promoted excessive broadening of the cell response profiles. Cells began to respond to both objects with increasing invariance. It was observed that a sparseness value of 0.04 creates broad response profiles that remain object selective. At this sparseness value, the SOM filter is able to distribute neurons successfully, therefore developing functional maps where a single cell response is similar to the average response of its neighbours. The network was robust to different learning rates.

The effect of different SOM filter widths was also explored. The primary effect of the SOM is to cause cells with similar response properties to develop close together in the output layer and cells

with different responses to form in separate clusters. Adjacent cells influence one another reciprocally, gradually becoming active to the particular views that their neighbours respond to best. This has the general effect of amplifying the Hebbian learning mechanism and broadens the cells' response profiles so that they learn to respond to an increasing number of similar views. Even with the narrow SOM, this effect still helps build invariant responses in the output layer relative to a competitive network with comparable parameters. As the SOM increases in width, the network learns to build increasingly invariant responses. However, when the SOM is very wide it forces the output cells to respond to all views of both objects due to overwhelming levels of mutual excitation between cells.

In summary, a mechanism has been presented for how the visual system could self-organise when presented with complex natural scenes, using only independent movement as a method to build view invariant, object specific, representations. Early research with VisNet focussed on presenting just one object at a time (Stringer and Rolls, 2000). More recent research required a large superset of a number of different objects from which to present random object pairings. For the first time, this research shows that it is possible for the network to develop completely view invariant, object selective, responses when trained with only two objects. Crucially, the objects must move independently of one another. A SOM network architecture was implemented for the first time within each layer of the VisNet model, and the result has shown that this causes cells with similar response profiles to form close together and develop completely view invariant responses. This observation is documented *in vivo*, where neurons in IT cortex responding preferentially to similar features, can be grouped together (Tsunoda et al., 2001; Kiani et al., 2007)

Chapter 5

The Formation of Separate Representations of Facial Identity and Expression

Results presented in the previous chapter show that the VisNet model is able to use independent motion to form invariant representations of individual objects, despite the fact that a pair of objects was always presented during training. More generally, this is an important result because it confirms that output neurons can use statistical decoupling as a method for developing separate representations of individual objects, or feature spaces.

Having demonstrated how the visual system might respond when presented with a pair of objects that move independently of each other, this chapter applies the same principles to learning about facial expression and identity. As will become clear, it is no longer appropriate for VisNet output neurons to respond invariantly. Instead, output neurons must learn to respond to a small subset of views that

form part of their preferred view space, and must not respond to views from another view space.

5.1 Introduction

Single unit recording studies in non-human primates have revealed that a number of the visual processing areas of the brain appear to encode facial identity and facial expression across separate subpopulations of neurons. For example, it has been shown that the inferior temporal gyrus (TE) contains cells that are primarily responsive to facial identity, the adjacent superior temporal sulcus (STS) contains cells that primarily respond to facial expression, and the cells on the lip of the sulcus (TEm) tend to respond to expression and identity (Hasselmo et al., 1989a). Cells responsive to facial identity are primarily found in inferior temporal cortex, while cells that respond to dynamic facial features such as facial expression are found in STS (Perrett et al., 1992). Orbitofrontal cortex (OFC) of non-human primates contains some cells that respond exclusively to changes in facial identity, while other cells respond exclusively to facial expression (Rolls et al., 2006). Similar cells have been found in the amygdala of non-human primates, which respond to either facial identity or facial expression (Gothard et al., 2007).

Further evidence of physically separate visual representations of facial identity and expression comes from fMRI adaptation (fMRIa) studies in humans. Using fMRIa, functional dissociations within the STS have been demonstrated (Winston et al., 2004). Specifically, cells in lateral right fusiform cortex and posterior STS were released from adaptation upon changes to facial identity, while cells in more anterior STS were released from adaptation upon changes to facial expression. These findings are consistent with other neuroimaging studies (e.g. Fox et al., 2009; Cohen Kadosh et al., 2010; Xu and Biederman, 2010).

How might the primate visual system develop physically separate representations of facial identity and expression given that during learning the visual system is always exposed to simultaneous combinations of facial identity and expression? Previous research has shown that principal component analysis can extract and categorise facial cues related to facial identity and expression (Calder et al., 2001; Gary et al., 2002). For a review see Calder and Young (2005). However, these computational methods are not based on biologically plausible models of brain function. In this chapter, it is shown for the first time how separate visual representations of facial identity and expression could develop in a biologically plausible neural network architecture using associative Hebbian learning.

In the simulations described below, images of faces with different identities and expressions are shown to a neural network model of the ventral visual pathway, VisNet (Wallis et al., 1993; Stringer and Rolls, 2000, 2002; Stringer et al., 2006, 2007; Stringer and Rolls, 2008). The version of the VisNet architecture used consists of a feedforward series of four self-organising maps (SOMs). As outlined in previous chapters, during learning the feedforward synaptic weights are updated by associative Hebbian learning. A key aspect of the model in terms of biological plausibility is that learning is unsupervised, that is, the network is not explicitly told the identity or expression of the current face during training.

A pilot study is first presented, exploring a limitation of the VisNet model. This study builds on the results presented within Chapter 4 and helps guide the direction of the main experiments where the VisNet network is trained with carefully controlled cartoon faces. These faces convey information about both facial identity and expression simultaneously. Finally, a follow-up study is presented in which life-like computer generated 3D faces are presented to the network, also conveying information about both facial identity and expression simultaneously.

The first pilot study explores the extent to which VisNet is able to develop separate representations of different features as they vary with respect to one another during training. More specifically, it builds on the pilot study of Chapter 4. A human face has many muscles and is capable of producing a huge range of expressions. Furthermore, even the most distant of acquaintances can be identified by their face. Accordingly, the first pilot study investigated whether VisNet is able to build separate representations of more than two orthogonal spaces during learning. The results of this study influence the experimental design in the main experiments. The second pilot study explores the extent to which the current version of VisNet is able to distinguish between the features that comprise more detailed stimuli. Faces with varying levels of detail are presented to the VisNet model during learning and the results dictate the stimuli design of the main experiments.

In the main experiments, the network is trained with carefully controlled cartoon faces, which convey information about both facial identity and expression concurrently. The face images comprise two types of continuously varying facial features. The eyes and nose convey where the face lies within a uni-dimensional continuum of identities, while the mouth and eyebrows convey where the face lies within a one-dimensional continuum of expressions. After training, the output layer of the network has developed separate clusters of cells that respond exclusively to either facial identity or facial expression. Individual neurons that learn to encode identity fire selectively to a small region of the space of identities regardless of expression. Similarly, individual neurons that learn to represent expression fire selectively to a small region of the space of expressions regardless of identity. Performance of the network is interpreted in terms of the learning properties of SOMs.

A more challenging task is presented as part of the follow-up study. The network is trained with life-like computer generated human faces where the facial expression and identity covary with

respect to one another. However, unlike the main experiments previously discussed, the expression and identity are no longer orthogonal visual spaces. That is, expression is encoded by some of the same features that encode identity, and vice versa. Although these facial stimuli are still carefully controlled and manipulated programmatically, they comprise a more ecologically valid training set. These experiments are conducted in recognition of the limitations of the main experiments, the results of which form the basis for some interesting discussion points.

How might the primate brain form separate visual representations of facial identity and expression?

Recent research has shown how VisNet can exploit the statistics of natural visual input in order to learn separate visual representations of objects, despite the fact that multiple objects were always presented together during training (Stringer et al., 2007; Stringer and Rolls, 2008). Furthermore, the results presented within Chapter 2 emphasise this in particular. During presentation, the features that make up an object occur together more frequently when compared to the features that make up different objects. The frequency with which the objects are presented together during training denotes the level of correlation between features from different objects. However, the features that comprise individual objects are always seen together and so are more highly correlated. In this situation, a competitive network will form representations of individual objects, rather than the combinations of objects seen during training. The same effects should also be present in SOMs.

In the main experiments that comprise this research, it is shown how similar mechanisms, which exploit the visual statistics of facial identity and expression, may cause VisNet to form separate representations of identity and expression during training with complete faces. Facial identity and expression are each modelled as separate spaces rather than discrete objects. An example is shown in

Figure 5.3, which shows a selection of cartoon faces composed from particular combinations of 40 identities and 40 expressions. In this simplified test case, identity and expression are represented by different facial features. Specifically, identity is represented by the shape and location of the eyes and nose, while expression is represented by the mouth and eyebrows. The space of identities varies along the horizontal dimension, while the space of expressions varies along the vertical dimension. Each space can be varied independently of the other. The choice to use these cartoon faces has several justifications. Firstly, the nameable features that comprise a face (eyes, nose, etc.) correspond to the brightness discontinuities of real faces. Secondly, the cartoon face stimuli are clearly perceived as faces by human viewers and single-unit recording has shown that face selective neurons in the macaque monkey respond to real faces and cartoon faces comparably (Freiwald et al., 2009). Thirdly, the individual features of a cartoon face can be carefully controlled. Finally, relatively few parameters are required to completely specify a cartoon face when compared to the individual pixel values of images of real faces. Together, these reasons suggest that cartoon faces can provide a very useful method to simplify the space of faces.

During training, all 1600 possible faces are presented to VisNet. During presentation, the features that make up a particular identity always occur together. The Hebbian learning rule will encourage output neurons to learn to respond to input features that tend to occur together simultaneously during training. However, the features that make up that identity are paired with each of the different expressions on different occasions. Therefore, the features that make up the identity occur together far more frequently than they occur with the features that make up any particular expression. This statistical decoupling between the particular identity and any of the expressions prevents an output neuron from learning to respond to a combination of the identity and any one of the expressions. In this situation,

a SOM will form representations of individual identities, rather than the combinations of identity and expression seen during training.

Similarly, the features that make up a particular expression always occur together. Moreover, the features that make up that expression are paired with each of the different identities on different occasions. Therefore, the features that make up the expression occur together far more frequently than they occur with the features that make up any particular identity. This prevents an output neuron from learning to respond to a combination of the expression and any one of the identities. A SOM will similarly form representations of individual expressions.

The effects of a SOM are investigated by introducing short-range excitatory connections and long-range inhibitory connections (von der Malsburg, 1973; Kohonen, 1982) to the original VisNet architecture. This lateral connectivity of the SOM uses a 'Mexican-hat' profile and encourages spatially proximal neurons to develop similar response properties due to their mutual excitation while neurons that are relatively far apart will experience mutual inhibition. This drives competition to create separate pools of functionally distinct neurons, which encode either the space of identities or the space of expressions.

The aim of the research has been to establish how this process might operate with idealised cartoon faces, in which facial identity and expression are one-dimensional spaces and represented by non-overlapping sets of visual features. It is proposed that, as long as there is sufficient statistical decoupling between facial identity and expression, then a similar mechanism will operate with more realistic face stimuli, in which the dimensions of identity and expression are more complex and the two spaces are represented by a common set of overlapping visual features.

All computer simulations are carried out using VisNet. As discussed above, experimental studies

in non-human primates have shown a dissociation between the representations of facial identity and expression across a number of visually-responsive brain areas both within and outside the ventral visual pathway. However, it is proposed that the learning mechanisms demonstrated in the model below, in which SOMs are able to exploit the statistical decoupling between facial identity and expression, may still operate across a more complex network of visually-responsive areas in the brain.

5.2 Model Architecture and Parameters

All the experiments presented as part of this chapter were conducted using the VisNet model architecture as described in Section 1.3.3. However, this research implements a SOM (von der Malsburg, 1973; Kohonen, 1982) within each layer in place of the original competitive network architecture. That is, in the model simulations, the local graded inhibition within a layer was adjusted to incorporate short-range excitation and longer-range inhibition. Accordingly, this VisNet model consists of a hierarchical series of four layers of SOMs, corresponding to V2, V4, the posterior inferior temporal cortex, and the anterior inferior temporal cortex. Images were filtered in accordance with the difference of Gaussian Equation 1.1 previously presented. The sparseness percentile and slope values are presented in Table 5.1. The lateral inhibition and excitation parameters are provided in Table 5.2. A range of learning rates were explored, ranging from 0.0001 through to 0.1. Unless otherwise stated, all simulations results presented below are with the learning rate $k = 0.01$.

Layer	1	2	3	4
Percentile	95.0	95.0	95.0	95.0
Slope β	190	40	75	26

Table 5.1: Sigmoid parameters

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

Table 5.2: Lateral inhibition and excitation parameters for the SOM.

5.3 Pilot Study

5.3.1 Method

Stimuli

The stimuli used as part of the pilot study include three independently moving blocks. Each of the three different blocks occupied a different portion of the VisNet retina. Each portion, or visual space, was divided into 15 different locations within which a block could be located. Each block was presented with the other two blocks in all possible locations, such that there were $15 \times 15 \times 15 = 3375$ possible image combinations that are presented to the network during training. Each image was filtered in accord with the normal filtering routines outlined in Chapter 1, Section 1.3.3.

Training procedure

Figure 5.1 presents an annotated example of the three blocks presented together on the VisNet retina. In this example, the 'Top Block' is presented in position 1, and over the course of training it will occupy all equispaced locations until it finishes in position 15, as annotated in the figure. For the purposes of illustration, the 'Middle Block' is presented in position 7 and the 'Bottom Block' is presented in position 15, the final position. Figure 5.1 presents just one configuration of a possible 3375 training images.

Training comprises the presentation of all possible 3375 training images, where initially all the

blocks will occupy position 1. As training commences, the 'Bottom Block' will shift from position 1 to position 2, and eventually into position 15, before being reset to position 1. Accordingly, the 'Middle Block' will shift to position 2 and the process will repeat, where the 'Bottom Block' shifts one position at a time, from position 1 to position 15. The 'Bottom Block' is then reset back to position 1 while the 'Middle Block' is moved into position 3 and the process is repeated once more. Once the 'Middle Block' reaches position 15, it too is reset such that both the 'Middle Block' and the 'Bottom Block' occupy position 1, while the 'Top Block' is shifted to position 2 and the whole process is repeated until the 'Top Block' is in position 15. At which point, the presentation of all 3375 training images is complete. The complete presentation of all 3375 training images comprises one training epoch. Each layer of the model is trained for 50, 100, 100, 75 epochs for VisNet layers 1 through 4, respectively. It should be noted that the stimulus presentation order was also randomised and had no qualitative effect on results.

Testing procedure

Having trained the VisNet model with all 3375 block combinations, network performance was tested. Each block was presented in isolation shifting across all 15 possible locations that together comprise the visual space. Accordingly, there was a total of 45 images that comprise testing, 15 images from each of the three blocks. During testing the firing rate of the neurons within the VisNet model were recorded.

5.3.2 VisNet Simulation Results

The results of the pilot study show that VisNet has failed to develop separate representations of the three individual block stimuli. Figure 5.2 presents response plots for a prototypical output cell before

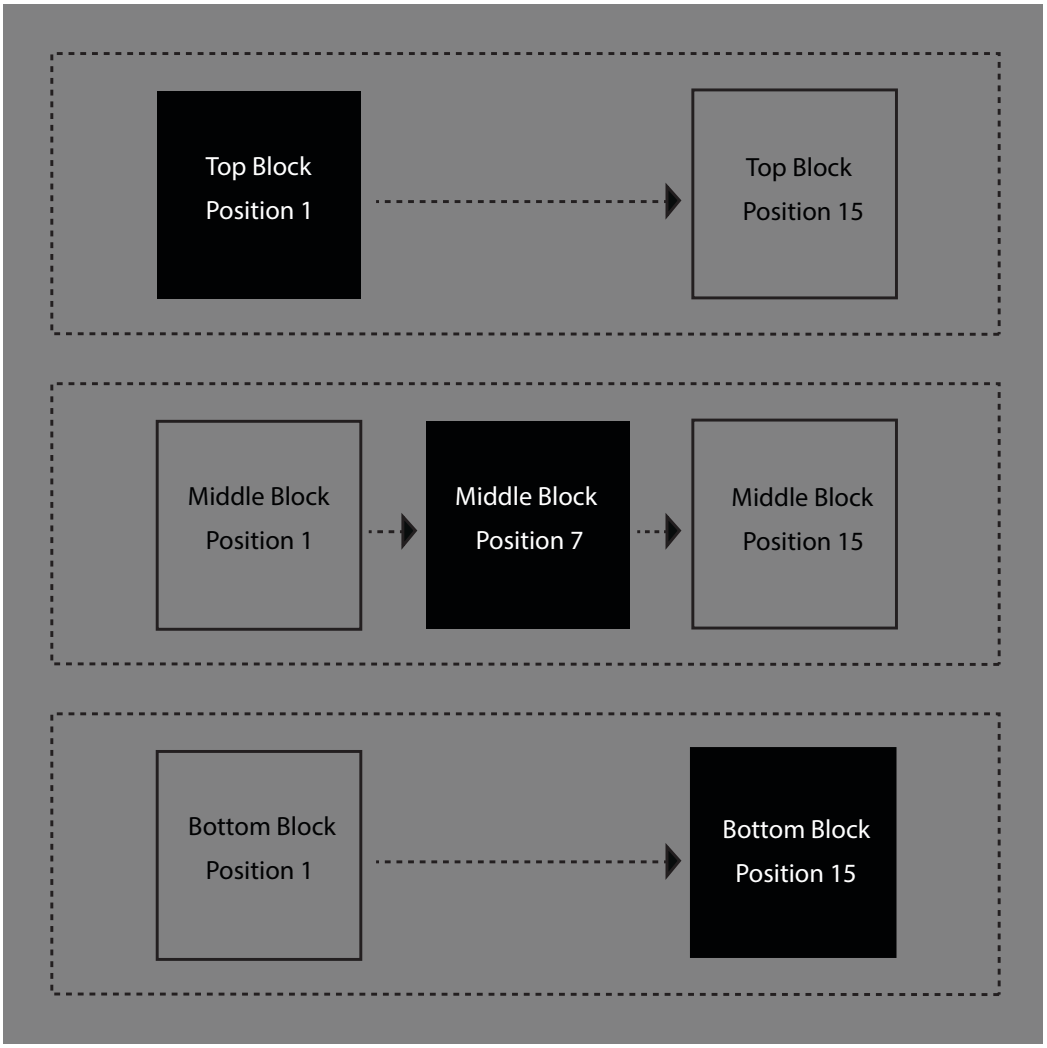

Figure 5.1: Three block stimuli. This figure presents an annotated example configuration of the three blocks used as part of the pilot study. The three visual spaces within which each block can be located are shown by the large dotted rectangles. The direction of movement is illustrated with arrows and the opaque squares represent the current location of each block: the 'Top Block' is located in position 1; the 'Middle Block' is located in position 7; while the 'Bottom block' is in position 15.

and after training. Unlike the pilot study presented within Chapter 4, the VisNet output neurons fail to respond exclusively to just one preferred stimulus. The same neurons tend to learn to respond to small portions of the view space for each of the three blocks, such as when the block is in position 1 through position 5. However, in some cases neurons only respond to one or two views. Furthermore, an output neuron may learn to respond to one block when it is only located within the first 5 positions that comprise its visual space, and then respond to a different block for a different part of the visual space. Although not presented, the lack of exclusivity means that an information analysis would show that output neurons convey relatively little information about a specific block. It should be noted that the responses to a small subset of locations is desirable.

These results have important implications for the main experiments presented later as part of this chapter. Most importantly, they suggest that the VisNet model would struggle to develop separate representations of three or more visual dimensions of a face if they were varied in a similar manner to that described above. That is, if three or more different parts of a face moved independently from one another, neurons in the output layer of the VisNet model would not necessarily learn to respond to each visual space exclusively. Accordingly, the stimuli used as part of the main experiments only contain two visual spaces. These in turn map to facial identity and facial expression.

Future studies should give further consideration to these results to better understand why the output neurons of the network have developed in this way when trained using three visual spaces. Having successfully shown in Section 4.3 that two visual spaces may be separated by the VisNet model in the manner outlined above, it is a logical progression to investigate the performance of the model with three or more visual spaces. If VisNet is unable to separate three or more visual spaces after further simulation, then the mechanism as a whole must be re-evaluated. In particular, whether

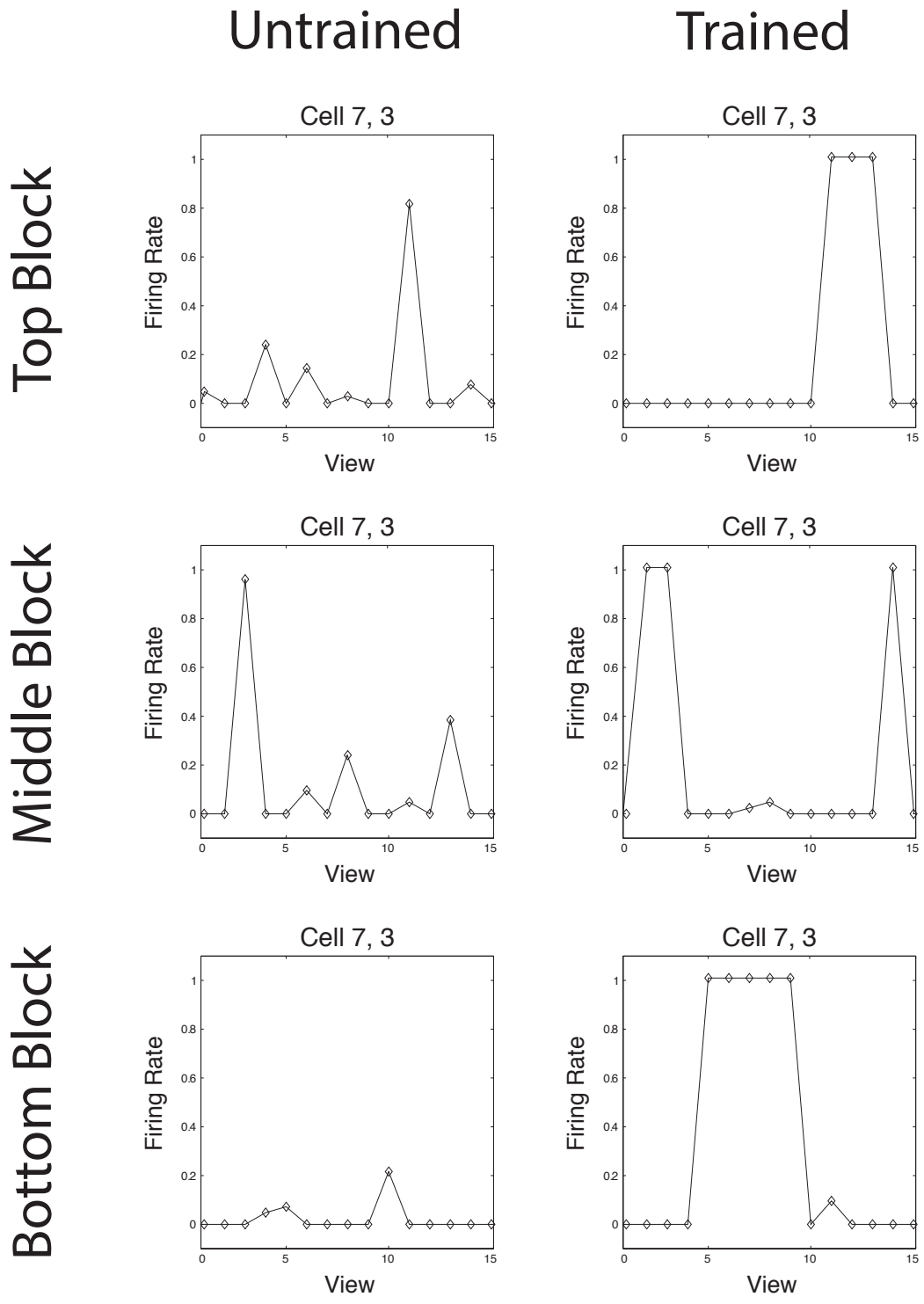

Figure 5.2: Three block cell response plots. This figure presents the cell response plots before and after training from a prototypical output cell 7, 3, in response to testing with each of the three blocks translating individually. It can be seen that the cell has failed to learn to respond exclusively to just one preferred block. Instead, after training, the cell responds to views from more than one block.

this mechanism for learning about individual objects could exist within the primate visual system.

5.4 Main Experiments

The findings from the pilot study demonstrate that within this training paradigm, only two independently varying visual spaces may be used. Although not reported, preliminary simulations with cartoon faces revealed that the highest spatial frequencies are important for the successful self-organisation of the model. More specifically, a high level of spatial resolution was necessary to separate the different images and avoid them being associated together onto the same output cells.

5.4.1 Method

Stimuli

Due to the findings from the pilot study, the Visnet model was trained with the computer-generated images of 2D faces that are shown in Figure 5.3. The use of carefully-constructed cartoon faces allowed for the controlled investigation of the hypothesis outlined in the Introduction of this chapter, whilst also satisfying the criteria for only two visual spaces varying independently. Accordingly, the simplest case was examined, in which the identity and expression spaces are one-dimensional and different facial features encode either identity or expression but not both.

These cartoon face stimuli were created as follows. First, 40 continuously varying transforms of identity were created by exclusively varying the shape and location of the nose and eyes (Figure 5.3: left to right). Then, 40 continuously varying transforms of expression were created by varying the shape and location of the mouth and eyebrows (Figure 5.3: top to bottom). The different identities and expressions were then combined to create 40 x 40 = 1600 combinations of facial identity and

expression. The final images were presented on a 128x128 pixel background each as a 256 grey scale image and are filtered in accord with the standard VisNet filtering routines.

Training procedure

To train the network, all identity and expression combinations were presented creating a total of 1600 possible faces. One training epoch comprises all 1600 face presentations. At each presentation of a face to the network, the activation levels of individual neurons are calculated, then their firing-rates are calculated and finally their synaptic weights updated in accordance with the procedure outlined in Section 1.3.3.

The network was trained one layer at a time from layer 1 through 4. The feedforward architecture means that each layer must wait for learning in the previous layer to converge before it can also converge. In previous studies, training the network one layer at a time has been found not to affect the qualitative performance of the network in comparison to training all four layers at once (Rolls and Milward, 2000). There are 50, 100, 100, 75 training epochs used for layers 1, 2, 3 and 4 respectively.

Testing procedure

After training all four layers of the network with the 1600 face images (each comprising unique combinations of identity and expression) the performance of the network was tested in two ways. First, the network was tested using the same complete faces that were presented during training. Secondly, network performance was also tested by presenting either identity or expression in isolation. In both cases, the firing-rates were recorded from all neurons within the model.

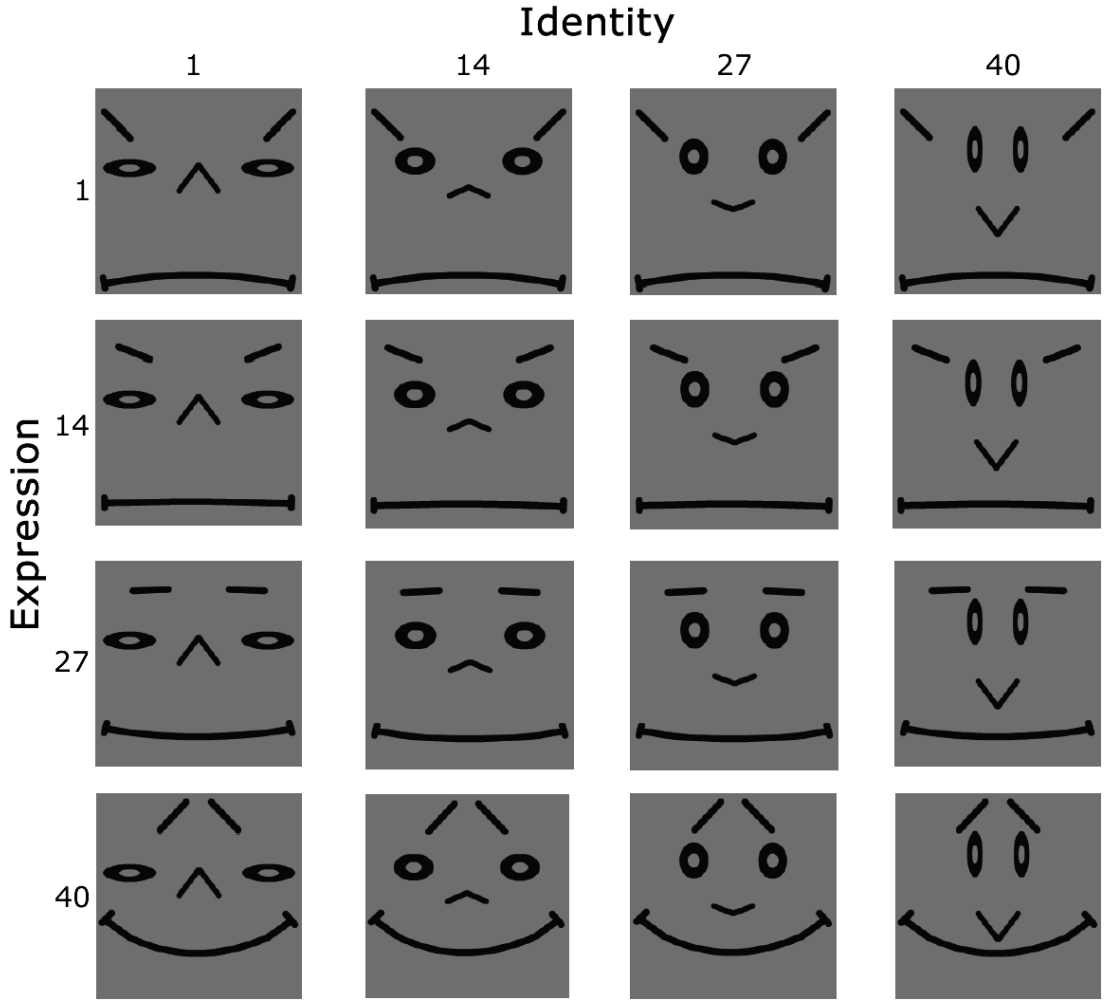

Figure 5.3: Cartoon face stimuli. A selection of 16 cartoon face stimuli presented to VisNet where identity (represented by the shape and location of the eyes and nose) changes along the horizontal dimension (left to right) and expression (represented by the shape and location on the mouth and eyebrows) changes along the vertical dimension (top to bottom). A total of 1600 faces were produced from all possible combinations of the 40 identities and 40 expressions.

Information measures

The network's ability to recognise varying identities or expressions during testing is also assessed using information theory (Rolls et al., 1997; Rolls and Milward, 2000). The visual face stimuli used were constructed from a large number (40) of different expressions and identities in order to represent near continuous spaces of expression and identity. However, VisNet's information analysis measures were originally designed for a discrete number of visual stimuli, each of which was presented over a range of different transforms (Rolls et al., 1997; Rolls and Milward, 2000). Therefore, in order to apply these information measures to the current work, it was necessary to quantise the identity and expression spaces as follows.

In order to analyse how much information about facial identity was carried by neurons within the fourth (output) layer of the network, the 40 identities were quantised into five contiguous blocks of eight. During testing, a central identity from within each of the five blocks was presented, which was then combined with all 40 expressions. If a subpopulation of neurons in the fourth layer had learned to represent exclusively facial identity, then each neuron within that subpopulation should ideally respond selectively to only one of the five quantised identities. Furthermore, each neuron should respond to its favoured identity invariantly over all of the 40 expressions that can be paired with that identity. In this case, such neurons will convey maximal information about facial identity.

In a separate analysis, the information measures were also used to assess how much information was carried by fourth layer neurons about facial expression. This was done by quantising and testing the expression space in a corresponding manner.

The single cell information measure has been used to show how much information is available from the response of a single cell about a stimulus that was presented over a range of different trans-

forms (Rolls et al., 1997; Rolls and Milward, 2000). The stimulus is taken to be one of the five quantised identities or one of the five quantised expressions. When assessing the amount of information conveyed about identity, then the five quantised identities are treated as the five stimuli, and the 40 expressions are treated as transforms of those five identity stimuli. Conversely, when assessing the amount of information conveyed about expression, then the five quantised expressions are treated as the five stimuli, and the 40 identities are treated as transforms of those five expression stimuli.

The single cell information measure for each cell shows the maximum amount of information that the cell conveys about any one stimulus over all transforms. The maximum amount of information that can be attained is $\log_2(N)$ bits, where N is the number of stimuli. For the case of five stimuli, the maximum amount of information is 2.32 bits.

5.4.2 VisNet Simulation Results

The VisNet model was trained on all possible combinations of expression and identity to create a total of 1600 faces. The results show that the network was able to form separate localised clusters of neurons in the fourth (output) layer which represented the two independent visual spaces, identity and expression, even though they are presented together as complete faces during training. Crucially, neurons in the fourth layer of the network learned to respond to different portions of each visual space. Some neurons learned to respond to a small part of the space of identity irrespective of expression. Similarly, other neurons learned to respond to a small part of the space of expression irrespective of identity.

Analysis of firing rate responses

Figure 5.4 shows neuronal firing rate responses to complete faces. The firing rate responses of an 8×8 subset of fourth (output) layer cells are shown before training (left) and after training (right) with the full set of 1600 complete faces. The 8×8 neurons are represented by 64 2D response plots, one per neuron. Each 2D response plot shows the neuron's firing-rate response to all 1600 possible faces. For each subplot, the x-axis is bound between 1 and 40, where each discrete point along the horizontal axis represents a different transform of the identity space. Similarly, each discrete point along the vertical axis represents a different transform of the expression space. The firing-rate of each neuron is then plotted for the different combinations of the 40 identities and 40 expressions. Stronger firing is depicted by a darker shading.

In the untrained condition, cells responded randomly due to the untrained random weight structure in the network. However, after training, some neurons responded to local portions of their preferred space, either identity or expression, irrespective of the position in the alternative space. This is evidenced by the appearance of vertical or horizontal 'bars' of activation in the subplots. If a neuron responds selectively to a local region of the identity space, then this results in a vertical bar in the subplot. Alternatively, if a neuron responds to a local region within the expression space, then this results in a horizontal bar. Furthermore, due to the effects of the SOM architecture, cells with similar response profiles form close together. This causes spatial clustering of cells that respond preferentially to either facial identity or facial expression. The firing rates tended to be rather binarised, i.e. near zero (white) or one (black), because of the steep slopes of the sigmoid transfer functions used in the present simulations.

The model was also tested with each of the 40 identities and 40 expressions used to generate the

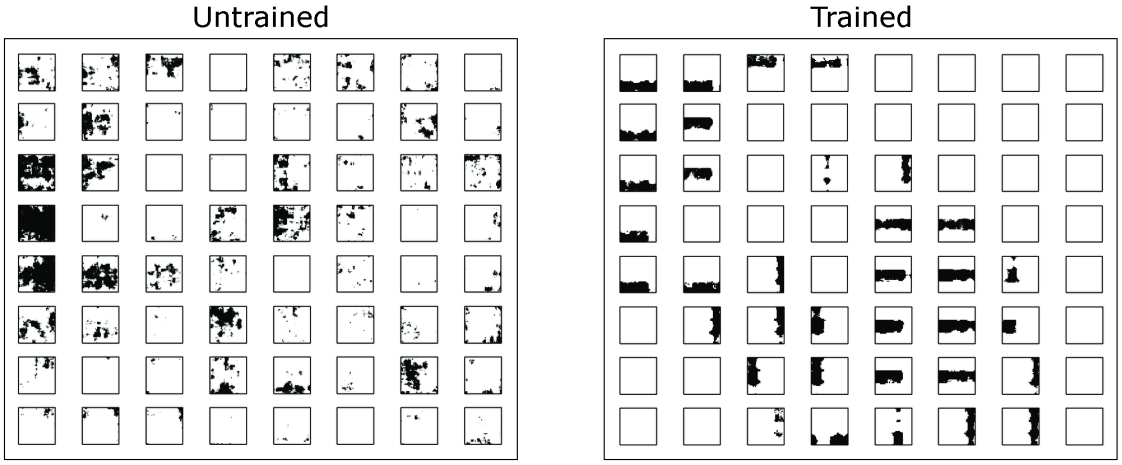

Figure 5.4: Neuronal firing rate responses to complete faces. The firing rate responses of an 8×8 sample of fourth (output) layer cells are shown before training (left) and after training (right) with the full set of 1600 complete faces. Each of the 8×8 sub-plots represents an individual cell in the output layer, and shows its firing-rate profile to all 1600 faces presented during testing. For each subplot, the horizontal axis denotes the position of the test face within the identity space and the vertical axis denotes the position within the expression space. The responses of the neuron are represented on a grey scale where black indicates high firing. After training, some individual neurons learned to respond to local portions of their preferred space, either identity or expression, irrespective of the position in the alternative space. This results in the appearance of vertical or horizontal 'bars' of activation in the subplots. If a neuron responds selectively to a local region of the identity space, then this results in a vertical bar in the subplot. In contrast, if a neuron responds to a local region within the expression space, then this results in a horizontal bar. Furthermore, due to the effects of the SOM architecture, cells with similar response profiles form close together. This causes spatial clustering of cells that respond preferentially to either facial identity or facial expression.

1600 faces presented during training. Figure 5.5 shows neuronal firing rate responses to either identity or expression. The network is first tested with each one of the 40 images that comprise the identity space, and then tested with the 40 images that comprise the expression space. The figure displays the responses of four different kinds of typical output cells: (A) a cell that responds exclusively to a portion of the identity space; (B) a cell that responds exclusively to a portion of the expression space; (C) a cell that responds to portions of both the identity and expression space; and (D) a cell that responds to multiple portions of either or both spaces. Cell responses to the 40 transforms of identity are plotted with the solid line, and cell responses to the 40 transforms of expression are plotted with the dashed line. For each cell, the firing-rate (0-1) is plotted on the y-axis against the identity/expression number (1-40) on the x-axis.

It was also investigated whether cells that responded to a part of either the expression space or the identity space had learned to respond to a conjunction of the corresponding input features. By tracing the synaptic weights after training from cells in layer four back to the input filters, it was found that this was indeed the case. For example, in the case of identity, it was found that individual cells in layer four, which responded selectively to a particular part of the identity space, received strengthened connections from corresponding conjunctions of both the eyes and the nose. In the case of expression, it was found that cells received strengthened connections from all parts of the mouth including the two corners of the mouth, although with weaker connections from the eyebrows.

Information Analysis

Figure 5.6 shows the results of analysing the amount of information about expression conveyed by fourth (output) layer cells before and after training on 1600 complete faces. The information analysis was performed as described in Section 1.3.5 and more specifically in Section 5.4.1. This analysis

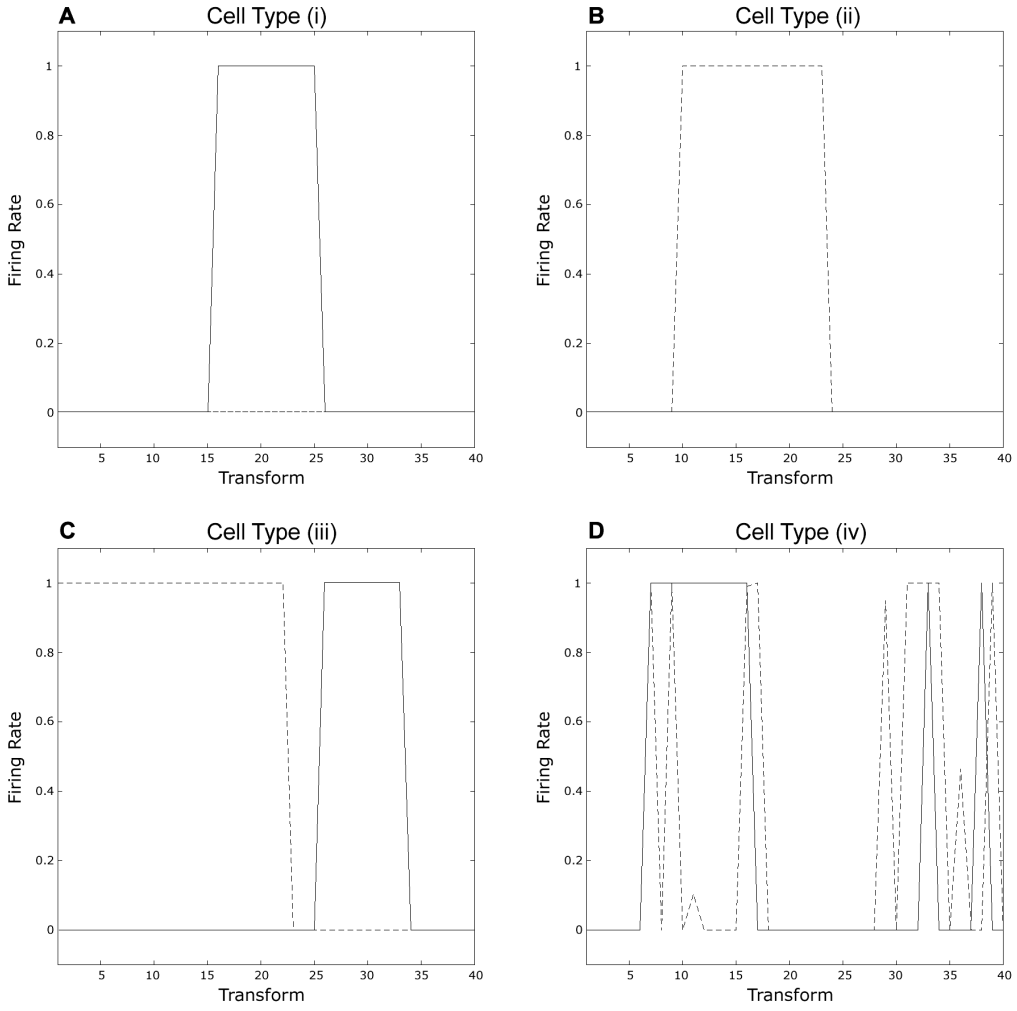

Figure 5.5: Neuronal firing rate responses to either identity or expression. First the network is trained with all 1600 complete faces. Then the network is tested with the visual features that represent either the 40 transforms of facial expression or the 40 transforms of facial identity presented separately. The figure displays the responses of four different kinds of typical output cells: (A) a cell that responds exclusively to a portion of the identity space; (B) a cell that responds exclusively to a portion of the expression space; (C) a cell that responds to portions of both the identity and expression space; and (D) a cell that responds to multiple portions of either or both spaces. Cell responses to the 40 transforms of identity are plotted with the solid line, and cell responses to the 40 transforms of expression are plotted with the dashed line. For each cell, the firing-rate (0-1) is plotted on the y-axis against the identity/expression number (1-40) on the x-axis.

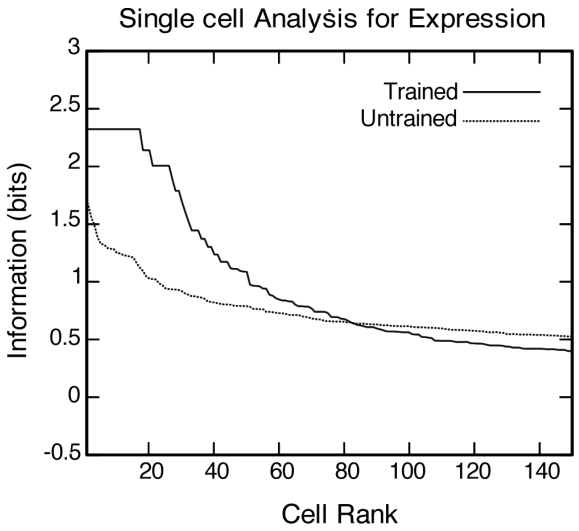

Figure 5.6: Information analysis for expression. Analysis of information about expression conveyed by fourth (output) layer cells before and after training on 1600 complete faces. The information analysis was performed as described in the Methods. The plot shows the amount of single cell information carried by individual output cells in rank order. In the trained condition, 17 neurons provided maximal information of 2.32 bits. In the untrained condition, no cells reached maximal information.

involved quantising the expression space into five separate contiguous blocks. The maximal amount of information possible in this case is 2.32 bits. Figure 5.6 shows the amount of single cell information carried by individual output cells in rank order. In the trained condition, 17 neurons provided the maximal information of 2.32 bits. These 17 neurons responded selectively to a particular quantised portion of the expression space irrespective of the identity. In the untrained condition, no cells reached maximal information.

It can be seen that training the network on all 1600 complete faces has led to a dramatic increase in the information carried by the fourth layer neurons to the expression of the current face irrespective of identity.

Figure 5.7 shows the results of analysing the amount of information about identity conveyed by fourth (output) layer cells before and after training on 1600 complete faces. This analysis involved quantising the identity space into five separate contiguous blocks as described in the Section 5.4.1.

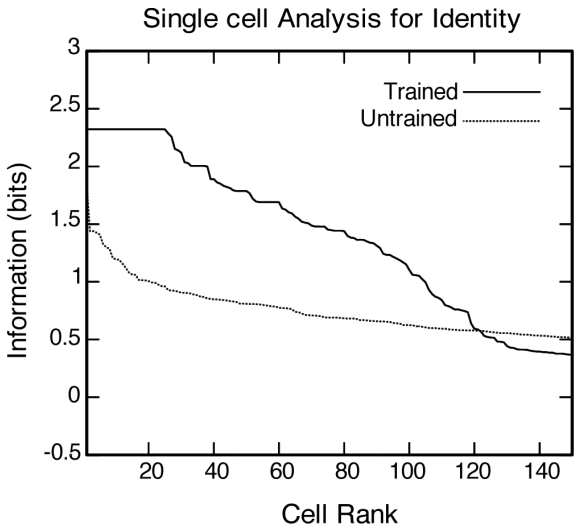

Figure 5.7: Information analysis for identity. Analysis of information about identity conveyed by fourth (output) layer cells before and after training on 1600 complete faces. The information analysis was performed as described in the Methods. The plot shows the amount of single cell information carried by individual output cells in rank order. In the trained condition, 25 neurons provided maximal information of 2.32 bits. In the untrained condition, no cells reached maximal information.

The plot shows the amount of single cell information carried by individual output cells in rank order. In the trained condition, 25 neurons provided the maximal information of 2.32 bits. In the untrained condition, no cells reached maximal information.

In summary, Figs. 5.6 and 5.7 confirm that, after training on the 1600 complete faces, the output layer cells have learned to convey large amounts of information about either facial identity or expression. This is because different subpopulations of cells in the output layer have learned to respond selectively to just one of the visual spaces, identity or expression, irrespective of the current transform of the other space. Some cells respond exclusively to a particular portion of the identity space, while other cells respond exclusively to a particular portion of the expression space.

5.5 Follow-up Study

The follow-up study presented as part of this section builds on the previous results obtained with cartoon faces from the pilot study. It makes use of more realistic 3D faces produced by professional face generation software. The lack of resolution of the VisNet retina was a contributing factor to the decision to create very simple cartoon faces in the pilot study. During some of the preliminary simulations (not shown), the cartoon faces were made more detailed. These preliminary simulations demonstrated that the same set of output neurons were unable to differentiate between the facial expressions and identities of the more detailed cartoon faces. Therefore, this follow-up study makes use of a VisNet model with more processing neurons and a larger retina. The following simulations form the basis of an avenue for future research.

5.5.1 Method

The simulations are performed using a larger version of the VisNet model built in response to the limitations outlined above, specifically, the constraints on image resolution. In principle, the model is identical apart from two important aspects. Firstly, the image filtering routines have been updated. Secondly, the number of neurons within each layer of the VisNet model has been increased.

Gabor Filtering

As previously outlined in Section 1.3.3, before the stimuli are presented to the VisNet input layer, they are pre-processed by a set of input filters which accord with the general tuning profiles of simple cells in V1. In all VisNet simulations to date, the input filters used are computed by weighting the difference of two Gaussians by a third orthogonal Gaussian. Each retinal location comprises a bank

of filters produced from the different spatial frequencies, orientations and polarity. In this scaled-up simulation at each retinal location there now exists a bank of Gabor filter outputs (Daugman, 1985) corresponding to a hypercolumn generated by

$$g\left(x, y; \lambda, \theta, \psi, \sigma, \gamma\right) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{x'}{\lambda} + \psi\right) \tag{5.1}$$

$$x' = x \cos\theta + y \sin\theta \tag{5.2}$$

$$y' = -x \sin\theta + y \cos\theta \tag{5.3}$$

for all combinations of $\lambda = 2, \gamma = 0.5, \sigma = 0.56\lambda, \theta \in \{0, \pi/4, \pi/2, 3\pi/4\}$ and $\psi \in \{0, \pi, -\pi/2, \pi/2\}$. The Gabor function provides a useful and reasonably accurate description of most spatial aspects of simple receptive fields (Jones and Palmer, 1987)[1].

Model Architecture

The scaled-up version of the VisNet model has a larger input layer and also more neurons within each of the four processing layers. More precisely, in place of the 128×128 input locations, there are now 512×512 locations; instead of the 32×32 neurons within each processing layer, there are now 128×128 neurons. The method with which a new untrained network is created follows the same procedures used for the previous version of VisNet. However, the number of connections between layers and the sampling radius were also increased. All other aspects of the model architecture remain constant. The exact parameters used for this study are presented in Table 5.3.

[1]These specific Gabor filtering code routines were programmed by Ben Evans, a member of the Oxford Foundation for Theoretical Neuroscience and Artificial Intelligence (OFTNAI) lab, prior to implementation into the scaled-up version of the VisNet model.

	Dimensions	Number of Connections	Radius
Layer 4	128x128	400	12
Layer 3	128x128	400	9
Layer 2	128x128	400	6
Layer 1	128x128	1000	6
Retina	512x512x32	-	-

Table 5.3: Network dimensions. The number of connections per neuron and the radius in the preceding layer from which 67% of connections are received

Stimuli

The scaled-up VisNet model is trained and tested with realistic computer-generated faces. These faces are created using the FaceGen software produced by Singular Inversions. This software allows for the systematic control over various facial characteristics. This level of control is sufficient to alter almost any part of the face, thus making it a perfect tool for generating faces that vary with respect to identity and expression.

Unlike the stimuli used in the previous experiments, the features that comprise identity and expression are no longer completely orthogonal. The degree to which the different features interact with one another is much harder to control. However, since these experiments are designed to investigate how a scaled-up version of VisNet would perform with more realistic human faces, this is in fact an important characteristic of the stimuli.

Figure 5.8 presents 9 example faces. As in the previous section, identity changes along the horizontal dimension (left to right) over 40 transforms, and expression changes along the vertical dimension (top to bottom) also over 40 transforms. In total, there are still 1600 faces with which the scaled-up VisNet model is presented. The final images were presented on a 512×512 pixel background each as a 256 grey scale image and are filtered in accord with the new Gabor filtering routines outlined above.

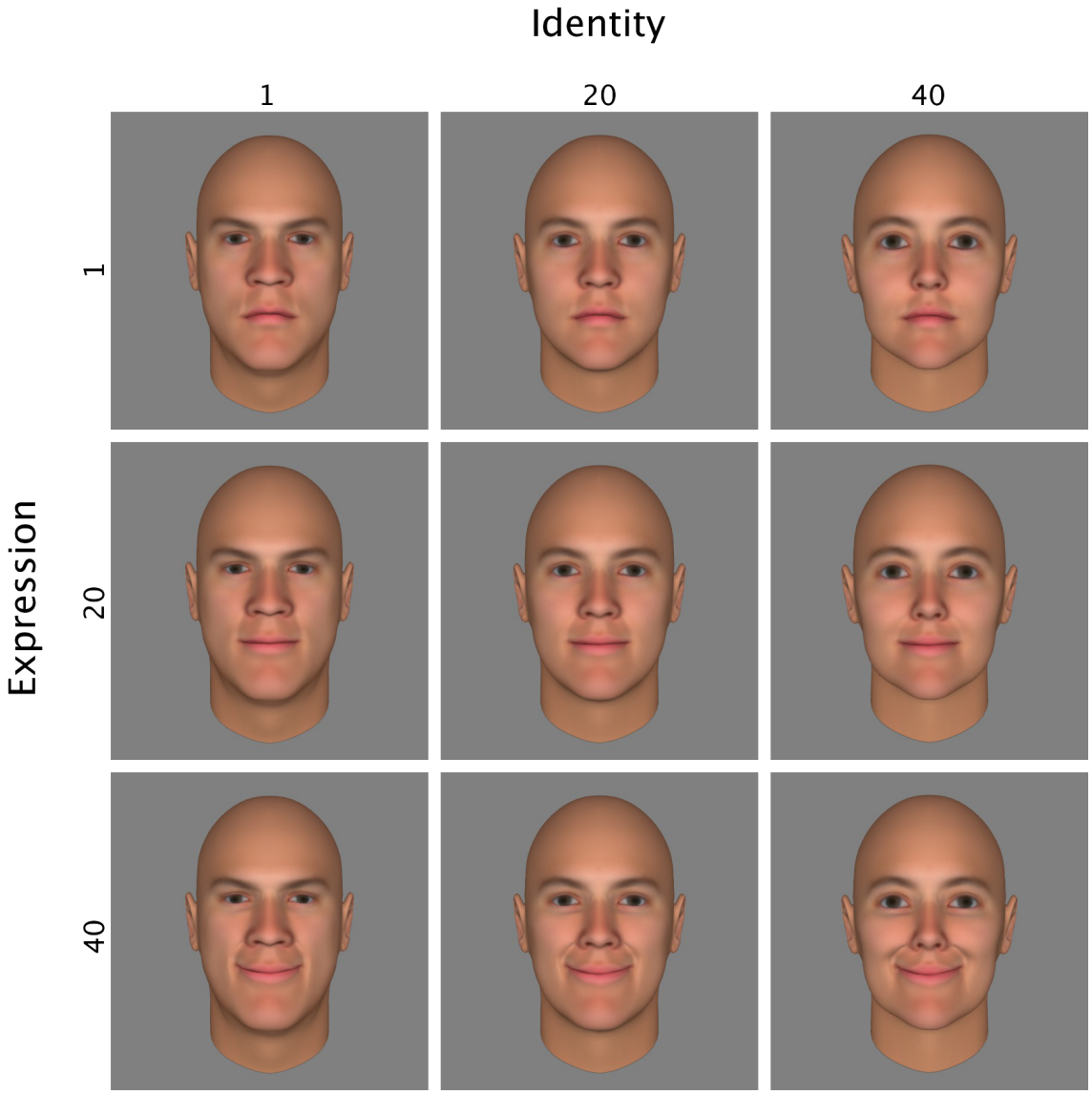

Figure 5.8: Realistic face stimuli. A selection of 9 example face stimuli used the train and test the scaled-up VisNet model are presented. Identity changes along the horizontal dimension (left to right) over 40 transforms, and expression changes along the vertical dimension (top to bottom) also over 40 transforms. These faces are created using the FaceGen software product and in total there are $40 \times 40 = 1600$ faces that together comprise the complete stimuli set.

Training procedure

The training procedure used in this simulation is identical to the training procedure used in the main experiments and presented as part of Section 5.4.1. In summary, all 1600 faces were presented to the scaled-up network. The order of face presentation was sequential, but randomised presentation was also attempted and did not qualitatively effect the results. One training epoch comprises all 1600 face presentations. At each presentation of a face to the network, the activation levels of individual neurons are calculated, then their firing-rates are calculated and finally their synaptic weights updated in accordance with the procedure outlined in Section 1.3.3. Layers 1 through 4 were trained sequentially for 50, 100, 100, 75 training epochs, respectively.

Testing procedure

To test the performance of the scaled-up model after training, all 1600 face images were presented to the network and the firing-rates were recorded from all neurons within the model. These results are directly comparable to those presented in the previous section. It should be noted that testing did not include presenting each visual space varying independently. This method of testing is no longer possible due to the manner in which the stimuli are created and also due to the fact that the two different visual spaces are no longer completely orthogonal.

5.5.2 VisNet Simulation Results

Analysis of Output Cell Firing Properties

Figure 5.9 shows neuronal firing rate responses to complete faces after having trained and tested the scaled-up VisNet model with all 1600 possible face stimuli. This figure is directly comparable with

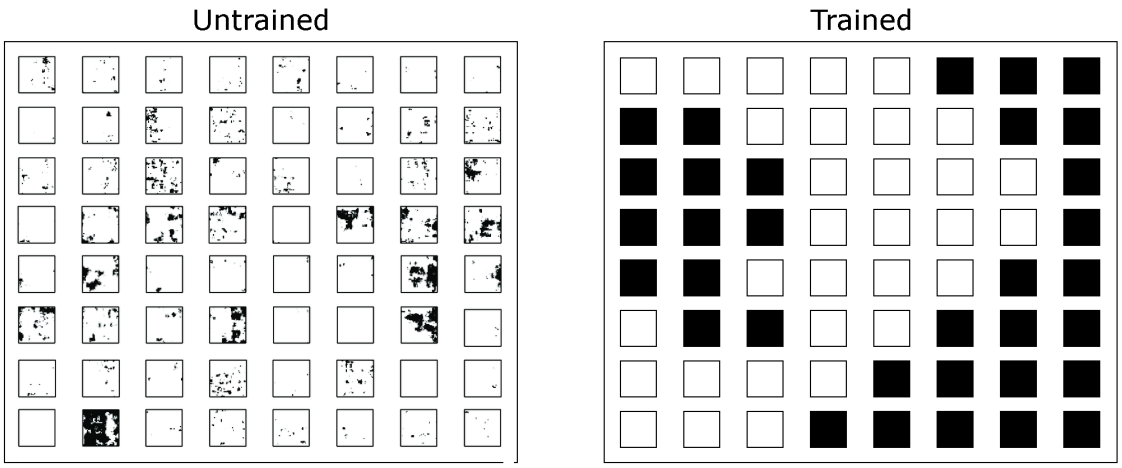

Figure 5.9: Neuronal firing rate responses from layer 4 to complete realistic faces. The figure shows an 8×8 subset of fourth (output) layer cells before training (left) and after training (right) with the full set of 1600 complete faces. Each 8×8 subset comprises 64 2D response sub-plots representing the firing rates produced by an individual neuron. For each sub-plot, the x-axis is bound between 1 and 40, where each discrete point represents a different transform of the identity space. This is also the case for expression, which is on the vertical axis. Stronger firing is depicted by a darker shading. Prior to training, neurons respond without preference to the different faces. After training, neurons have learned to respond to all the different faces, irrespective of identity or of facial expression.

Figure 5.4 presented in Section 5.4.2. In summary, the figure shows an 8×8 subset of fourth (output) layer cells before training (left) and after training (right) with the full set of 1600 complete faces. Each 8×8 subset comprises 64 2D response sub-plots representing the firing rates produced by an individual neuron. For each sub-plot, the x-axis is bound between 1 and 40, where each discrete point represents a different transform of the identity space. This is also the case for expression, which is on the vertical axis. Stronger firing is depicted by a darker shading.

It can be seen that prior to training, output neurons respond sporadically, and fail to respond in a meaningful or organised way. After training, the neurons have learned to respond to all the views of the faces, and fail to make any differentiation between the different views of each face. This is shown by dark firing rates across all possible combinations of facial expression and identity, resulting in completely darkened cell response plots.

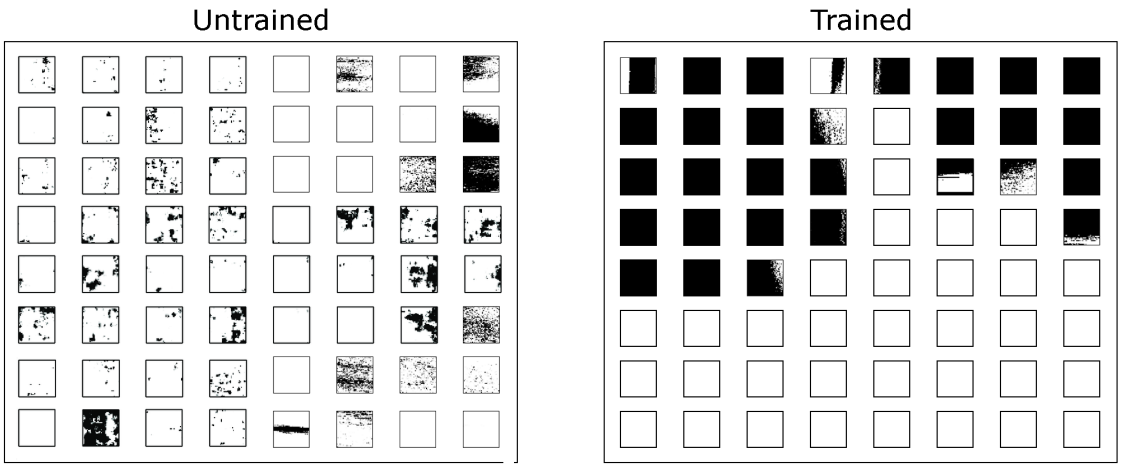

Figure 5.10: Neuronal firing rate responses from layer 1 to complete realistic faces. The figure shows an 8×8 subset of first layer cells before training (left) and after training (right) with the full set of 1600 complete faces. Prior to training, some neurons respond without preference to the different faces. However, there are some neurons that already respond to a large number of faces. After training, most neurons have learned to respond to all the different faces, irrespective of identity or of facial expression. Conventions as in Figure 5.9.

Analysis of Cell Firing Properties in Earlier Layers

The responses of cells in the output layer of the model are dependent upon the responses of cells in the intermediate and earlier layers. For example, if every neuron within the third layer of VisNet responds to every face stimuli regardless of which features are visible on the retina, then the fourth layer neurons will only be able to respond in a similar manner; they too will respond invariantly across all the different face stimuli.

Figure 5.10 shows the cell responses of neurons from the first layer of the scaled-up VisNet model before and after training. All conventions are as in Figure 5.9. Even prior to training there are some neurons that respond to almost all of the 1600 different faces. After training, the majority of first layer neurons respond to every face regardless of facial expression or identity. Although not presented, further inspection of neurons within the intermediate layers of the scaled-up VisNet model reveals that they too respond to all faces without regard to facial expression or identity. In conclusion, the

results reveal that VisNet has been unable to develop separate representations of facial expression and facial identity.

5.6 Discussion

The following section will first address the results obtained from the main experiments. Consideration is then given to what was learned from the pilot study. Next, the results of the follow-up study, which used more realistic face stimuli, are reviewed. Finally, avenues for future research are presented and special emphasis is placed on further use of the FaceGen software and use of realistic faces.

A number of experimental studies have shown that physically separated subpopulations of neurons develop in the visually responsive areas of the primate brain that respond primarily either to facial identity or to facial expression (Hasselmo et al., 1989a; Perrett et al., 1992; Rolls et al., 2006; Gothard et al., 2007). The results of the main experiment simulations with VisNet were confirmed by examination of the firing rate responses of the output cells, as well as analysis of the information carried by these cells about facial identity and expression. This is a non-trivial problem because the visual system is always exposed to combinations of facial identity and expression during learning. The model solves the problem by exploiting how SOMs self-organise their output neuronal responses when they are trained on statistically independent visual input spaces such as facial identity and expression.

Experimental studies have demonstrated that a dissociation between representations of facial identity and expression occurs across a number of visually responsive areas of the primate brain, both within and outside the ventral visual pathway. VisNet has demonstrated biologically plausible learning mechanisms, specifically how learning in SOMs is shaped by the statistical independence between facial identity and expression, that may operate across these different but connected brain areas. Fur-

ther simulation work could explore how more architecturally detailed models of these interconnected areas of the brain develop their heterogeneous representations of facial identity and expression.

The results presented here are robust. Further simulations showed that altering the learning rates over three orders of magnitude did not qualitatively alter the findings. Also, the effects of varying levels of global competition within the model were explored by adjusting the sparseness percentile values. The model performed well with a sparseness of 4% to 10% within each layer, that is when 4% to 10% of the neurons were allowed to fire during any given stimulus presentation. Large variations in learning rate, sparseness, and the width of the SOM still produced spatially organised clusters of cells that were responsive to either facial identity or expression.

A spatial filter $I_{a,b}$ that simulated short-range excitatory and long-range inhibitory connections was introduced to the VisNet model to create the SOM network architecture. There is some anatomical evidence demonstrating that lateral inhibitory connections in V1 tend to project over a smaller proximity than excitatory connections (Fitzpatrick et al., 1985; Bosking et al., 1997; Tucker and Katz, 2003); this is the opposite of the SOM architecture that has been added to the original VisNet model. However, if lateral inhibitory connections are fast acting, a 'Mexican-hat' style functional architecture may still be achieved (Kang et al., 2003). The SOM architecture implemented in this doctoral thesis is similar to others that are successful in replicating a range of experimental findings (Sirosh and Miikkulainen, 1994; Choe and Miikkulainen, 1998; Bednar and Miikkulainen, 2006).

The tuning profiles of cells shown in Figure 5.5 are consistent with non-human primate single cell recording data from Hasselmo et al. (1989b); in their study, only 6.7% of cells responded to facial expression and identity. However, more recent research has shown that in the monkey amygdala, 64% of recorded cells responded to both facial expression and identity (Gothard et al., 2007). This cell type

is presented in Figure 5.5 (iii). Indeed, fMRIa studies with human subjects report clusters of cells in the fusiform face area (FFA) that respond to both identity and expression (Xu and Biederman, 2010). This type of heterogeneous cell type is consistent with the simulations reported in this chapter, which produced four types of cell in the output layer.

A number of studies have revealed that neonates are able to mimic certain facial gestures (Meltzoff and Moore, 1977, 1983). Furthermore, neonates show preference for familiar faces over novel faces, suggesting they can recognise different identities (Walton et al., 1992). However, the question of innate functional segregation remains an open issue. The results presented within this chapter show how a functional segregation between facial identity and expression could be learned through visual experience without requiring any innate functional segregation.

The VisNet architecture has a number of limitations in terms of biological accuracy. For example, the current VisNet model lacks back-projections, which are a major feature of the ventral visual pathway in the primate brain. In future developments of the VisNet model, the effects of these additional connections during learning will be explored. Also, because the current model is purely feedforward, it is possible to train the layers one at a time from layer 1 to 4. With the introduction of back-projections, all of the layers will need to be trained simultaneously. Moreover, the neurons in the current model are rate-coded with rather binarised firing rates. Future studies will implement integrate and fire neurons, which explicitly model the emission of action potentials. This in turn will allow the exploration of the effects of spike time dependent plasticity on visually-guided learning in the network.

The pilot study served to outline an important limitation of the VisNet model. It was shown that VisNet output neurons are unable to separate three distinct visual spaces from one another when the

spaces vary independently. This limitation was addressed as part of the follow-up study.

The follow-up study used more realistic human faces. In the main experiments, the use of carefully designed cartoon faces allowed for exploring the simplified situation in which the visual spaces of facial identity and expression are independent, one-dimensional, and non-overlapping on the retina. However, real faces are more complex in a number of ways. Firstly, expression may not be entirely independent of identity. Secondly, identity and expression are not simple one-dimensional spaces. Thirdly, individual facial features such as the mouth may convey information about both identity and expression simultaneously. In this case, the visual representations of identity and expression are encoded by overlapping sets of retinal input neurons in a more distributed manner. The follow-up study tested how the learning mechanisms demonstrated in the main experiment simulations might work with face images that are more realistic in all of these respects. The more realistic human faces were produced by the FaceGen modeller software package, which is used in a number of fMRI studies investigating the separation of identity and expression (Xu and Biederman, 2010).

The results of the follow-up study showed that the VisNet output neurons were unable to learn separate representations of facial expression and identity during learning with more realistic FaceGen faces. Analysis of the earlier VisNet layers was also performed and the cell response plots were presented. It was shown that many of these cells also respond invariantly across all of the 1600 faces. If the majority of cells in the early layers of VisNet respond to all the stimuli presented during training, it follows that subsequent VisNet layers will also learn to respond to all stimuli. More specifically, if the neurons within the first layer of the VisNet model learn to respond to all faces, then no information is provided to subsequent layers regarding the state of the stimulus presented. For example, whether the facial expression is happy or sad. Accordingly, subsequent layers are unable to

use such information to develop selective representations since this information is not present within their inputs.

The high degree of invariance created in only the first layer of VisNet is due to the fact that some facial features presented did not vary during the course of learning. That is, features that were present in all 1600 face images, such as the face outline or eyebrows, were present in the same retinal location. The neurons in the first layer of the VisNet model have small receptive fields, but if many of the features that occupy these receptive fields never change, the neurons are continuously stimulated during training. When combined with Hebbian learning, the competitive nature of the VisNet layers causes the same cells to become active and thus compete with the other cells within the same layer. This can prevent these others cells from learning about the individual facial features.

The primate visual system is able to cope with the challenge outlined above, that is, it can ignore the contrast boundaries of the face outline and instead rely on the details provided by the inner facial features. More specifically, upon gazing at a face, monkeys primarily look towards the eyes and other inner facial features compared to areas surrounding the face (Nahm et al., 1997). On the other hand, VisNet neuronal outputs are biased towards representing the contrast of the face boundary. Accordingly, future studies with the scaled-up version of VisNet should continue to use face stimuli produced by the FaceGen modeller. In conjunction with this, future research should also investigate a biologically plausible mechanism for focussing on certain feature categories that could suppress the constant activation caused by a contrast boundary produced by a facial outline. This may require back-projections and perhaps some form of attentional mechanism. This will allow the VisNet model to develop representations that reflect the varying inner facial features and hopefully distinguish between them. The type of encoding that develops will be paramount to the direction of future research. More

specifically, it is likely that the method of encoding facial expression or identity will be through a heterogeneous population of cells. Additional analytical procedures may be required to accommodate this type of cell response.

Eventually the model should be trained and tested with real faces. This will allow for assessing the network's ability to generalise when tested with faces not presented during training. This may require VisNet's retina to be further increased in size in order to provide a sufficiently detailed input representation of the visual features encoding facial identity and expression.

Chapter 6

Continuous Mapping of Different Face Views: View Specific and Location Invariant Responses

The previous chapter presented results obtained when training and testing the VisNet model for the first time with a scaled-up network architecture. With the advent of new technology, it is now possible to investigate the performance of VisNet with much larger input images, and this increase in image resolution provides far more detail to the scaled-up VisNet model. Furthermore, the number of neurons within each processing layer was also increased, as were other key parameters such as the number of connections between layers. Specifically, it is now possible to run simulations with a 512×512 pixel retina, and with 128×128 neurons within each of the four processing layers. This type of network architecture was previously too computationally demanding to implement. The benefits afforded by this updated model are crucial for ongoing research with VisNet. For example, the previous chapter

used high resolution realistic face stimuli in order to investigate how VisNet may learn to separate facial expression and identity. The experiments in this current chapter make use of this larger version of VisNet in an attempt to exploit these benefits. Unless otherwise stated, throughout this chapter the term VisNet will be used to reference this scaled-up version of the model, and also general VisNet principles where scale is irrelevant.

It has previously been shown within Section 1.3.1 that the majority of the past VisNet research has focussed on presenting one stimulus at a time during training. Typically, the primary objective was to understand how the VisNet model might learn to develop an invariant representation of the stimulus presented upon the retina during the course of training. Throughout the other chapters of this thesis, an attempt has been made to move away from presenting just one object on the retina at a time in order to understand more about how the primate visual system may learn in a more complex world of multiple objects. However, with the implementation of the larger version of VisNet, single stimuli simulations are revisited in the context of experimental research.

The main objective of this chapter is to produce similar results to those presented by Wang et al. (1998) obtained during *in vivo* optical imaging studies with monkeys. Furthermore, the different types of neural representations that form in the scaled-up version of the VisNet model during training are investigated. More specifically, this chapter reports the implementation details, and simulation results, of the new VisNet model when trained with a high resolution, realistic face stimulus. There are two training paradigms, the first of which explores the continuous mapping of different face views by training the network with a series of images of the face stimulus rotating in depth in a specific central location upon the VisNet retina. The second training paradigm is similar, but the face translates between four different retinal locations in addition to rotating in depth. This second paradigm explores

whether the model is able to develop view specific responses that are also location invariant.

6.1 Introduction

Experimental research has investigated the functional organisation in the monkey visual cortex. It has been shown that different TE neurons are activated preferentially by specific individual objects, while other neurons respond selectively to faces (Wang et al., 1998). More specifically, Wang et al. (1998) reported that during an optical imaging study the activation spot changed gradually along the cortical surface as a stimulus face was rotated in depth. These results demonstrate two important points. Firstly, many cells in anterior IT cortex do not respond with view invariance to a preferred stimulus, instead responding to a preferred stimulus when presented at certain views. Secondly, cells with similar stimulus selectivity are clustered together in regions and some parameters of object features are continuously mapped. Additional results from the same study showed that some of the same neurons in TE also demonstrated location invariance. That is, neurons would respond to the same view of a face even if it was presented in different locations upon the retina. This is an interesting finding because it highlights the fact that there are cells that respond with view specificity but also location invariance.

However, as previously mentioned in Chapter 1, Op De Beeck and Vogels (2000) have shown that there are neurons with varying receptive field sizes withn IT, therefore capable of responding with varying degrees of location invariance. Although the size of the receptive field may change with the number (or size) of stimuli presented during testing, a robust result was reported; in most cases neurons responded maximally when a preferred stimulus was presented within a certain location inside the receptive field. Clearly, IT neurons are also capable of encoding spatial information even in larger

receptive fields based on their level of activation. Further research has also shown that there is strong evidence supporting the fact that position information is present in the population response of neurons in IT (Hung et al., 2005).

Further evidence outlining the types of spatial information that are present within the ventral visual stream comes from a comparison study between anterior inferotemporal cortex (AIT) and lateral intraparietal cortex (LIP) of conscious, active monkeys (Sereno and Lehky, 2011). It is shown that population encoding in LIP is able to perfectly reconstruct stimulus locations within a low-dimensional representation. On the other hand, AIT neurons provide less accurate low-dimensional reconstructions of stimulus locations despite often demonstrating spatially selective responses. It is argued that the dorsal processing may be egocentric, using a variety of body based 'coordinate' spatial frames of representation, while ventral processing is more allocentric, providing a 'categorical' spatial representation required for identifying objects in scenes (e.g. to the right of, above, connected to).

The types of responses in IT cortex are heterogeneous and varied. In considering various experimental results, it is clear that some neurons in this region may respond with location invariance, while others will encode information about location. Likewise, some neurons may learn to respond with view invariance, but others will only respond to a particular view of a stimulus. These different types of transform specific, or invariant, responses both rely on the local configuration of the feature parts that together comprise the stimulus. It is not only the presence, but also the spatial arrangement of features that is important when determining the current view of a stimulus. For example, if the spatial arrangement of the stimulus features is incorrect, the stimulus may appear as if it is presented from a different view point, or alternatively it may appear to be an entirely different stimulus altogether. The manner with which such a feature space is encoded is crucial for understanding how objects are

recognised and represented within the primate visual cortex.

Freiwald et al. (2009) have shown that neurons in the middle face patch of the macaque temporal lobe respond well to simple cartoon faces. They explored the effect of varying the different combinations of features that together comprised different cartoon faces. The results of their study reveal that neurons in this cortical region do not require the presence of a whole face and will respond to different face parts and interactions between parts (eyes, nose, mouth, etc.). Interestingly, they showed that neurons demonstrated a preference for face aspect ratio over both internal and external facial features. The tuning profiles of these cells was ramp-like, that is, neurons responded vigorously to a particular extreme and did not respond well to the opposite extreme. For example, the face's inter-eye distances ranged from almost cyclopean to abutting the edges of the face, and some neurons would respond maximally for one extreme, but hardly respond to the other.

It has been made clear that the pursuit of invariant cells will only reveal part of the full story for how IT neurons encode objects in the visual world. However, the experimental results presented above also indicate that even an attempt to characterise IT cells by the 'critical features' to which they best respond would also be incomplete. Instead, consideration must be given to individual cell tuning preferences to specific features - bearing in mind that a minimal response is equally important to encoding stimulus features as a maximal one - and feature interactions.

The work by Freiwald et al. (2009) investigated responses of neurons located in the middle face patch of the macaque temporal lobe. Additional research has shown that the degree of view invariance (to faces) is dependent upon the specific anatomical area of the face patch, such as medial lateral (ML) compared to anterior medial (AM) (Freiwald and Tsao, 2010). More specifically, face patches differed qualitatively in how they represented identity across head orientations whereby neurons in ML were

view specific while neurons in AM were more view invariant. Accordingly, careful consideration must be given to the types of encoding that form at all stages of the ventral visual pathway.

The computational experiments presented below explore the types of neural representations that form in the scaled-up version of the VisNet model, comprising four self-organising map layers. The primary aim of these experiments is to investigate the continuous mapping of different face views, as well as the ability to encode view specific responses with location invariance. Both types of neural behaviour are reported by the *in vivo* experimental research of Wang et al. (1998). The results are presented in the context of the original experimental findings, and consideration is given to the type of neural representations that form, and the type of information that is encoded. It is shown that VisNet is successfully able to develop neurons that respond comparably to those found in the experimental work of Wang et al. (1998). However, the types of encoding that develop throughout the layers of the VisNet model do not compare well to the results of Freiwald et al. (2009). Future computational research is suggested, with the goal of reconciling these differences.

6.2 Method

As previously outlined, the simulations within this chapter make use of the scaled-up version of the VisNet model architecture. This updated model was presented for the first time in the previous chapter. A general overview of the model architecture is provided next, outlining some of the specific differences between the scaled-up version of the VisNet model and the previous version.

6.2.1 Model Architecture and Parameters

The scaled-up version of the VisNet model now has a 512×512 pixel retina, where each pixel corresponds to a particular retinal location. At each retinal location there are a bank of Gabor filtered outputs. The filtering procedures are outlined in Section 5.5.1. Furthermore, within each processing layer of the VisNet model there are now 128×128 neurons. The manner with which the VisNet model is initialised is identical to the previous version, whereby the synaptic connections between layers are randomly created. The initial weight of these synapses is also set at random, bound between 0 and 1. Because the network itself is larger, the number of connections between layers and the sampling radius were also increased. The parameters that govern these principles are presented in Table 6.1.

	Dimensions	Number of Connections	Radius
Layer 4	128x128	400	65
Layer 3	128x128	400	24
Layer 2	128x128	400	9
Layer 1	128x128	1000	6
Retina	512x512x32	-	-

Table 6.1: Network dimensions. The number of connections per neuron and the radius in the preceding layer from which 67% of the total number of connections are received.

The sparseness percentile and slope values are presented in Table 6.2. Instead of a competitive network, a SOM network architecture is introduced within each of the four layers of the model. These simulations mark the first time that a trace learning rule and SOM network architecture have both been present within VisNet. The lateral inhibition and excitation parameters are provided in Table 6.3. For all simulations, a learning rate of $k = 0.01$ was implemented.

Layer	1	2	3	4
Percentile	95.0	95.0	95.0	95.0
Slope β	190	40	75	26

Table 6.2: Sigmoid parameters

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

Table 6.3: Lateral inhibition and excitation parameters for the SOM.

6.2.2 Trace Learning Rule

A trace learning rule (Foldiak, 1991) is implemented in each of the following simulations. This is instead of a simple Hebbian learning rule. The trace learning rule is a modified Hebb-type rule and is outlined in Section 1.3.4 and presented in Equations 1.8 and 1.9. Trace learning was developed on the principle that, in the natural visual environment, changes in viewing conditions are more common than changes in object identity. Accordingly, the cortex can use a temporal trace of previous cell activity to achieve invariance by associating together visual inputs that appear close together in time.

6.2.3 Stimuli

The face stimuli used in the simulations presented as part of this chapter are comparable to those implemented in the follow-up study of Chapter 5. The stimuli were created using the FaceGen software package by Singular Inversions. This software allows for the control of over 150 facial characteristics, including age, race, gender. For the purposes of these simulations, a random face was generated within conservative bounds. A frontal view of this face is presented in Figure 6.1 and forms the basis of the other training stimuli presented within each of the studies.

All stimuli are filtered in accord with the updated filtering routines (Equation 5.1) outlined in Chapter 5, Section 5.5.1. Prior to filtering, each image is desaturated and saved as a 256 grey scale image on a 512×512 pixel background.

Figure 6.1: Frontal face view. This figure shows the frontal face view of the stimulus used to train and test VisNet during the simulations presented below. The face stimulus was created in FaceGen by Singular Inversions and is a 2D image export of a 3D model. From this 3D model, different rotational views can be exported as 2D images. This specific example is also the facial view referred to as view 90.

Figure 6.2: Face stimulus rotating in depth. This figure presents the same face stimulus rotating in depth around the vertical axis. In total, 180 face stimuli are created as the face rotates over 180 degrees. The central face presented as part of this figure is a complete frontal view, identical to that presented in Figure 6.1. It is 90th view from the complete 180 image sequence.

6.3 Continuous Mapping of Different Face Views

Past research has shown that neural activation is consistent for different views of the same face, but the centre of the neural activity shifts along the cortex in one direction with the rotation of the face (Wang et al., 1998). The first set of simulations presented below are designed to investigate the performance of VisNet in the context of these results. More specifically, now that the VisNet model is larger both in terms of retinal input and processing layers, it is hypothesised that a SOM network architecture will help build a continuous map of neuronal activity similar to that revealed by *in vivo* optical imaging (Wang et al., 1998).

6.3.1 Method

Stimuli

Figure 6.2 shows the face stimuli used to train and test the network during the following simulations. This stimulus rotates in depth around the vertical axis and a total of 180 images are created, one for each degree of rotation. The rotating head is designed to occupy as much of the retina as possible, and is located centrally on the retina.

Training Procedure

During training, each of the 180 images is presented to the network one at a time. Each of the four SOM layers is trained in turn starting with the first layer, for 50, 100, 100, and 75 epochs, respectively. As with the previous version of VisNet, upon image presentation the activation level of individual neurons within a layer are calculated, then their firing rates are calculated, and the feedforward synaptic weights between layers w_{ij} are updated according to Equation 1.8, where the trace $\bar{y}^{\tau}$ is updated according to Equation 1.9. The presentation of all 180 images constitutes one training epoch. The learning rate of the model was set to $k = 0.01$ and sparseness was set to 0.05. Other parameters were explored but had no quantitative effect upon the simulation results and are not reported in this chapter.

Testing Procedure

The testing procedure is similar to the training procedure. During testing, all 180 images are presented to the network. However, learning does not take place during testing and the synaptic weights are not updated. Upon each individual image presentation, the firing rates of the VisNet neurons are recorded from each layer of the model. It is typical to produce polar plots for simulations with rotating stimuli. However, during the simulations the stimulus does not complete a full rotation because only 180 views are presented and therefore the cell responses are presented as line plots.

6.3.2 VisNet Simulation Results

Having trained and tested the model as outlined above, the neurons in the output layer have learned to respond in a number of different ways. Instead of producing only invariant cells, the responses are heterogeneous, whereby some cells learned to respond to specific views of the face stimulus. This

modelling result reflects the *in vivo* optical imaging results of Wang et al. (1998). Perhaps more importantly, the output neurons were organised such that neurons with similar response profiles were positioned close to one another. This organisation is typical when a SOM network architecture is implemented. Across the output layer of the model, there was a continuous mapping of different face views. More specifically, a gradual shift of the centre of activation in one direction along the output layer is observed in response to the rotation of the face stimulus. Initially, neurons demonstrate relatively view specific response profiles, having learned to respond preferentially to a contiguous portion of the view space. However, the centre of the activity shifts along the output layer creating overlapping response profiles. Neurons learned to become more invariant, before eventually becoming view specific again.

Figure 6.3 exemplifies these results and presents an 80×80 portion of the output layer. The x-axis and y-axis mark the cell co-ordinates within this square region. There are five different colour activation maps that form elongated rings on the output layer: blue; cyan; yellow; orange; and brown. Each activation map is produced by cells that respond maximally with specific view preferences. More specifically, the total 180 ° view space was divided into five equal parts: 1 ° through 36 °; 37 ° through 72 °; 73 ° through 108 °; 109 ° through 144 °; and 145 ° through 180 °. All neurons inside the blue activation map have learned to respond maximally to all of the first 36 images of the view space, 1 ° through 36 °. All those inside the cyan have learned to respond maximally to all views 37 ° to 72 °. Those inside the yellow respond maximally to 73 ° through 108 °. Those inside the orange respond maximally to 109 ° through 144 °. Those inside the brown respond maximally to 145 ° through 180 °. The activation maps overlap, reflecting the fact that some neurons develop broader response profiles and respond to larger contiguous portions of the view space. Accordingly, some neurons fall within

all five activation maps, reflecting the fact that they respond to all viewing angles and are invariant. The different types of cell responses are presented in Figure 6.4.

Figure 6.4 shows four prototypical cell responses. Cell 42, 50 responds to all 180 views that together comprise the complete view space. However, other cells respond to contiguous regions of the view space, exemplified by cells 30, 71; 51, 52; and 41, 9. These view specific neurons only fire when the face is presented from a specific view point and as such, their firing provides information regarding the angle at which the face is presented on the retina.

6.4 Location Invariance with View Specificity

Wang et al. (1998) showed that the same region of the anterior part of the inferotemporal cortex (cytoarchitechtonic area TE) responds to a specific view of a face regardless of the retinal location. Their experimental results revealed that 91.8% of the activated regions was common to three different retinal locations separated by 10°. The following simulations explore the performance of VisNet in a similar paradigm. In particular, special consideration is given to the differences in training procedure, which is shown to have a notable impact on network performance. These variations receive comment in the discussion.

6.4.1 Method

Stimuli

During the following simulations, two different training procedures are implemented in order to explore the effect each has on the network's ability to self-organise during training. The primary difference between the two training procedures is the stimuli. Although both procedures make use of the

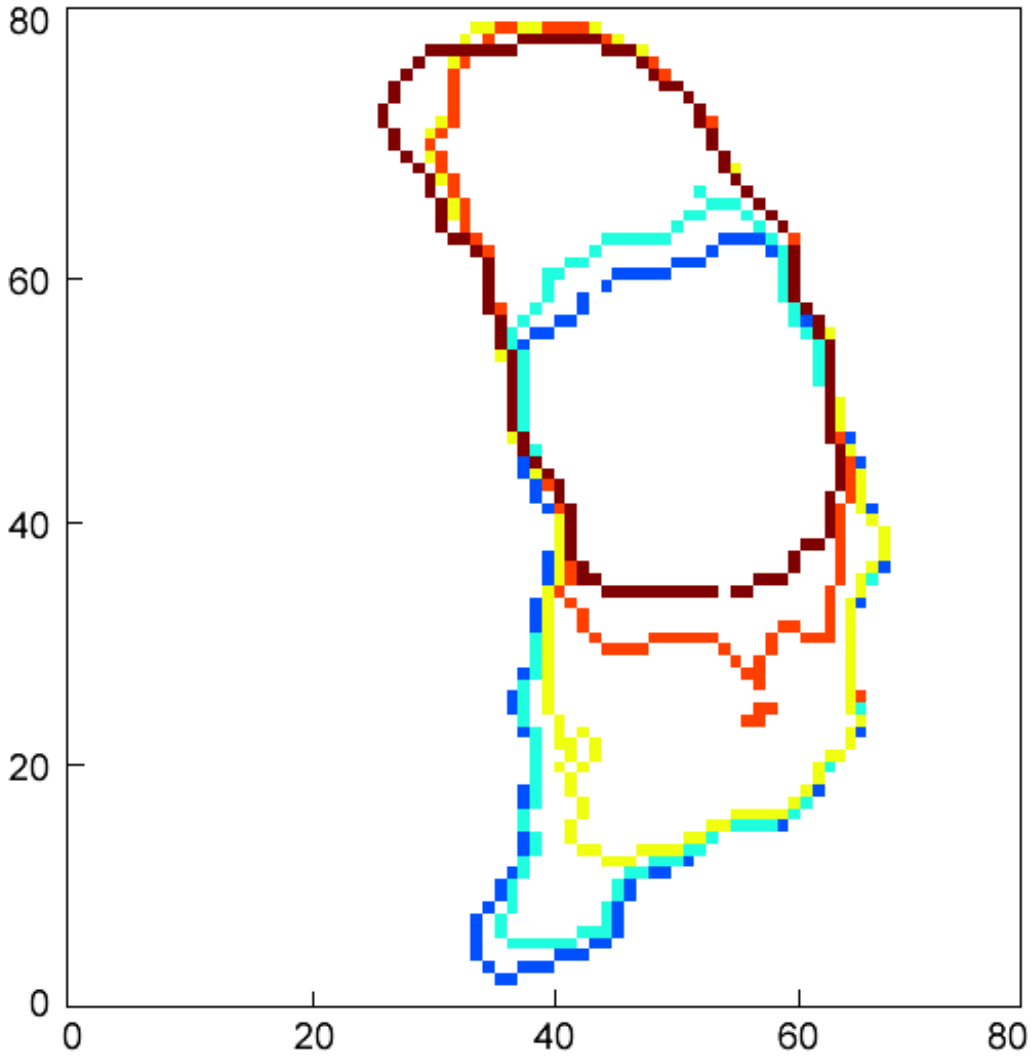

Figure 6.3: Continuous mapping of different face views. This figure presents firing rates in an 80×80 portion of the VisNet output layer. There are five different colour activation maps that form elongated rings on the output layer: blue; cyan; yellow; orange; and brown. Cells inside each ring respond with certain view preferences; blue, 1 ° through 36 °; cyan, 37 ° through 72 °; yellow, 73 ° through 108 °; orange, 109 ° through 144 °; and brown, 145 ° through 180 °. The x-axis and y-axis mark the cell co-ordinates within this square region.

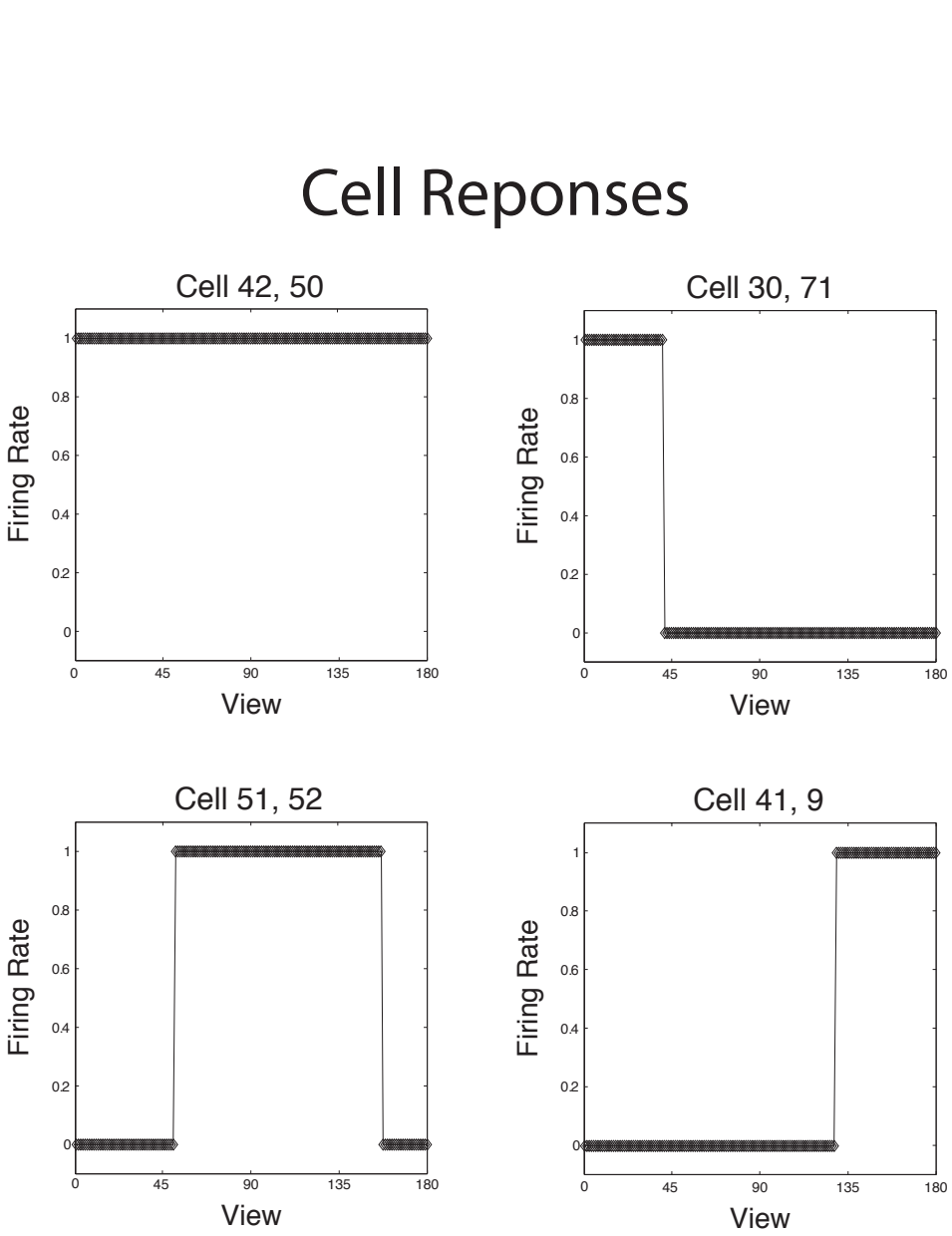

Figure 6.4: Cell response plots. This figure presents the cell response plots of four different prototypical output cells. It can be seen that cell 42, 50 responds invariantly with respect to the viewing angle of the face, while cells 30, 71; 51, 52; and 41,9 all respond with view specificity to contiguous portions of the view space.

same face views as previously presented in Figure 6.1 and Figure 6.2, the location at which the face is presented on the retina varies.

Training Procedures

Similar to the previously presented simulations, the training procedure comprises the presentation of all 180 images of the face as it rotates through 180°. However, the rotating face is now presented in four different retinal locations during training. Two training procedures are implemented, which vary only with respect to the four different locations used. Both training procedures use a trace learning rule to update the synaptic weights between layers, identical to the previous simulations. Prior to the presentation of a new face view, the trace is reset.

Also common to both training procedures is the manner in which the face images are presented. The first view of the face is presented in the top left of the four possible locations. It is then shifted clockwise through each of the three remaining locations. Accordingly, the same view of the face is presented in each of the four possible locations in turn. Once complete, the next view of the face is presented in the same way. This process is repeated until all 180 views of the face have been presented in this manner, and together this completes one training epoch.

When an image is presented to the network, the activation levels of individual neurons within a layer are calculated followed by their firing rates. The feedforward synaptic weights between layers w_{ij} are updated according to the trace learning rule already outlined. This process is repeated for each of the four layers in turn. The network is trained one layer at a time starting with layer 1 and finishing with layer 4, for 50, 100, 100, and 75 epochs, respectively.

The first training procedure presents the face stimuli in four locations that vary by a few pixels. Figure 6.5 shows each location in relation to the other different locations. At each image location,

a marker is annotated upon each section of the figure designating where all four possible locations reside. The opacity of each face is lowered to 50% to allow the annotations to display clearly. It can be seen that each location is offset either horizontally and/or vertically from the others by 10 pixels. Accordingly, there is overlap between the faces as they move between the four possible locations. The second training procedure presents the face stimuli in four locations that are orthogonal with respect to one another. Figure 6.6 shows that each rotating face occupies a different quadrant of the retina.

Testing Procedures

The testing procedures for the two different training procedures are identical. The images used to train the model are presented in the same manner during the testing procedure. However, synaptic weights are not updated during testing. The firing rates from each neuron within the model are recorded in response to each image presentation.

6.4.2 VisNet Simulation Results

First, the results obtained from the first training and testing procedure are presented. Figure 6.7 shows prototypical output cell response plots after testing with an arbitrary facial image, view 90 (frontal face view), when presented in all four possible locations. This cell has learned to respond invariantly with respect to location, that is, it has learned to respond to the same facial view in all four locations.

Figure 6.8 shows response plots produced by the same cell when presented with the head rotating through all 180° in all four locations. It is clear that this prototypical output cell has learned to respond to approximately identical contiguous portions of the facial view space regardless of the location within which the face is presented. There are similar cells that have learned to respond to different portions of the view space such that the entire view space is represented in this way. For

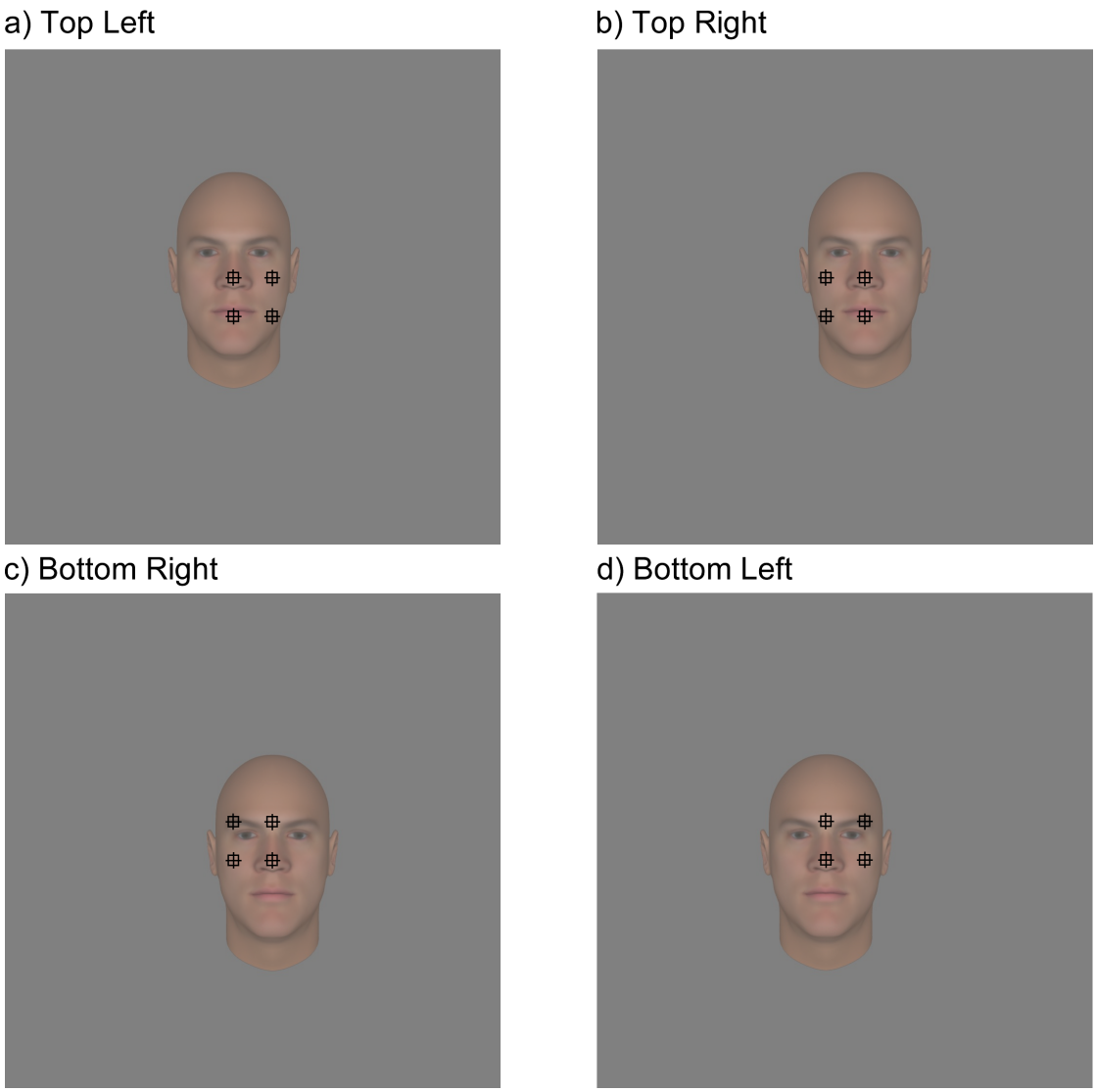

Figure 6.5: First training procedure. This figure shows the four possible locations with annotated points marking the centre of each. Each location varies by 10 pixels in the horizontal and/or vertical axis. The figure comprises four images: a) top left; b) top right; c) bottom right; and d) bottom left. Each image contains the markers designating the centre point of all four image locations. The faces are presented with 50% opacity to allow the central markers to display clearly. During training, all 180 views of the face are presented at each location, every view being presented at all four locations consecutively.

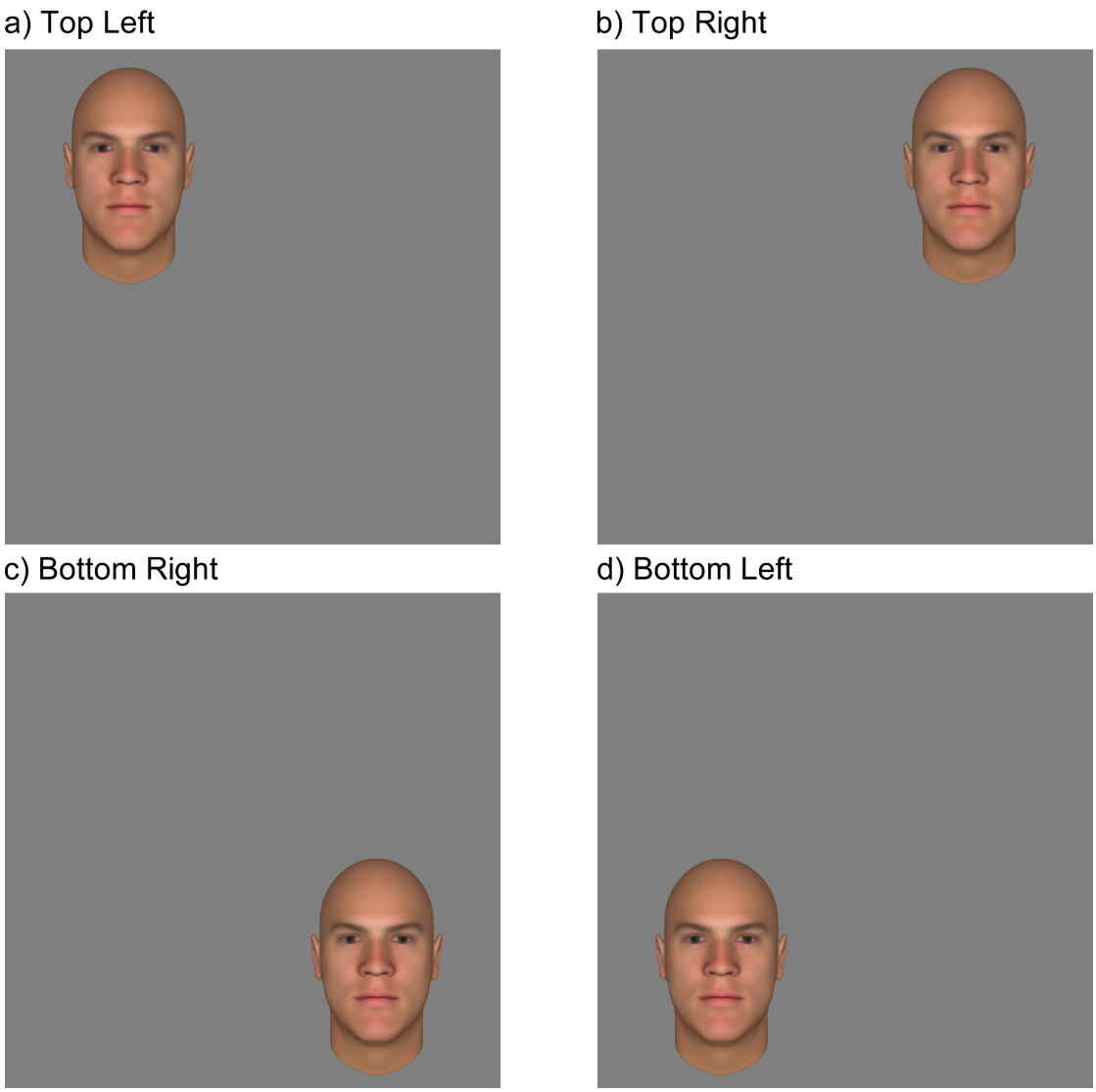

Figure 6.6: Second training procedure. This figure shows the four possible locations, which are orthogonal with respect to one another: a) top left; b) top right; c) bottom right; and d) bottom left. Training conventions as in the first training procedure.

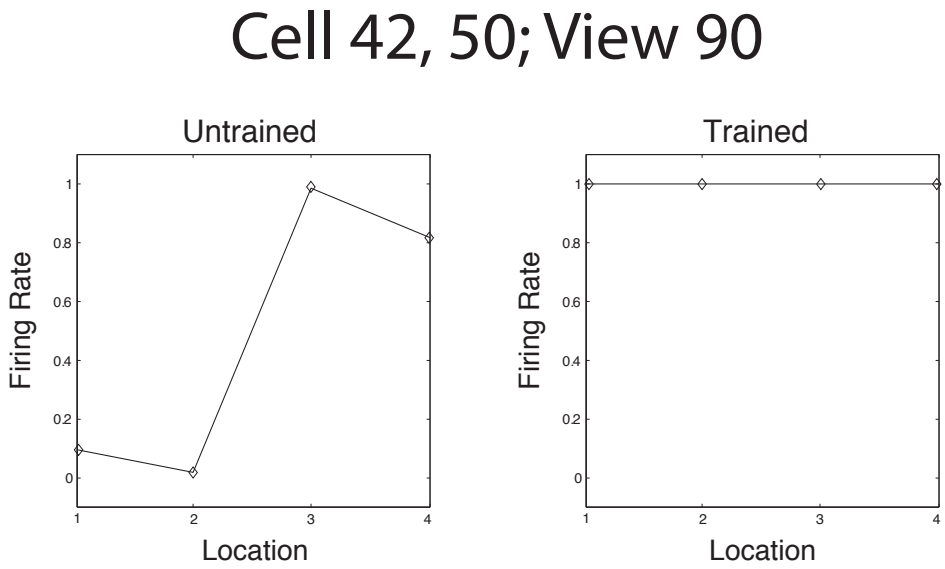

Figure 6.7: Location invariant, cell response plots. This figure presents the response of cell 42, 50 both before and after training having been tested with view 90 of the face stimulus. It is shown that prior to training, the cell does not respond with location invariance, instead responding maximally to the face view when it is presented in one of the four locations. However, after training the cell has learned to respond invariantly, that is, it responds maximally to view 90 of the face stimulus regardless of the location within which it is presented.

comparison, Figure 6.9 presents the untrained responses from cell 42, 50 when the face stimuli are presented in the top left location during testing rotating through all 180°.

Results obtained during the second training and testing procedure are presented below. Figure 6.10 shows the output cell response for Cell 32, 30 to view 90 of the face stimulus across all four locations. It can be seen that after training the output neuron has failed to learn to respond to the same view in all four different locations. The untrained network is presented for comparison. It was generally found that output layer cells learned to respond to only one location.

At individual locations, the output cells that had learned to respond to that location did so in an identical manner to those neurons presented in Figure 6.4 but with slightly broader response profiles. Neurons learned to respond to contiguous portions of the view space. However, due to a combination of the topographical connectivity between layers and the SOM architecture within a layer, the output

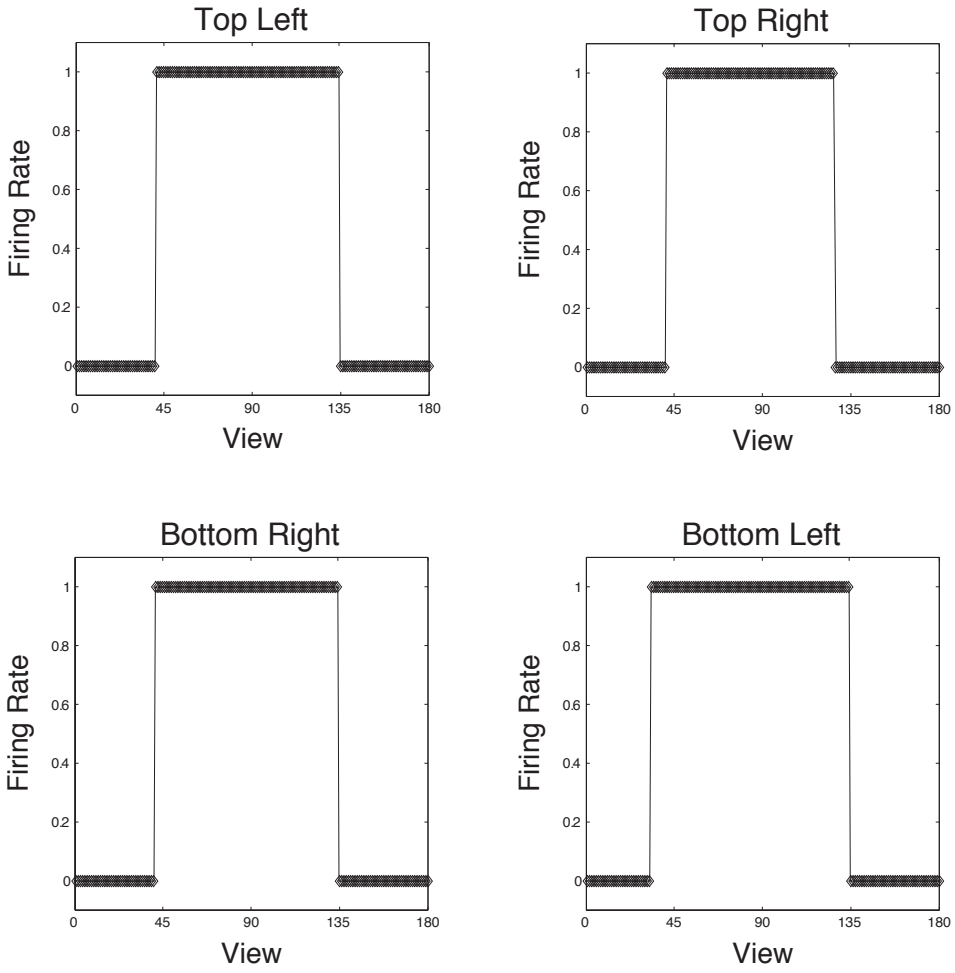

Figure 6.8: View specific, location invariant, cell responses plots. This figure presents four response plots for cell 42, 50. Each response plot reflects one of the four possible locations within which the face stimuli could be presented. The x-axis on each response plot measures the viewing angle from which the face is presented, while the y-axis measures the cell's firing rate. It can be seen that the cell responds comparably to the face stimuli regardless of which position within which they are presented. More specifically, cell 42, 50 approximately responds to views 40 through 135 of the face as it rotates in depth irrespective of location.

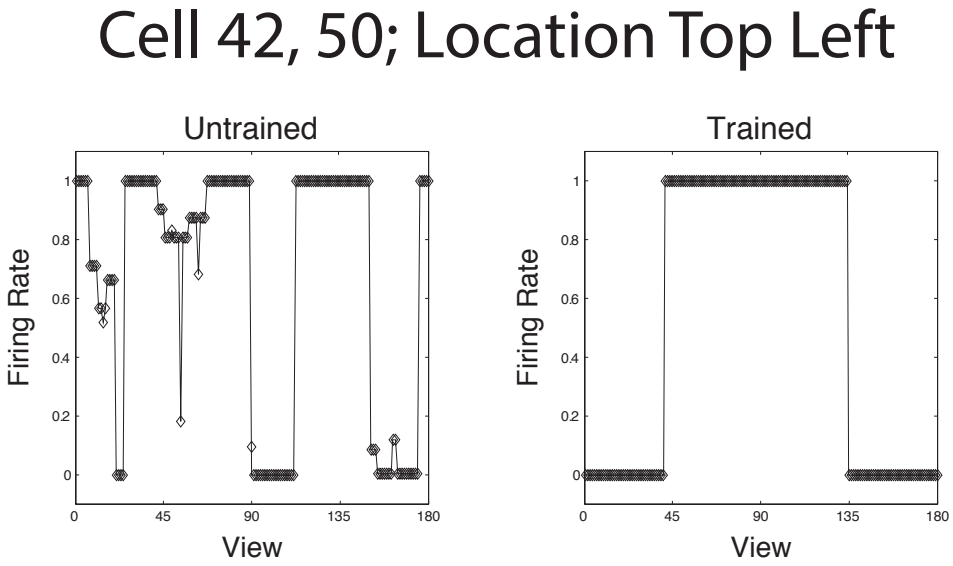

Figure 6.9: Untrained comparison, cell response plot. This figure presents the responses of cell 42, 50 when presented with the face stimulus rotating in depth in the top left of the four possible locations. It is shown that prior to training, this cell responds sporadically to different views of the rotating face. However, after training, cell 42, 50 has learned to respond to a contiguous region of the view space, and does not respond to any of the other views outside of its preferred contiguous view space.

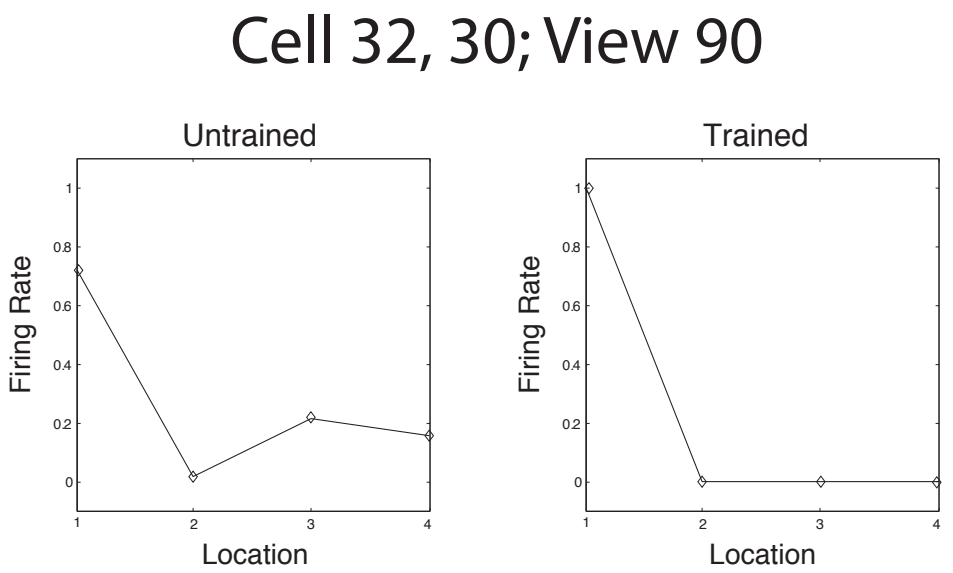

Figure 6.10: Orthogonal locations, cell response plots. This figure presents the cell response plots of cell 32, 30 in response to view 90 of the face stimuli. The network has been trained with the face rotating in depth across four orthogonal locations. It can be seen, prior to training, that the cell does not respond maximally, but produces residual firing to all four locations. After training, the cell has failed to learn to respond to the same view of the face stimulus regardless of location. Instead, the cell has only learned to respond to the face stimulus when it is presented in location 1 (top left).

layer produced four activation maps that directly reflect the location of the stimuli presented on the retina during testing. Accordingly, the output layer comprises four separate maps, each of which corresponds to one of the four separate locations within which the face stimuli are presented during training (Figure 6.11). Each of the activation maps comprises a heterogeneous population of cells that respond preferentially to specific views of the face stimuli, and in some cases all the views. In summary, the four different maps (one for each location) have developed because the network has failed to learn to develop location invariant responses, and instead the VisNet output neurons have only learned to respond with view specificity.

Together, the results show that the face stimuli need to be trained continuously over the retina in order for the experimentally observed view specific, location invariant, cell responses to develop during training. When the face stimuli are not trained continuously, but are trained orthogonally in different retinal locations, separate maps form in the output layer of the VisNet model. In this case, output neurons fail to learn to develop location invariant response properties.

6.5 Discussion

The simulations presented as part of this research chapter demonstrate that the VisNet model develops comparable results to the *in vivo* optical imaging studies presented by Wang et al. (1998). The first set of simulations presented above show that the scaled-up version of VisNet develops neurons that learned to respond to specific views, as well as to all the views of the rotating face. The second set of simulations showed that output neurons were able to develop view specific and location invariant responses. The particular success of these results was dependent upon the specific training paradigm.

The results from the first set of simulations investigating the continuous mapping of different

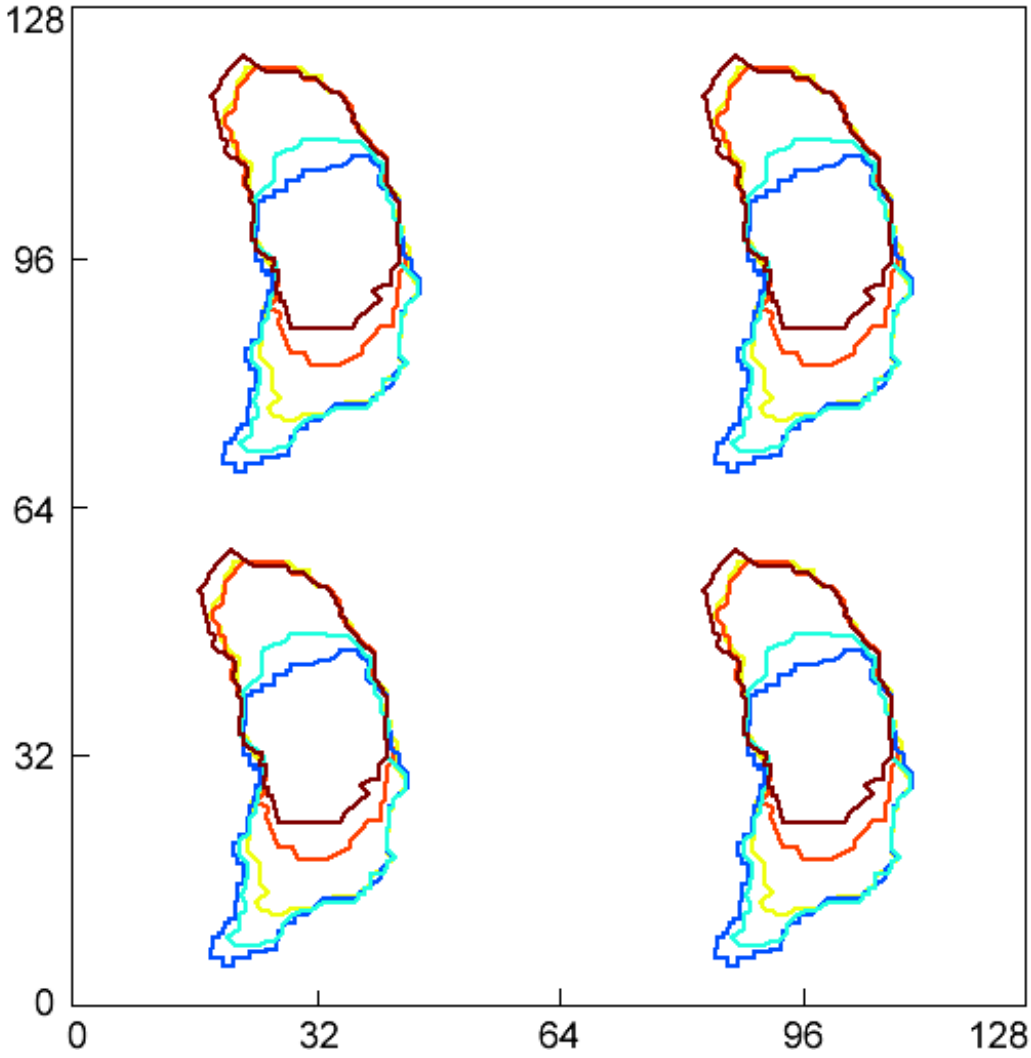

Figure 6.11: Output layer response maps. The output layer comprises four separate response maps, each of which corresponds to one of the four separate locations within which the face stimuli are presented during training. Neurons within each map demonstrate heterogeneous response properties, very similar to those presented in Figure 6.3. Within each location specific map, some neurons respond to specific views of the rotating face, while other neurons learn to respond to all views of the rotating face.

face views are markedly similar to the optical imaging study by Wang et al. (1998). This is the first time that a heterogeneous population of cells has been produced with VisNet accurately reflecting *in vivo* experimental research. Although a trace learning rule was implemented during training, the CT learning mechanism could still operate due to the spatial overlap between the views of the rotating face. Accordingly, with both a trace rule and CT learning mechanism in operation, the VisNet output layer may have been expected to produce only view invariant responses. The results show that the SOM network architecture was still able to force a map-like output representation in spite of this.

It should be noted that previous VisNet research investigating CT learning has produced results whereby output neurons learned to respond to different contiguous blocks of a view space (Stringer et al., 2006). However, the fragmentation of the view space was primarily due to the catastrophically different views of the object presented during rotation. This prevented the CT learning effect from binding together the different views of the same object. It is shown here for the first time that a genuine activation map has formed in the output layer of the scaled-up VisNet model. There are a number of important considerations that may explain why this has happened.

Firstly, the VisNet retina is 16 times larger than the original VisNet model. This provides a far greater resolution at which the training and testing images are presented. The increase in resolution affords more fine grained detail within each image, and this additional detail can be used by the processing layers of the model. Secondly, each processing layer comprises 16 times as many neurons. Although it is not necessarily clear that more neurons within each layer in itself will produce qualitatively different results when compared to fewer neurons, neurons with less common responses will be more frequent in absolute terms, allowing for a more robust analysis. Finally, these two previous considerations may have a qualitative impact on the performance of the SOM network architecture.

A SOM architecture is expected to form activation maps, but previous results presented as part of Chapter 4 within this thesis reveal that this is not always the case. The type of output representation produced by a SOM is dependent upon more than just the excitatory and inhibitory connections within a layer. Other parameters such as the level of global competition are important. More specifically, the emergent behaviour produced by the increased number of competing neurons within each layer may be different to that with fewer neurons.

The activation map observed in the output layer of the VisNet model during the first set of simulations is compelling, especially in the context of experimental results. However, evidence for the existence of a SOM as implemented within the VisNet model requires further investigation. The implementation of a SOM network architecture is motivated by the categoric feature maps observed in IT cortex. Similarly, SOMs have been particularly successful in the LISSOM (laterally interconnected synergetically self-organising map) model. More specifically, the self-organisation of ocular dominance columns similar to those found in V1 (Sirosh and Miikkulainen, 1994), orientation and direction preference maps, and also long range excitatory collaterals (Miikkulainen et al., 2005) resembling those recorded *in vivo* (Bosking et al., 1997) are all encouraging. However, there is evidence that the short range lateral connections within V1 are actually inhibitory and not excitatory (Tucker and Katz, 2003). Furthermore, Bosking et al. (1997) show that longer range connections may actually connect areas with similar response preferences. Although these experimental studies were not performed with primates, it is clear that the precise connectivity that underpins the formation of maps in the visual cortex requires further investigation.

It should be noted that previous single-cell recording studies have shown that non-face objects are not represented by single groups of cells (Kobatake and Tanaka, 1994). Instead, a combination of

multiple groups of cells were found, representing partial features of the object images. However, the modelling results presented within this chapter should generalise well to any type of object, and not just to faces. Therefore, further consideration must be given to the differences between face stimuli and non-face stimuli, specifically how the different types of representations form during learning.

The results presented above also demonstrate that VisNet is able to develop neurons with location invariant responses that are also view specific. It was shown that given appropriate training, output neurons will learn to respond to specific views of a rotating face regardless of the location within which the face views are presented. These results are comparable to those of Wang et al. (1998) who showed that neurons in TE learned to respond with location invariance to a specific (frontal) view of the face. However, the specific training paradigm implemented during learning had a significant qualitative impact on the results. More specifically, the success of this result requires the different locations to have a relatively high degree of overlap. It was shown that when the different locations were completely orthogonal, output neurons failed to develop location invariant responses. These neurons learned to respond with view specificity, but they also responded to only one of the four locations. Consideration is now given to the difference between these two results.

In order for VisNet output neurons to develop view specific responses that are also location invariant, a single activation map must form in the early layers of the model, and persist throughout each subsequent layer. It has been shown that this is challenging when the locations are orthogonal with respect to one another and this is due to a number of contributing factors. The SOM network architecture causes neurons with similar response profiles to develop close together. The different VisNet processing layers are aligned directly above one another, as is made clear in Figure 1.3. Due to the topographical connectivity between layers, neurons in a receiving layer are more likely to become

active directly above those active neurons in the preceding layer. This produces separate activation maps, one per location, throughout all four layers of the VisNet model. This occurs in spite of an increase in the size of the connectivity radius (from which 67% of connections from the preceding layer are received) in the last two layers of the VisNet model. In summary, the activation maps are kept separate due to a combination of the SOM architecture, topographical connectivity and global competition.

These results are different to those that would be obtained with a traditional competitive network architecture. A competitive network, as implemented in past VisNet research, will not produce cells that cluster together with similar response properties unless the topographical connectivity between layers is extremely strong. To exemplify this point, consider a fully-connected network produced without topographical connectivity. Past VisNet research implements a diluted connectivity similar to that used in the simulations presented within this research chapter. However, in a fully-connected network neurons would not exhibit any form of organisational structure in terms of their response preferences. Accordingly, VisNet output neurons would receive connections from any part of the preceding layer with equal probability and could learn to become location invariant. With that said, the network would suffer from a CT learning effect and bind together the different views of the face as it rotates in depth. This may ultimately produce neurons that do not convey any information about view or location, because they are completely invariant with respect to both.

The introduction of the SOM architecture is responsible for producing activation maps that comprise neurons with view specific responses. As previously mentioned, the SOM can have the effect of preventing cells with *only* invariant responses from forming. Instead, it is able to 'spread out' activity into maps. However, it has already been discussed that if the locations of objects are too far apart, the

output layer neurons of the VisNet model are unable to bind together views presented in the different locations. Successful results were only obtained when the different locations have significant overlap, helping to produce a single activation map that persists throughout each of the VisNet processing layers.

6.5.1 Future work

It has been shown that neurons in the primate visual system respond heterogeneously to visual stimuli. The results presented as part of this research chapter successfully model results produced during *in vivo* optical imaging studies (Wang et al., 1998). However, more research must be conducted in order to establish the exact nature of the VisNet neural responses and the information that they encode. Past experimental research has shown that IT neurons respond to interesting facial feature characteristics, such as their relative aspect ratio (Freiwald et al., 2009). More specifically, neurons respond to different face parts and interactions between parts (eyes, nose, mouth, etc.). They appear to encode a 'categorical' spatial representation (above, next to, etc.), useful for object recognition within a scene (Sereno and Lehky, 2011).

In order to investigate the exact nature of neural encoding within the VisNet model, a different approach to training must be undertaken. For example, a large range of different frontal face views could be presented to the model. This may include many hundreds of genuinely different faces. Testing will comprise of different combinations of prototypical facial features and their combinations, similar to the systematic variations explored by Sereno and Lehky (2011). As has been made clear, careful consideration must then be given to the types of neural representations that form in each of the VisNet processing layers. In order to analyse these neural responses, similar techniques to those used by Sereno and Lehky (2011) may be employed, more specifically, a population-based statistical

approach (namely, multidimensional scaling). Accordingly, a direct comparison can be made between the neural responses of the VisNet model and the responses recorded during the experimental research. It will be interesting to establish exactly how neurons within the different layers respond, and whether they are similar to the experimental results. For example, do they encode information such as the relative aspect ratio of facial characteristics? It will then be possible to systematically explore how these neural representations form, and why this is the preferred method of self-organisation.

Chapter 7

Conclusion

How does the brain build visual representations of individual objects even when multiple objects are present within a scene during learning? This important question must be addressed when attempting to understand natural visual scene processing. Over successive stages, the primate ventral visual system develops neurons that respond with view, size and position invariance to objects including faces. A major challenge is to explain how invariant representations of individual objects could develop when visual input is derived from environments containing multiple objects.

This doctoral thesis presents progress in understanding how neurons in the output layer of the VisNet model could learn to respond to individual objects, despite the fact that objects are never presented in isolation during training. By more accurately simulating the statistics of natural environments, it was shown that the statistics of how objects can occur in different combinations, on different occasions, can play a useful role during learning. Furthermore, it was shown how the independent motion of just two different objects may also be utilised by the VisNet architecture to form separate representations of individual objects always presented together to VisNet. Consideration has also been given to the types of neural responses that develop in the VisNet model, and the different types of

object-feature encoding that develop during training. This latter work is an important area for future research.

Accordingly, the overall objective of the research presented in this thesis has been;

- Ecological Validity: Increase the complexity of the VisNet training environment to more accurately reflect the scenes occurring in natural vision.
- Architectural and Learning Mechanisms: Investigate how combinations of CT and trace learning rules can operate within a SOM architecture.
- Neural Responses: Consider the different types of neural encoding that develop during training; output neurons are not always invariant, but carry information about specific transforms.

Previous research with VisNet has focussed on the presentation of one object at a time to the VisNet model during learning. There are a number of limitations to this approach which have received discussion in the various chapters of this thesis. For example, in natural vision, individual objects are rarely presented in isolation and the primate visual system has to learn about each object in complex environments containing multiple objects. Furthermore, previous VisNet research has not considered what individual neurons in the higher layers of VisNet might encode about the shape or form of the objects that they represent. Are neurons in the higher layers of the ventral pathway encoding the visual form of the objects by some means, perhaps using a form of distributed coding across a population of cells? The remainder of this chapter will discuss the strengths and limitations of the research presented in this thesis as a whole, and will suggest potential directions for future research based on the results previously presented.

The first two chapters of this doctoral thesis extend previous research by Stringer et al. (2007) and by Stringer and Rolls (2008). It is shown that statistical decoupling of features belonging to different

objects enables VisNet to learn independent, separate representations of each object presented during training despite the fact that objects were always presented as part of a pairing. Previous research had failed to explore the extent to which VisNet must be trained with object pairings in order to learn object specific representations. An important contribution of the first research chapter was to show that the amount of training necessary to achieve this decoupling effect is reasonably small.

Presenting object pairs to the network during training is an improvement over previous VisNet research with single objects. The natural visual world comprises many objects, and together these objects produce a scene. If we are to better understand, through computational modelling, how the primate visual system is able to learn about individual objects in the natural visual world, then our models must make use of realistic training environments.

A notable limitation of the particular approach described within this thesis is that two objects together do not make up a scene. A real scene may comprise numerous different objects, some that are always seen together or rarely seen together, some that move with respect to one another or tend to always move in the same way. The difference between presenting just one object to the network during learning, versus presenting two objects, may appear slight. However, the addition of a second object during training, as opposed to just a 'noisy background', is in fact essential if one wishes to study the emergent behaviour of a model that learns using the statistics of the natural visual world; for example, it is only with two or more objects that a partial occlusion such as that presented in Chapter 3 can be performed.

An important question that must be addressed in future research is how Visnet can learn to use the statistics of the visual environment when more than two objects are presented translating across the retina during an independent motion training paradigm. More specifically, the results of the pilot

study in Section 5.3 suggest that this is more challenging than expected, given the results of the first two research chapters. If this mechanism is to exist in the primate visual system as a method for learning about individual objects, then it must perform successfully in an environment where more than just pairs of objects are presented translating during learning.

The results presented in the third research chapter suggest that independent motion may help play an important role when object selective representations form during learning. When there are many different object pairings, and thus a potentially large amount of statistical decoupling between the individual objects, then object selective responses can form when different objects are presented with one another during learning. However, even though relatively few pairings are necessary for this decoupling effect to occur, if there are only two objects presented during training, then the output neurons of the VisNet model will be unable to differentiate one object from another. Accordingly, a novel solution is necessary to address this challenge. Independent motion of two objects provides a similar statistical decoupling mechanism to that which is provided by presenting different pairings during learning. However, independent motion does not rely on many different pairings in order to be efficient. Instead, provided the different views belonging to each object do not correlate with one another, then a CT learning effect is able to bind together the views from each object, respectively. The result is that representations of different objects are learned by different output neurons. A limitation of this approach to learning is discussed next.

In natural vision, objects do not always move with respect to one another, nor with respect to the viewer. Sometimes objects are simply static when seen together, and yet in many situations we are still able to learn to recognise the individual objects. A typical situation of this kind might occur when we view novel faces in a photograph. The current VisNet model would not be able to solve such a

problem because it relies on the statistical decoupling of features of different objects that can occur through independent movement in order to tell them apart. However, the OFTNAI laboratory has recently shown that this more difficult problem can be solved using spiking neural network dynamics. In such a model, the emission of individual action potentials is simulated, and the synaptic plasticity can be dependent on the timings of the pre- and post-synaptic spikes (Bi and Poo, 1998). Preliminary results suggest that the learning mechanisms discussed as part of the first two chapters of this thesis are likely to exist in a spiking neural network. As a consequence, the benefits afforded by implementing some form of spiking neuronal model extend the results presented within this doctoral thesis, and do not undermine them.

VisNet output neurons form object selective and invariant responses by virtue of the visual training environment to which they are exposed. Neurons within the VisNet model are able to use the frequency with which object features occur together in order to develop output representations that are object selective and invariant. There is no need for an attentional mechanism during learning for output neurons to develop object selective responses in this manner; the type of learning demonstrated in this thesis is completely passive and very robust. Accordingly, it seems likely that such a robust mechanism for differentiating stimuli, and stimuli features, could exist in the primate visual system. One potential application of this mechanism could be in learning about facial expression separately from facial identity, and vice versa, and this is explored in the fourth research chapter.

Neurophysiological evidence reveals that neurons in different areas of IT show preference to facial expression, independent of facial identity, and vice versa. As already outlined in Chapter 5, neurons in TE respond primarily to facial identity while neurons in the STS primarily respond to facial expression. Perhaps unsurprisingly, neurons located on the lip of the sulcus (TEm) tend to respond to

expression and identity (Hasselmo et al., 1989a). Given that facial expression and facial identity are not presented separately during learning, how is it that the primate visual system is able to develop selective responses to each 'visual space'? The results presented as part of the fourth research chapter extend the work of the previous research chapter and provide a possible solution to this problem. It was shown that due to the statistical independence of the features that make up facial expression and facial identity, VisNet output neurons were able to develop selective responses to contiguous portions of the facial identity view space, while other neurons learned to develop responses to contiguous portions of the facial expression view space. It should be noted that these neurons do not learn to respond invariantly, and nor should they. Instead, these output neurons learned to respond to a small subset of either expression or identity view spaces. If these neurons had learned to respond to the whole of their preferred view space, then there would be no information encoded in their responses regarding the exact facial expression on display (or facial identity, depending on the neurons' preference).

These results are important, not only because they reveal a potential mechanism for learning about facial identity separately from facial expression, but because they reflect a new approach to understanding object recognition with the VisNet model. More specifically, output neurons should still learn stimulus selective responses (e.g. respond only when presented with the facial expression view space, and not the spatial identity space), but neurons that demonstrate invariant response properties are no longer the most preferable. Instead, a heterogenous population of neuronal responses is necessary in order to encode the different facial expressions, or facial identities.

Neurons in the output layer of the VisNet model learned to respond preferentially to a subset of contiguous views from just one visual space. Carefully controlled cartoon faces were presented during training and testing and it was found that this level of control was necessary in order to create the ap-

propriate level of statistical independence required during learning. Additional research must confirm whether this is a limiting factor of the training paradigm. Furthermore, in the case of the cartoon faces, the different view spaces that reflect facial expression and facial identity were highly orthogonal. In natural vision, a real primate face is composed of overlapping visual spaces that together encode facial identity and facial expression.

Consequently, performance of the VisNet model could be explored by presenting to the network hundreds of different realistic faces during training. The different faces should exhibit different facial expressions. In this way, expression and identity are no longer orthogonal with respect to one another, and the training environment is far more rich. This type of training paradigm is similar to that presented as part of the follow-up study (Section 5.5). A relevant limitation to this approach was revealed by these results; when realistic faces are presented to the VisNet model during training, there are a number of facial features that do not often change, regardless of facial identity or facial expression (e.g. eyebrows). As a result, a CT learning effect persists during training and all the different faces are bound together and learned by the same subset of output neurons. Without further simulation, it is difficult to ascertain whether this issue may be alleviated by further improvement to the VisNet retina.

The increase in size of the VisNet retina was a fundamental improvement necessary in order to investigate model performance in response to more realistic face stimuli. A larger retina provides higher resolution and therefore more detail is present within each of the images presented during training. It is shown that neurons in the VisNet processing layers are able to benefit from this finer detail. Previously, a lack of resolution provided by the VisNet retina meant changes in the stimuli must be relatively large and salient if output neurons are to differentiate the stimuli. Now, neurons are able to learn to respond to relatively subtle changes in visual input. Therefore, it is possible that

when combined with different realistic faces, an even larger VisNet retina will be able to produce output responses that can distinguish the small variations between different eyebrows. An alternative method was also suggested. That is, it was proposed that some form of attentional mechanism may be necessary in order to suppress the activation that these persistent facial features produce. For such a mechanism to exist within the VisNet model, architectural changes may be necessary and these are discussed later.

The final research chapter exploits the new benefits afforded by the larger VisNet model. The primary goal of this chapter was to simulate the development of cell activity patterns reported by Wang et al. (1998), whereby neurons in IT cortex demonstrated heterogeneous response properties. In particular, some TE neurons responded only to specific views of a rotating face, while other TE neurons learned to respond more invariantly. Furthermore, the *in vivo* experimental results revealed that some TE neurons learned to respond to specific views of the rotating face, but with location invariance. The results produced when simulating the VisNet model are comparable to the experimental results. An activation map was formed in the output layer of the model when VisNet was trained with a face rotating in depth around the vertical axis, and in a separate set of simulations, VisNet developed view specific, location invariant, neuronal representations in the output layer.

Much of the work presented in this thesis investigates the performance of the VisNet model during CT learning paradigms. A short discussion of the pros and cons of CT learning follows.

CT learning requires spatial overlap between the different transforms presented during training. Depending on the experimental paradigm, this requirement can become a limiting factor, dictating the performance of the model. However, it may also be desirable since it does not allow the model to learn arbitrary associations between transforms of completely different objects, and more specifically,

objects that only occur close together in time, but not in space. For example, if two consecutive views of a bird and lion are presented in close temporal proximity during training, CT learning does not encourage the network to develop a single output representation of the two different objects. Furthermore, even if visual fixation moves rapidly and randomly between different views of different objects by saccades, CT learning will still develop invariant representations of the individual objects. Compared to trace learning, CT learning greatly increases the number of transforms of a given object that can be learned. It also lacks the eta parameter present in the trace learning rule; the eta term is sensitive to the presentation sequence length and can require fine-tuning.

CT learning is a versatile algorithm, ideal for unsupervised training, and only relies on continual synaptic modification of just the feed-forward connection weights between layers using an associative (e.g. Hebbian) learning rule. It can cause the same output neuron to learn to respond to two very different input patterns, such as two markedly different views of the same object, which is a desirable characteristic in an invariance learning paradigm. It is biologically plausible in that it requires a standard Hebbian associative learning rule, and has been replicated with ease in numerous types of neural networks. For example, CT learning has been shown to work in one-layer neural network models, as well as multi-layer neural network models such as those presented in this thesis, and results that can be consistently reproduced demonstrate that the underlying mechanism is extremely robust. In addition to working well with a basic competitive network architecture, CT learning also functions well with a self-organising map architecture.

A process with the general characteristics of CT learning could be a general property of the functional architecture of the cerebral cortex. Indeed, the modelling work with VisNet suggests that CT learning is such a pervasive process that it can often cause features that should be represented sep-

arately by the output layer of the network to become represented by the same output neurons. This was exemplified particularly clearly in the final experiments presented in Chapter 5 and is certainly a current limitation of the CT learning process.

In Chapter 5, different facial features were bound together into a single representation. The network failed to learn about facial expression separately from facial identity. The CT learning mechanism caused the same output neurons of the network that learned to respond to features of the identity space, to also learn to respond to features of the facial expression space. In particular, the output neurons learned to respond to facial features common across all transforms, such as the hair line and ears. Different CT learning parameters were explored and no solution was found. Accordingly, if CT learning is to exist in the brain, it may require modulation by another mechanism so as to avoid binding together persistent facial features and allow separate representations of different facial dimensions to form.

Past work with VisNet has focussed on the useful property of invariance learning. However, the last two research chapters of this thesis have focussed on the study of how neurons can learn to encode transform specific information. In particular, view specific information was successfully encoded by the activation map produced in the output layer of the VisNet model during simulations presented in the final research chapter. However, the exact same principles apply for location specific information. This type of transform specific information has received significant attention in the experimental literature. As a consequence, there is a significant body of experimental research that can guide simulations with VisNet, thus making it an exciting area for future modelling research.

As discussed in the Introduction chapter, neurons respond to increasingly more complex combinations of object features through successive stages of the ventral visual pathway. The types of

neural representations that form during learning at different stages of the ventral visual pathway can be studied using the VisNet model. For example, Pasupathy and Connor (2002) show that V4 neurons respond to complex patterns including curvature, as well as orientation. Perhaps of most interest is the fact that these neurons encode 'object-relative position of boundary fragments' such as 'concavity on the right'. Different neurons in V4 preferred different curvatures and positions, suggesting that a population of V4 neurons can encode whole stimuli in this way. This type of encoding is comparable to suggestions by Sereno and Lehky (2011), who argue that neurons encode a 'categorical' spatial representation (above, next to, etc.).

An interesting future study with VisNet that could perhaps develop these findings further would be to present the same 'blob stimuli' used by Pasupathy and Connor (2002). The types of neural representations that form at different stages of the VisNet model could be directly compared to the experimental results. It is hypothesised that the same kinds of neural representations observed experimentally would form in the VisNet model during learning. The previous results presented as part of this thesis (object statistical decoupling) will also work at the level of object features, such as boundary contours. That is, the same mechanisms that help output neurons form object selective responses can also help form object feature selective responses. Accordingly, individual neurons in the higher layers of VisNet should learn to respond to naturally distinguishable features of stimuli, such as to individual boundary contours or low order combinations of boundary features. This hypothesis is consistent with results observed in TE (Yamane et al., 2006).

Yamane et al. (2006) showed that monkey TE neurons are sensitive to a particular spatial arrangement of parts. Neurons will become active when local features are arranged in a particular spatial configuration revealing that these neurons represent not only the local features but also the global

features such as the spatial relationship among object parts. Similarly, it has also been shown that IT neurons selective to faces, preferentially encode face aspect ratio and inter-eye distance, thus suggesting that facial layout geometry is important. As part of the future research with 'blob stimuli', testing paradigms that explore the impact of altering local feature combinations must be created. This way, it will be possible to comprehensively explore the different kinds of spatial information that are encoded in the VisNet model.

The current VisNet model architecture is quite appropriate for performing future simulations similar to those outlined above. VisNet is biologically inspired and employs local, biologically plausible, learning rules. However, the biological accuracy of the model is limited when compared to that of a model that uses spiking neural network dynamics. Therefore, the different challenges addressed by the VisNet model must be carefully considered to ensure that they are appropriate for this type of modelling. For example, it is argued that the experiments within this thesis produce results that are relevant in a more biologically accurate model such as a spiking neural network. Nonetheless, certain experiments do require adjustments to the model architecture. In particular, if future research is to investigate the role of attention during learning, then a time accurate model implementing back-projection may be necessary.

An updated model investigating attentional mechanisms should implement back-projections in a similar manner to how normal feedforward connections are currently implemented within VisNet. More specifically, back-projections should not be hard-wired and should be configured during learning. VisNet employs rate coded neurons, and is therefore quite computationally economical to simulate. In principle, the implementation of back-projections as described would not require considerable modification to the VisNet program. For example, the model can remain discrete, as opposed to con-

tinuous, and rate coded. Such a model will allow for the study of attentional mechanisms, whereby output neurons could be stimulated during testing, or even during learning, in order to study the self-organisation of the VisNet model. Even with the increased size of the VisNet retina, this approach is possible despite current computational constraints.

It is important to not only consider the limitations and improvements that can be made to the VisNet model, but also important to understand the limitations of the stimulus set. Specifically, whether it is possible to separately infer the limitations of the model from the limitations of the stimulus set used during training and testing.

The importance of the training environment has been highlighted in past VisNet research. Stringer and Rolls (2000) investigated the performance of the VisNet model when trained with different faces presented on cluttered backgrounds during learning. VisNet failed to learn about the individual faces presented during training, and instead learned about the backgrounds upon which they were displayed. When considering the limitations of the stimulus set, its ecological validity must be considered in the context of how it relates to an overall hypothesis. A clear hypothesis must dictate the choice of stimuli used to train the model and the different ways that the stimuli are presented together. Failure to provide a clear hypothesis makes it very difficult to infer whether subsequent performance limitations lie in the model architecture itself, or the stimulus set.

For example, carefully controlled pilot studies were presented in this thesis to explore whether the VisNet architecture could, in principle, learn as hypothesised. Even if VisNet performs as expected when using an overly simplified set of stimuli, it is still difficult to claim what level of performance VisNet will demonstrate when presented with more complex stimuli. If the model is able to perform as expected when presented with carefully controlled simple stimuli, but subsequently fails if the

training stimuli are made more complex (e.g. more detail, higher resolution, more visually complex), then it may be argued that this failure is due to the change in training stimuli. In this case, the initial limitation lies in the use of a simplified stimulus set, that is, they are not indicative of the real-world environment. However, as they are made more ecologically valid, model performance breaks down. It could now be argued that this is a limitation of the model itself when dealing with more complex and ecologically valid stimuli, and that the model may require additional neurons, more layers or perhaps even more qualitative alterations if it is to avoid failure with the more complicated training stimuli. Also, it is worth considering that it can be difficult to maintain enough control over the stimulus set so as to always conduct a fair test when changing from simple stimuli to more complex stimuli. For example, as stimuli are made more intricate, they can also be made qualitatively different, thus providing a different challenge for the model. This can make it more complex to infer where the limitations lie.

In summary, when VisNet fails to learn about ecological robust stimuli, it becomes difficult to infer whether this limitation is due to aspects relating to the stimuli and training paradigm, or whether it is a fundamental limitation of the network architecture. VisNet is a complex, dynamic system and its behaviour will emerge from the different interactions of the model architecture and the stimuli, and the training and testing paradigm.

The results presented within this thesis explore the performance of the VisNet model when presented with multiple objects during learning. An important consideration is to determine what is learned about the brain versus what is learned about VisNet. As a whole, this thesis presents results from experiments that explore the dynamics of biologically plausible learning mechanisms operating as part of a biologically plausible network architecture. The VisNet architecture comprises a relatively

straightforward neural network utilising only feed-forward connections that are modified through basic associative learning rules. Due to the simplicity of the model, it is reasonable that the mechanisms explored within this thesis could be instantiated in the primate brain, specifically the visual cortex. As previously stated, given the local feed-forward connectivity along with topological maps that are both present in early visual cortex, a process with the general characteristics of CT learning is implied to be a general property of the functional architecture of the cerebral cortex. Accordingly, it is argued that the results of this thesis may have general applicability to learning in the primate cortical visual areas.

Despite robust results supporting general purpose mechanisms, it is important that the results presented within this thesis are not automatically assumed to be valid in the context of the primate visual system. The primate visual system is clearly far more complex than the VisNet model. Notably, the VisNet model is rate coded and also lacks a number of known cortical connections, such as back-projections, as well as connections between different brain areas, such as the dorsal and ventral visual pathways. These omissions are potential shortcomings of the VisNet model and they make it difficult to extend to the brain any specific aspects of what is learned about the VisNet model. To help understand what is specifically learned about VisNet versus what is learned about the brain, experimental predictions, grounded in the results from the simulation studies, must be made. Predictions that can be validated in carefully controlled experimental investigation will help establish precisely what is learned about VisNet and what is learned about the brain.

At times, experimental predictions can be made that are not practical because the work to investigate them may require the measurement of single cell firing rates that is not currently viable in primate brains. However, it is also possible to make experimental predictions that can be tested by

simply observing the behaviour of a conscious human participant, and measuring performance using straightforward tests.

Past research has shown that common motion can cause infants to perceive an object, whose top and bottom were visible but whose centre was occluded, as a unit that was connected behind the occluder (Slater et al., 1990). When the object's parts moved independently of one another, this was not the case. This is congruent with results from Chapter 4 of this thesis, where the VisNet model built separate representations of independently moving objects, but developed a single representation when they moved in lock-step.

An example of a new experimental prediction is as follows. Results from Chapter 2 suggest that the frequency with which objects are seen together can play an important role in determining how successful the VisNet model is at developing individual representations for each object. If the different objects are not presented frequently enough with the other objects that are also presented during training, output neurons fail to learn about individual objects, and instead they learn to represent pairs of objects (or triples, etc.). A straightforward experimental prediction based on this result is that human participants will perform similarly in a carefully controlled experimental paradigm. For example, if novel 2D objects, such as wire-frame objects, are presented to participants in different combinations of pairings, triples, etc. and participants must eventually conclude which features comprise one object versus another, successful performance will be a function of the frequency with which objects are presented with one another during the training phase. If objects are not presented frequently enough with other objects, participants will either be unable, or find it very challenging, to discern which features together comprise an object. Careful consideration would also need to be given to how the spatial location of different objects could provide additional information to participants that would help them

solve this challenge more easily and avoid testing the underlying prediction.

It is not always possible to produce practical experimental predictions from the VisNet model, but it is very important that, wherever possible, simulation work is designed with experimental predictions in mind. Computational modelling work is most valuable when the results produced can help guide experimental work and the two distinct avenues of research can complement one another.

In summary, this thesis has explored how VisNet, a biologically plausible neural network model of the ventral visual pathway, may learn to produce some cell firing properties observed empirically in the latter stages of the primate ventral visual pathway. Two general themes emerge from the research presented throughout the thesis chapters. Firstly, that the ecological validity of the training environment can have a important impact on the types of representations that form during learning. Secondly, that the types of neural representations that form during learning are heterogeneous, providing information about specific transforms, and do not only learn to respond invariantly to all transforms. Although these simulations do not explore the performance of the VisNet model in genuinely complex natural scene environments, a principled first step away from individual object presentations to more scene-like training environments has been taken. A new, scaled-up version of the VisNet architecture was shown to have a qualitative impact on the types of output representations that formed during learning, despite the fact that it implemented identical computational principles to the previous, original version of VisNet. Future research, such as the role of back-projections during learning, or the study of neural encoding in the intermediate layers of the VisNet model, was also considered.

Bibliography

Aggelopoulos, N. C. and Rolls, E. T. (2005). Scene perception: inferior temporal cortex neurons encode the positions of different objects in the scene. *The European Journal of Neuroscience*, 22(11):2903–2916. PMID: 16324125.

Amaral, D. G. and Price, J. L. (1984). Amygdalocortical projections in the monkey (macaca fascicularis). *The Journal of Comparative Neurology*, 230(4):465–496.

Anzai, A., Peng, X., and Van Essen, D. C. (2007). Neurons in monkey visual area v2 encode combinations of orientations. *Nat Neurosci*, 10(10):1313–1321.

Bailey, C. H., Gouras, P., Kandel, E. R., and Schwartz, J. H. (1985). The retina and phototransduction. In *Principles of Neural Science*, pages 63–81. Elsevier Science Publishing Co., Inc. New York, New York, 2 edition.

Barbas, H. (1988). Anatomic organization of basoventral and mediodorsal visual recipient prefrontal regions in the rhesus monkey. *The Journal of Comparative Neurology*, 276(3):313–342.

Barlow, H. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, 1(4):371–394.

Baylis, G., Rolls, E., and Leonard, C. (1987). Functional subdivisions of the temporal lobe neocortex. *The Journal of Neuroscience*, 7(2):330 –342.

Bednar, J. A. and Miikkulainen, R. (2006). Joint maps for orientation, eye, and direction preference in a self-organizing model of v1. *Neurocomputing*, 69(10-12):1272–1276.

Bell, A. J. and Sejnowski, T. J. (1997). The independent components of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338.

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. PMID: 3575582.

Blumberg, J. and Kreiman, G. (2010). How cortical neurons help us see: visual recognition in the human brain. *Journal of Clinical Investigation*, 120(9):3054–3063.

Booth, M. and Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb. Cortex*, 8(6):510–523.

Bosking, W. H., Zhang, Y., Schofield, B., and Fitzpatrick, D. (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *The Journal of Neuroscience*, 17(6):2112 –2127.

Brincat, S. L. and Connor, C. E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nat Neurosci*, 7(8):880–886.

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.

Calder, A. J. and Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nat Rev Neurosci*, 6(8):641–651.

Callaway, E. M. (1998). Local circuits in primary visual cortex of the macaque monkey. *Annual Review of Neuroscience*, 21:47–74. PMID: 9530491.

Cerella, J. (1986). Pigeons and perceptrons. *Pattern Recognition*, 19(6):431–438.

Choe, Y. and Miikkulainen, R. (1998). Self-organization and segmentation in a laterally connected orientation map of spiking neurons. *Neurocomputing*, 21(1-3):139–158.

Cohen Kadosh, K., Henson, R. N. A., Cohen Kadosh, R., Johnson, M. H., and Dick, F. (2010). Task-dependent activation of face-sensitive cortex: An fMRI adaptation study. *Journal of Cognitive Neuroscience*, 22(5):903–917.

Connor, C. E., Preddie, D. C., Gallant, J. L., and Van Essen, D. C. (1997). Spatial attention effects in macaque area v4. *The Journal of Neuroscience*, 17(9):3201 –3214.

Croner, L. J. and Kaplan, E. (1995). Receptive fields of p and m ganglion cells across the primate retina. *Vision Research*, 35(1):7–24.

Damasio, A. R., Tranel, D., and Damasio, H. (1990). Face agnosia and the neural substrates of memory. *Annual Review of Neuroscience*, 13(1):89–109. PMID: 2183687.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America. A, Optics and image science*, 2(7):1160–1169. PMID: 4020513.

De Monasterio, F. M. and Gouras, P. (1975). Functional properties of ganglion cells of the rhesus monkey retina. *The Journal of Physiology*, 251(1):167–195. PMID: 810576 PMCID: 1348381.

De Valois, R. L., Albrecht, D. G., and Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research*, 22(5):545–559. PMID: 7112954.

Dean, P. (1976). Effects of inferotemporal lesions on the behavior of monkeys. *Psychological Bulletin*, 83(1):41–71. PMID: 828276.

Denison, R. N. and Silver, M. A. (2011). Distinct contributions of the magnocellular and parvocellular visual streams to perceptual selection. *Journal of Cognitive Neuroscience*, 24(1):246–259.

Derrington, A. M. and Lennie, P. (1984). Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. *The Journal of Physiology*, 357:219–240. PMID: 6512690 PMCID: 1193256.

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience*, 3(1):1–8.

Desimone, R., Albright, T., Gross, C., and Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *The Journal of Neuroscience*, 4(8):2051 –2062.

Desimone, R. and Schein, S. J. (1987). Visual properties of neurons in area v4 of the macaque: sensitivity to stimulus form. *Journal of Neurophysiology*, 57(3):835 –868.

DiCarlo, J. J. and Maunsell, J. H. R. (2003). Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *Journal of Neurophysiology*, 89(6):3264–3278. PMID: 12783959.

Do, Q., Lozo, P., and Jain, L. (2005). A vision system for partially occluded landmark recognition. In Zhang, S. and Jarvis, R., editors, *AI 2005: Advances in Artificial Intelligence*, volume 3809 of *Lecture Notes in Computer Science*, pages 1246–1252. Springer Berlin / Heidelberg.

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.

Ferster, D. and Miller, K. D. (2000). Neural mechanisms of orientation selectivity in the visual cortex. *Annual Review of Neuroscience*, 23(1):441–471.

Fitzpatrick, D., Lund, J. S., and Blasdel, G. G. (1985). Intrinsic connections of macaque striate cortex: afferent and efferent connections of lamina 4C. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 5(12):3329–3349. PMID: 3001242.

Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Comput.*, 3(2):194–200.

Fox, C. J., Moon, S. Y., Iaria, G., and Barton, J. J. (2009). The correlates of subjective perception of identity and expression in the face network: An fMRI adaptation study. *NeuroImage*, 44(2):569–580.

Freiwald, W. A. and Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science (New York, N.Y.)*, 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.

Fujita, I., Tanaka, K., Ito, M., and Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, 360(6402):343–346.

Gary, G. C., Branson, K. M., and Calder, A. J. (2002). Do expression and identity need separate representations? *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*, pages 238—243.

Gattass, R., Sousa, A., and Gross, C. (1988). Visuotopic organization and extent of v3 and v4 of the macaque. *J. Neurosci.*, 8(6):1831–1845.

Gegenfurtner, K. R., Kiper, D. C., and Levitt, J. B. (1997). Functional properties of neurons in macaque area v3. *Journal of Neurophysiology*, 77(4):1906 –1923.

Ghose, G. M. and Ts'O, D. Y. (1997). Form processing modules in primate area v4. *Journal of Neurophysiology*, 77(4):2191 –2196.

Gibson, J. J. (1950). *The perception of the visual world*. Houghton Mifflin.

Gilbert, C. D. and Sigman, M. (2007). Brain states: top-down influences in sensory processing. *Neuron*, 54(5):677–696. PMID: 17553419.

Goodale, M. A. and Milner, A. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25.

Gothard, K. M., Battaglia, F. P., Erickson, C. A., Spitler, K. M., and Amaral, D. G. (2007). Neural responses to facial expression and face identity in the monkey amygdala. *Journal of Neurophysiology*, 97(2):1671 –1683.

Gregory, R. L. (1997). *Eye and Brain: The Psychology of Seeing*. OUP Oxford, 5 edition.

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. PMID: 4621506.

Hadj-Bouziane, F., Bell, A. H., Knusten, T. A., Ungerleider, L. G., and Tootell, R. B. H. (2008). Perception of emotional expressions is independent of face selectivity in monkey inferior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 105(14):5591–5596. PMID: 18375769 PMCID: 2291084.

Harris, A. E., Ermentrout, G. B., and Small, S. L. (1997). A model of ocular dominance column development by competition fortrophicfactor. *Proceedings of the National Academy of Sciences of the United States of America*, 94(18):9944–9949. PMID: 9275231 PMCID: 23304.

Hasselmo, M. E., Rolls, E. T., and Baylis, G. C. (1989a). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural Brain Research*, 32(3):203–218.

Hasselmo, M. E., Rolls, E. T., Baylis, G. C., and Nalwa, V. (1989b). Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research. Experimentelle Hirnforschung. Exprimentation Crbrale*, 75(2):417–429. PMID: 2721619.

Hawken, M. J. and Parker, A. J. (1987). Spatial properties of neurons in the monkey striate cortex. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 231(1263):251–288.

Hebb, D. O. (1950). Organization of behavior. *Journal of Clinical Psychology*, 6(3):307–335.

Hecht, E. (1987). *Optics*. Addison-Wesley, 2nd edition.

Hendry, S. H. C. and Reid, R. C. (2000). The koniocellular pathway in primate vision. *Annual Review of Neuroscience*, 23(1):127–153.

Hertz, J. A., Krogh, A. S., and Palmer, R. G. (1991). *Introduction To The Theory Of Neural Computation, Volume I*. Westview Press.

Hess, R., Baker, C., and Zihl, J. (1989). The "motion-blind" patient: low-level spatial and temporal filters. *The Journal of Neuroscience*, 9(5):1628 –1640.

Hinton, G., Dayan, P., Frey, B., and Neal, R. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158 –1161.

Hinton, G. E. (2010). Learning to represent visual input. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537):177 –184.

Hinton, G. E. and Ghahramani, Z. (1997). Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 352(1358):1177–1190. PMID: 9304685 PMCID: 1692002.

Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574 –591.

Hubel, D. H. and Wiesel, T. N. (1961). Integrative action in the cat's lateral geniculate body. *The Journal of Physiology*, 155(2):385 –398.

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.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243. PMID: 49664574966457 PMCID: 1557912.

Hummel, J. E. and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review;Psychological Review*, 99(3):480–517.

Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science (New York, N.Y.)*, 310(5749):863–866. PMID: 16272124.

Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J Neurophysiol*, 73(1):218–226.

Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233 –1258.

Kang, K., Shelley, M., and Sompolinsky, H. (2003). Mexican hats and pinwheels in visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 100(5):2848–2853.

Kiani, R., Esteky, H., Mirpour, K., and Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *J Neurophysiol*, 97(6):4296–4309.

Kobatake, E. and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J Neurophysiol*, 71(3):856–867.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69.

Kreiman, G., Hung, C. P., Kraskov, A., Quiroga, R. Q., Poggio, T., and DiCarlo, J. J. (2006). Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. *Neuron*, 49(3):433–445. PMID: 16446146.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 16(1):37–68. PMID: 13035466.

Lamme, V. A. and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11):571–579.

Leventhal, A., Rodieck, R., and Dreher, B. (1981). Retinal ganglion cell classes in the old world monkey: morphology and central projections. *Science*, 213(4512):1139 –1142.

Levine, M. W. (2000). *Fundamentals of Sensation and Perception*. OUP Oxford, 3 edition.

Levitt, J. B., Kiper, D. C., and Movshon, J. A. (1994). Receptive fields and functional architecture of macaque v2. *Journal of Neurophysiology*, 71(6):2517 –2542.

Livingstone, M. and Hubel, D. (1988). Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, 240(4853):740 –749.

Logothetis, N. K., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5):552–563.

Logothetis, N. K. and Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, 19:577–621. PMID: 8833455.

Marr, D. and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 200(1140):269–294. ArticleType: research-article / Full publication date: Feb. 23, 1978 / Copyright 1978 The Royal Society.

Martin, K. A. (1992). Parallel pathways converge. *Current Biology*, 2(10):555–557.

Mel, B. W. (1997). SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation*, 9(4):777–804. PMID: 9161022.

Meltzoff, A. N. and Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, 198(4312):75 –78.

Meltzoff, A. N. and Moore, M. K. (1983). Newborn infants imitate adult facial gestures. *Child Development*, 54(3):702–709. ArticleType: research-article / Full publication date: Jun., 1983 / Copyright 1983 Society for Research in Child Development.

Merigan, W. H. and Maunsell, J. H. (1990). Macaque vision after magnocellular lateral geniculate lesions. *Visual Neuroscience*, 5(04):347–352.

Michler, F., Eckhorn, R., and Wachtler, T. (2009). Using spatiotemporal correlations to learn topographic maps for invariant object recognition. *Journal of Neurophysiology*, 102(2):953 –964.

Miikkulainen, R., Bednar, J. A., Choe, Y., and Sirosh, J. (2005). *Computational Maps in the Visual Cortex*. Springer.

Miller, K. D. (1994). Models of activity-dependent neural development. *Progress in Brain Research*, 102:303–318. PMID: 7800821.

Mishkin, M. and Pribram, K. H. (1954). Visual discrimination performance following partial ablations of the temporal lobe. i. ventral vs. lateral. *Journal of Comparative and Physiological Psychology*, 47(1):14–20. PMID: 13130724.

Moran, J. and Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, 229(4715):782–784.

Murphy, P. C., Duckett, S. G., and Sillito, A. M. (1999). Feedback connections to the lateral geniculate nucleus and cortical response properties. *Science*, 286(5444):1552 –1554.

Nahm, F. K. D., Perret, A., Amaral, D. G., and Albright, T. D. (1997). How do monkeys look at faces? *J. Cognitive Neuroscience*, 9(5):611623.

Obermayer, K. and Blasdel, G. (1993). Geometry of orientation and ocular dominance columns in monkey striate cortex. *The Journal of Neuroscience*, 13(10):4114 –4129.

Op De Beeck, H. and Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. *The Journal of Comparative Neurology*, 426(4):505–518. PMID: 11027395.

Panzeri, S. and Schultz, S. R. (2001). A unified approach to the study of temporal, correlational, and rate coding. *Neural Computation*, 13(6):1311–1349.

Panzeri, S., Schultz, S. R., Treves, A., and Rolls, E. T. (1999). Correlations and the encoding of information in the nervous system. *Proceedings of the Royal Society B: Biological Sciences*, 266(1423):1001–1012. PMID: 10610508 PMCID: PMC1689940.

Parker, J. R. (1996). *Algorithms for Image Processing and Computer Vision*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition.

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. (2002). Population coding of shape in area v4. *Nat Neurosci*, 5(12):1332–1338.

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., Benson, P. J., and Rolls, E. T. (1992). Organization and functions of cells responsive to faces in the temporal cortex [and discussion]. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 335(1273):23 –30.

Perrett, D. I., Smith, P. A., Potter, D. D., Mistlin, A. J., Head, A. S., Milner, A. D., and Jeeves, M. A. (1984). Neurones responsive to faces in the temporal cortex: studies of functional organization, sensitivity to identity and relation to perception. *Human Neurobiology*, 3(4):197–208. PMID: 6526706.

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*, 204(2):255–270.

Perry, V. H., Oehler, R., and Cowey, A. (1984). Retinal ganglion cells that project to the dorsal lateral geniculate nucleus in the macaque monkey. *Neuroscience*, 12(4):1101–1123. PMID: 6483193.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107.

Reynolds, J. H., Chelazzi, L., and Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas v2 and v4. *The Journal of Neuroscience*, 19(5):1736 –1753.

Rolls, E. and Treves, A. (1990). The relative advantages of sparse versus distributed encoding for associative neuronal networks in the brain. *Network: Computation in Neural Systems*, 1(4):407–421.

Rolls, E. T. (1984). Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. *Human Neurobiology*, 3(4):209–222. PMID: 6526707.

Rolls, E. T. (1991). Neural organization of higher visual functions. *Current Opinion in Neurobiology*, 1(2):274–278. PMID: 1821190.

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. Experimentelle Hirnforschung. Exprimentation Crbrale*, 65(1):38–48. PMID: 3803509.

Rolls, E. T., Cowey, A., and Bruce, V. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philosophical Transactions: Biological Sciences*, 335(1273):11–21.

Rolls, E. T., Critchley, H., Browning, A., and Inoue, K. (2006). Face-selective and auditory neurons in the primate orbitofrontal cortex. *Experimental Brain Research*, 170(1):74–87.

Rolls, E. T. and Deco, G. (2002). *Computational neuroscience of vision*. Oxford University Press, Oxford.

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 Stringer, S. M. (2000). On the design of neural networks in the brain by genetic evolution. *Progress in Neurobiology*, 61(6):557–579.

Rolls, E. T. and Stringer, S. M. (2006). Invariant visual object recognition: A model, with lighting invariance. *JOURNAL OF PHYSIOLOGY - PARIS*, 100:43—62.

Rolls, E. T., Tovee, M. J., Purcell, D. G., Stewart, A. L., and Azzopardi, P. (1994). The responses of neurons in the temporal cortex of primates, and face identification and detection. *Experimental Brain Research. Experimentelle Hirnforschung. Exprimentation Crbrale*, 101(3):473–484. PMID: 7851514.

Rolls, E. T. and Treves, A. (1998). *Neural Networks and Brain Function*. Oxford University Press, USA, 1 edition.

Rolls, E. T. and Treves, A. (2011). The neuronal encoding of information in the brain. *Progress in Neurobiology*, 95(3):448–490.

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*, 4:309–333.

Royer, S. and Pare, D. (2003). Conservation of total synaptic weight through balanced synaptic depression and potentiation. *Nature*, 422(6931):518–522.

Rumelhart, D. E. and Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, 9(1):75–112.

Sakai, H. M., Wang, J. L., and Naka, K. (1995). Contrast gain control in the lower vertebrate retinas. *The Journal of General Physiology*, 105(6):815–835. PMID: 7561745.

Sato, T., Uchida, G., and Tanifuji, M. (2009). Cortical columnar organization is reconsidered in inferior temporal cortex. *Cerebral Cortex*, 19(8):1870 –1888.

Schiller, P. H., Logothetis, N. K., and Charles, E. R. (1990). Functions of the colour-opponent and broad-band channels of the visual system. *Nature*, 343(6253):68–70.

Schroeder, C., Tenke, C., Arezzo, J., and Vaughan Jr., H. (1990). Binocularity in the lateral geniculate nucleus of the alert macaque. *Brain Research*, 521(1-2):303–310.

Schwartz, E. L., Desimone, R., Albright, T. D., and Gross, C. G. (1983). Shape recognition and inferior temporal neurons. *Proceedings of the National Academy of Sciences*, 80(18):5776 –5778.

Schwartz, G. and Berry, Michael J, n. (2008). Sophisticated temporal pattern recognition in retinal ganglion cells. *Journal of Neurophysiology*, 99(4):1787–1798. PMID: 18272878.

Sereno, A. B. and Lehky, S. R. (2011). Population coding of visual space: Comparison of spatial representations in dorsal and ventral pathways. *Frontiers in Computational Neuroscience*, 4. PMID: 21344010 PMCID: PMC3034230.

Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., Rosen, B. R., and Tootell, R. B. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science (New York, N.Y.)*, 268(5212):889–893. PMID: 7754376.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Shapley, R. and Perry, V. (1986). Cat and monkey retinal ganglion cells and their visual functional roles. *Trends in Neurosciences*, 9(0):229–235.

Sillito, A. M., Cudeiro, J., and Murphy, P. C. (1993). Orientation sensitive elements in the corticofugal influence on centre-surround interactions in the dorsal lateral geniculate nucleus. *Experimental Brain Research. Experimentelle Hirnforschung. Exprimentation Crbrale*, 93(1):6–16. PMID: 8467892.

Sirosh, J. and Miikkulainen, R. (1994). Cooperative self-organization of afferent and lateral connections in cortical maps. *Biological Cybernetics*, 71(1):65–78.

Slater, A., Morison, V., Somers, M., Mattock, A., Brown, E., and Taylor, D. (1990). Newborn and older infants' perception of partly occluded objects. *Infant Behavior and Development*, 13(1):33 – 49.

Smirnakis, S. M., Berry, M. J., Warland, D. K., Bialek, W., and Meister, M. (1997). Adaptation of retinal processing to image contrast and spatial scale. *Nature*, 386(6620):69–73.

Soo, F. S., Schwartz, G. W., Sadeghi, K., and Berry, M. J. (2011). Fine spatial information represented in a population of retinal ganglion cells. *The Journal of Neuroscience*, 31(6):2145 –2155.

Stringer, S., Perry, G., Rolls, E., and Proske, J. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biological Cybernetics*, 94(2):128–142.

Stringer, S. M. and Rolls, E. T. (2000). Position invariant recognition in the visual system with cluttered environments. *Neural Networks*, 13(3):305–315.

Stringer, S. M. and Rolls, E. T. (2002). Invariant object recognition in the visual system with novel views of 3D objects. *Neural Computation*, 14:25852596. ACM ID: 639721.

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*, 21(7):888–903.

Stringer, S. M., Rolls, E. T., and Tromans, J. M. (2007). Invariant object recognition with trace learning and multiple stimuli present during training. *Network: Computation in Neural Systems*, 18(2):161.

Sry, G., Vogels, R., and Orban, G. A. (1993). Cue-invariant shape selectivity of macaque inferior temporal neurons. *Science (New York, N.Y.)*, 260(5110):995–997. PMID: 8493538.

Tanaka, K. (1993a). Column structure of inferotemporal cortex: visual alphabet or differential amplifiers? In *Proceedings of 1993 International Joint Conference on Neural Networks, 1993. IJCNN '93-Nagoya*, volume 2, pages 1095– 1099 vol.2. IEEE.

Tanaka, K. (1993b). Neuronal mechanisms of object recognition. *Science (New York, N.Y.)*, 262(5134):685–688. PMID: 8235589.

Tanaka, K. (1996). Representation of visual features of objects in the inferotemporal cortex. *Neural Networks*, 9(8):1459–1475.

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.

Tootell, R. B., Silverman, M. S., De Valois, R. L., and Jacobs, G. H. (1983). Functional organization of the second cortical visual area in primates. *Science (New York, N.Y.)*, 220(4598):737–739. PMID: 6301017.

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. *J Neurophysiol*, 72(3):1049–1060.

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. *Nat Neurosci*, 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*, 311(5761):670–674.

Tsao, D. Y., Moeller, S., and Freiwald, W. A. (2008). Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences of the United States of America*, 105(49):19514–19519. PMID: 19033466.

Tsunoda, K., Yamane, Y., Nishizaki, M., and Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nat Neurosci*, 4(8):832–838.

Tucker, T. R. and Katz, L. C. (2003). Spatiotemporal patterns of excitation and inhibition evoked by the horizontal network in layer 2/3 of ferret visual cortex. *Journal of Neurophysiology*, 89(1):488 –500.

Ullmann, J. (1992). Analysis of 2-d occlusion by subtracting out. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(4):485 –489.

Ungerleider, L., Mishkin, M., Ingle, D., Goodale, M., and Mansfield, R. (1982). Two cortical visual systems. In *Analysis of Visual Behaviour*, pages 549–586. MIT Press.

Ungerleider, L. G. and Haxby, J. V. (1994). What and where in the human brain. *Current Opinion in Neurobiology*, 4(2):157–165.

Vanduffel, W., Tootell, R. B., Schoups, A. A., and Orban, G. A. (2002). The organization of orientation selectivity throughout macaque visual cortex. *Cerebral Cortex*, 12(6):647 –662.

von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14(2):85–100. PMID: 4786750.

Wallis, G., Rolls, E., and Foldiak, P. (1993). Learning invariant responses to the natural transformations of objects. volume 2, pages 1087– 1090 vol.2.

Wallis, G. and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, 51(2):167–194.

Walton, G. E., Bower, N., and Bower, T. (1992). Recognition of familiar faces by newborns. *Infant Behavior and Development*, 15(2):265–269.

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, Y., Fujita, I., Tamura, H., and Murayama, Y. (2002). Contribution of GABAergic inhibition to receptive field structures of monkey inferior temporal neurons. *Cerebral Cortex*, 12(1):62 –74.

Winston, J. S., Henson, R., Fine-Goulden, M. R., and Dolan, R. J. (2004). fMRI-Adaptation reveals dissociable neural representations of identity and expression in face perception. *Journal of Neurophysiology*, 92(3):1830 –1839.

Xu, X. and Biederman, I. (2010). Loci of the release from fMRI adaptation for changes in facial expression, identity, and viewpoint. *Journal of Vision*, 10(14).

Yamane, Y., Tsunoda, K., Matsumoto, M., Phillips, A. N., and Tanifuji, M. (2006). Representation of the spatial relationship among object parts by neurons in macaque inferotemporal cortex. *J Neurophysiol*, 96(6):3147–3156.

Ying, Z. and Castaon, D. (2002). Partially occluded object recognition using statistical models. *International Journal of Computer Vision*, 49(1):57–78.

Yukie, M. and Iwai, E. (1988). Direct projections from the ventral TE area of the inferotemporal cortex to hippocampal field CA1 in the monkey. *Neuroscience Letters*, 88(1):6–10.

Zeki, S. (1973). Colour coding in rhesus monkey prestriate cortex. *Brain Research*, 53(2):422–427.