Memory, Attention, and Decision-Making

A Unifying Computational Neuroscience Approach

Edmund T. Rolls
University of Oxford
Department of Experimental Psychology
Oxford
England
Preface

The overall aim of this book is to provide insight into how memory systems in the brain work; how the operation of these systems is fundamental to understanding many aspects of brain function including perception, decision-making, and action selection; and how interactions between these systems provide an account of many cognitive phenomena including attention and emotion. It is shown that to do this, a neurocomputational approach is needed. The book provides Appendices that describe many of the building blocks of the neurocomputational approach, and that are designed to be useful for teaching.

To understand how the brain works, including how it functions in memory, attention, and decision-making, it is necessary to combine different approaches, including neural computation. Neurophysiology at the single neuron level is needed because this is the level at which information is exchanged between the computing elements of the brain. Evidence from the effects of brain damage, including that available from neuropsychology, is needed to help understand what different parts of the system do, and indeed what each part is necessary for. Neuroimaging is useful to indicate where in the human brain different processes take place, and to show which functions can be dissociated from each other. Knowledge of the biophysical and synaptic properties of neurons is essential to understand how the computing elements of the brain work, and therefore what the building blocks of biologically realistic computational models should be. Knowledge of the anatomical and functional architecture of the cortex is needed to show what types of neuronal network actually perform the computation. And finally the approach of neural computation is needed, as this is required to link together all the empirical evidence to produce an understanding of how the system actually works. This book utilizes evidence from all these disciplines to develop an understanding of how different types of memory, perception, attention, and decision-making are implemented by processing in the brain.

I emphasize that to understand memory, perception, attention, and decision-making in the brain, we are dealing with large-scale computational systems with interactions between the parts, and that this understanding requires analysis at the computational and global level of the operation of many neurons to perform together a useful function. Understanding at the molecular level is important for helping to understand how these large-scale computational processes are implemented in the brain, but will not by itself give any account of what computations are performed to implement these cognitive functions. Instead, understanding cognitive functions such as object recognition, memory recall, attention, and decision-making requires single neuron data to be closely linked to computational models of how the interactions between large numbers of neurons and many networks of neurons allow these cognitive problems to be solved. The single neuron level is important in this approach, for the single neurons can be thought of as the computational units of the system, and is the level at which the information is exchanged by the spiking activity between the computational elements of the brain. The single neuron level is therefore, because it is the level at which information is communicated between the computing elements of the brain, the fundamental level of information processing, and the level at which the information can be read out (by recording the spiking activity) in order to understand what information is being represented and processed in each brain area.

With its focus on how the brain works at the computational neuroscience level, this book
is distinct from the many excellent books on neuroscience that describe much evidence about brain structure and function, but do not aim to provide an understanding of how the brain works at the computational level. This book aims to forge an understanding of how some key brain systems may operate at the computational level, so that we can understand how the brain actually performs some of its complex and necessarily computational functions in memory, perception, attention, and decision-making.

A test of whether one’s understanding is correct is to simulate the processing on a computer, and to show whether the simulation can perform the tasks of memory systems in the brain, and whether the simulation has similar properties to the real brain. The approach of neural computation leads to a precise definition of how the computation is performed, and to precise and quantitative tests of the theories produced. How memory systems in the brain work is a paradigm example of this approach, because memory-like operations which involve altered functionality as a result of synaptic modification are at the heart of how all computations in the brain are performed. It happens that attention and decision-making can be understood in terms of interactions between and fundamental operations in memory systems in the brain, and therefore it is natural to treat these areas of cognitive neuroscience as well as memory in this book. The same fundamental concepts based on the operation of neuronal circuitry can be applied to all these functions, as is shown in this book.

One of the distinctive properties of this book is that it links the neural computation approach not only firmly to neuronal neurophysiology, which provides much of the primary data about how the brain operates, but also to psychophysical studies (for example of attention); to neuropsychological studies of patients with brain damage; and to functional magnetic resonance imaging (fMRI) (and other neuroimaging) approaches. The empirical evidence that is brought to bear is largely from non-human primates and from humans, because of the considerable similarity of their memory and related systems, and the overall aims to understand how memory and related functions are implemented in the human brain, and the disorders that arise after brain damage.

The overall aims of the book are developed further, and the plan of the book is described, in Chapter 1, Section 1.1.

Part of the material described in the book reflects work performed in collaboration with many colleagues, whose tremendous contributions are warmly appreciated. The contributions of many will be evident from the references cited in the text. Especial appreciation is due to Gustavo Deco, Simon M. Stringer, and Alessandro Treves who have contributed greatly in an always interesting and fruitful research collaboration on computational aspects of brain function, and to many neurophysiology and functional neuroimaging colleagues who have contributed to the empirical discoveries that provide the foundation to which the computational neuroscience must always be closely linked, and whose names are cited throughout the text. Much of the work described would not have been possible without financial support from a number of sources, particularly the Medical Research Council of the UK, the Human Frontier Science Program, the Wellcome Trust, and the James S. McDonnell Foundation. The book was typeset by the author using LaTeX and WinEdt.

The cover shows part of the picture Psyche Opening the Golden Box painted in 1903 by J. W. Waterhouse. The metaphor is to look inside the system of the mind and the brain, in order to understand how the brain functions, and thereby better to understand and treat its disorders. Updates to the publications cited in this book are available at http://www.oxcns.org.

I dedicate this work to the overlapping group: my family, friends, and colleagues – in salutem praesentium, in memoriam absentium.
Contents

1 Introduction 1
  1.1 Introduction and overview 1
  1.2 Neurons 4
  1.3 Neurons in a network 5
  1.4 Synaptic modification 7
  1.5 Long-term potentiation and long-term depression 8
  1.6 Distributed representations 13
    1.6.1 Definitions 13
    1.6.2 Advantages of different types of coding 14
  1.7 Neuronal network approaches versus connectionism 15
  1.8 Introduction to three neuronal network architectures 16
  1.9 Systems-level analysis of brain function 17
  1.10 The fine structure of the cerebral neocortex 22
    1.10.1 The fine structure and connectivity of the neocortex 22
    1.10.2 Excitatory cells and connections 22
    1.10.3 Inhibitory cells and connections 24
    1.10.4 Quantitative aspects of cortical architecture 26
    1.10.5 Functional pathways through the cortical layers 28
    1.10.6 The scale of lateral excitatory and inhibitory effects, and the concept of modules 30
  1.11 Backprojections in the cortex 31
    1.11.1 Architecture 31
    1.11.2 Learning 32
    1.11.3 Recall 34
    1.11.4 Semantic priming 35
    1.11.5 Attention 35
    1.11.6 Autoassociative storage, and constraint satisfaction 35

2 The hippocampus and memory 37
  2.1 Introduction 37
  2.2 Systems-level functions of the hippocampus 38
    2.2.1 Systems-level anatomy 38
    2.2.2 Evidence from the effects of damage to the hippocampus 40
    2.2.3 The necessity to recall information from the hippocampus 42
    2.2.4 Systems-level neurophysiology of the primate hippocampus 43
    2.2.5 Head direction cells in the presubiculum 49
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.6</td>
<td>Perirhinal cortex, recognition memory, and long-term familiarity</td>
<td>51</td>
</tr>
<tr>
<td>2.3</td>
<td>A theory of the operation of hippocampal circuitry as a memory system</td>
<td>57</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Hippocampal circuitry</td>
<td>58</td>
</tr>
<tr>
<td>2.3.2</td>
<td>CA3 as an autoassociation memory</td>
<td>58</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Dentate granule cells</td>
<td>75</td>
</tr>
<tr>
<td>2.3.4</td>
<td>The learning of spatial view and place cell representations from</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>visual inputs</td>
<td></td>
</tr>
<tr>
<td>2.3.5</td>
<td>Linking the inferior temporal visual cortex to spatial view and place</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>cell representations</td>
<td></td>
</tr>
<tr>
<td>2.3.6</td>
<td>CA1 cells</td>
<td>82</td>
</tr>
<tr>
<td>2.3.7</td>
<td>Backprojections to the neocortex – a hypothesis</td>
<td>86</td>
</tr>
<tr>
<td>2.3.8</td>
<td>Backprojections to the neocortex – quantitative aspects</td>
<td>88</td>
</tr>
<tr>
<td>2.3.9</td>
<td>Simulations of hippocampal operation</td>
<td>90</td>
</tr>
<tr>
<td>2.4</td>
<td>Tests of the theory</td>
<td>91</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Dentate gyrus (DG) subregion of the hippocampus</td>
<td>91</td>
</tr>
<tr>
<td>2.4.2</td>
<td>CA3 subregion of the hippocampus</td>
<td>94</td>
</tr>
<tr>
<td>2.4.3</td>
<td>CA1 subregion of the hippocampus</td>
<td>101</td>
</tr>
<tr>
<td>2.5</td>
<td>Evaluation of the theory of hippocampal function</td>
<td>105</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Quantitative aspects of the model</td>
<td>105</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Tests of the theory by hippocampal system subregion analyses</td>
<td>106</td>
</tr>
<tr>
<td>2.5.3</td>
<td>Comparison with other theories of hippocampal function</td>
<td>109</td>
</tr>
<tr>
<td>3</td>
<td>Reward- and punishment-related learning; emotion and motivation</td>
<td>113</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>113</td>
</tr>
<tr>
<td>3.2</td>
<td>Associative processes involved in reward- and punishment-related</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>learning</td>
<td></td>
</tr>
<tr>
<td>3.2.1</td>
<td>Pavlovian or classical conditioning</td>
<td>116</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Instrumental learning</td>
<td>118</td>
</tr>
<tr>
<td>3.3</td>
<td>Overview of brain processes involved in reward and punishment</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td>learning</td>
<td></td>
</tr>
<tr>
<td>3.4</td>
<td>Representations of primary reinforcers</td>
<td>124</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Taste</td>
<td>125</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Smell</td>
<td>125</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Pleasant and painful touch</td>
<td>126</td>
</tr>
<tr>
<td>3.4.4</td>
<td>Visual stimuli</td>
<td>128</td>
</tr>
<tr>
<td>3.5</td>
<td>Representing potential secondary reinforcers</td>
<td>129</td>
</tr>
<tr>
<td>3.5.1</td>
<td>The requirements of the representation</td>
<td>130</td>
</tr>
<tr>
<td>3.5.2</td>
<td>High capacity</td>
<td>133</td>
</tr>
<tr>
<td>3.5.3</td>
<td>Objects, and not their reward and punishment associations, are</td>
<td>133</td>
</tr>
<tr>
<td></td>
<td>represented in the inferior temporal visual cortex</td>
<td></td>
</tr>
<tr>
<td>3.5.4</td>
<td>Object representations</td>
<td>135</td>
</tr>
<tr>
<td>3.6</td>
<td>The orbitofrontal cortex</td>
<td>136</td>
</tr>
</tbody>
</table>
3.6.1 Historical background 136
3.6.2 Topology 137
3.6.3 Connections 139
3.6.4 Effects of damage to the orbitofrontal cortex 140
3.6.5 Neurophysiology and functional neuroimaging of the orbitofrontal cortex 142
3.6.6 The human orbitofrontal cortex 175
3.6.7 A computational basis for stimulus–reinforcer association learning and reversal in the orbitofrontal cortex 184
3.6.8 Executive functions of the orbitofrontal cortex 186
3.7 The amygdala 187
3.7.1 Connections of the amygdala 188
3.7.2 Effects of amygdala lesions 190
3.7.3 Neuronal activity in the primate amygdala to reinforcing stimuli 196
3.7.4 Responses of these amygdala neurons to reinforcing and novel stimuli 202
3.7.5 Neuronal responses in the amygdala to faces 204
3.7.6 Evidence from humans 205
3.7.7 Amygdala summary 210
3.8 The cingulate cortex 211
3.8.1 Anterior or perigenual cingulate cortex, reward, and affect 212
3.8.2 Mid-cingulate cortex, the cingulate motor area, and action–outcome learning 217
3.9 Human brain imaging investigations of mood and depression 219
3.10 Output pathways for reward- and punisher-guided behaviour, including emotional responses 220
3.10.1 The autonomic and endocrine systems 220
3.10.2 Motor systems for implicit responses, including the basal ganglia, reinforcement learning, and dopamine 221
3.10.3 Output systems for explicit responses to emotional stimuli 248
3.10.4 Basal forebrain and hypothalamus 249
3.10.5 Basal forebrain cholinergic neurons 249
3.10.6 Noradrenergic neurons 251
3.10.7 Opiate reward systems, analgesia, and food reward 252
3.11 Effects of emotion on cognitive processing and memory 253
3.12 Laterality effects in human reward and emotional processing 257
3.13 Summary 259
4 Invariant visual object recognition learning 262
4.1 Introduction 262
4.2 Invariant representations of faces and objects in the inferior temporal visual cortex 262
4.2.1 Processing to the inferior temporal cortex in the primate visual system 263
4.2.2 Translation invariance and receptive field size 264
4.2.3 Reduced translation invariance in natural scenes, and the selection of a rewarded object 265
4.2.4 Size and spatial frequency invariance 268
4.2.5 Combinations of features in the correct spatial configuration 269
4.2.6 A view-invariant representation 270
4.2.7 Learning in the inferior temporal cortex 274
4.2.8 Distributed encoding 276
4.2.9 Face expression, gesture, and view 281
4.2.10 Specialized regions in the temporal cortical visual areas 281

4.3 Approaches to invariant object recognition 285
4.3.1 Feature spaces 286
4.3.2 Structural descriptions and syntactic pattern recognition 287
4.3.3 Template matching and the alignment approach 289
4.3.4 Invertible networks that can reconstruct their inputs 290
4.3.5 Feature hierarchies 290

4.4 Hypotheses about object recognition mechanisms 295

4.5 Computational issues in feature hierarchies 298
4.5.1 The architecture of VisNet 299
4.5.2 Initial experiments with VisNet 307
4.5.3 The optimal parameters for the temporal trace used in the learning rule 314
4.5.4 Different forms of the trace learning rule, and their relation to error correction and temporal difference learning 315
4.5.5 The issue of feature binding, and a solution 324
4.5.6 Operation in a cluttered environment 335
4.5.7 Learning 3D transforms 342
4.5.8 Capacity of the architecture, and incorporation of a trace rule into a recurrent architecture with object attractors 347
4.5.9 Vision in natural scenes – effects of background versus attention 354
4.5.10 The representation of multiple objects in a scene 362
4.5.11 Learning invariant representations using spatial continuity: Continuous Spatial Transformation learning 364
4.5.12 Lighting invariance 365
4.5.13 Invariant global motion in the dorsal visual system 367

4.6 Further approaches to invariant object recognition 367

4.7 Visuo-spatial scratchpad memory, and change blindness 370

4.8 Processes involved in object identification 372

4.9 Conclusions 373

5 Short-term memory 375
5.1 Cortical short-term memory systems and attractor networks 375
5.2 Prefrontal cortex short-term memory networks, and their relation to percep-tual networks 378
5.3 Computational details of the model of short-term memory 381
5.4 Computational necessity for a separate, prefrontal cortex, short-term memory system 383
5.5 Synaptic modification is needed to set up but not to reuse short-term memory systems 384
5.6 What, where, and object–place combination short-term memory in the prefrontal cortex 384

6 Attention, short-term memory, and biased competition 386
6.1 Introduction 386
6.2 The classical view: the spotlight metaphor and feature integration theory 387
6.3 Biased competition – single neuron studies 391
  6.3.1 Neurophysiology of attention 392
  6.3.2 The role of competition 394
  6.3.3 Evidence for attentional bias 396
  6.3.4 Non-spatial attention 396
  6.3.5 High-resolution buffer hypothesis 398
6.4 Biased competition – fMRI 398
  6.4.1 Neuroimaging of attention 399
  6.4.2 Attentional effects in the absence of visual stimulation 399
6.5 A basic computational module for biased competition 401
6.6 Architecture of a model of attention 402
6.7 Simulations of basic experimental findings 407
  6.7.1 Simulations of single-cell experiments 408
  6.7.2 Simulations of fMRI experiments 410
6.8 Object recognition and spatial search 411
  6.8.1 Dynamics of spatial attention and object recognition 414
  6.8.2 Dynamics of object attention and visual search 416
6.9 The neuronal and biophysical mechanisms of attention 417
6.10 Linking computational and psychophysical data: ‘serial’ vs ‘parallel’ processing 421
  6.10.1 ‘Serial’ vs ‘parallel’ search 421
  6.10.2 Visual conjunction search 424
6.11 Linking computational and neuropsychological data on attention 429
  6.11.1 The neglect syndrome 429
  6.11.2 A model of visual spatial neglect 430
  6.11.3 Disengagement of attention in neglect 437
  6.11.4 Extinction and visual search 438
6.12 Conclusions 440
6.13 Attention – a formal model 443
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Probabilistic decision-making</td>
<td>449</td>
</tr>
<tr>
<td>7.1</td>
<td>Introduction</td>
<td>449</td>
</tr>
<tr>
<td>7.2</td>
<td>The neuronal data underlying the vibrotactile discrimination</td>
<td>451</td>
</tr>
<tr>
<td>7.3</td>
<td>Theoretical framework: a probabilistic attractor network</td>
<td>452</td>
</tr>
<tr>
<td>7.4</td>
<td>Stationary multistability analysis: mean-field</td>
<td>455</td>
</tr>
<tr>
<td>7.5</td>
<td>Non-stationary probabilistic analysis: spiking dynamics</td>
<td>458</td>
</tr>
<tr>
<td>7.6</td>
<td>Properties of this model of decision-making</td>
<td>465</td>
</tr>
<tr>
<td>7.7</td>
<td>Applications of this model of decision-making</td>
<td>469</td>
</tr>
<tr>
<td>7.8</td>
<td>The integrate-and-fire formulation used in the model of decision-making</td>
<td>470</td>
</tr>
<tr>
<td>7.9</td>
<td>The model parameters used in the simulations of decision-making</td>
<td>472</td>
</tr>
<tr>
<td>7.10</td>
<td>The mean-field approach used in the model of decision-making</td>
<td>474</td>
</tr>
<tr>
<td>8</td>
<td>Action selection by biased attractor competition in the prefrontal cortex</td>
<td>475</td>
</tr>
<tr>
<td>8.1</td>
<td>Introduction</td>
<td>475</td>
</tr>
<tr>
<td>8.2</td>
<td>A hierarchical attractor model of action selection</td>
<td>476</td>
</tr>
<tr>
<td>8.3</td>
<td>Setting up the synaptic connectivity for the prefrontal cortex</td>
<td>480</td>
</tr>
<tr>
<td>8.4</td>
<td>Pharmacology of attention, decision-making, and action selection</td>
<td>483</td>
</tr>
<tr>
<td>8.5</td>
<td>Application to a neurodynamical systems hypothesis of schizophrenia</td>
<td>485</td>
</tr>
<tr>
<td>8.6</td>
<td>Conclusions</td>
<td>494</td>
</tr>
<tr>
<td>9</td>
<td>Reward, decision, and action reversal using attractor dynamics</td>
<td>496</td>
</tr>
<tr>
<td>9.1</td>
<td>Stimulus–reinforcer association learning and reversal</td>
<td>496</td>
</tr>
<tr>
<td>9.2</td>
<td>Reversal of action selection</td>
<td>504</td>
</tr>
<tr>
<td>9.3</td>
<td>Sequence memory</td>
<td>506</td>
</tr>
<tr>
<td>9.4</td>
<td>Conclusions</td>
<td>508</td>
</tr>
<tr>
<td>10</td>
<td>Decision-making</td>
<td>509</td>
</tr>
<tr>
<td>10.1</td>
<td>Selection of mainly autonomic responses, and their classical conditioning</td>
<td>509</td>
</tr>
<tr>
<td>10.2</td>
<td>Selection of approach or withdrawal, and their classical conditioning</td>
<td>509</td>
</tr>
<tr>
<td>10.3</td>
<td>Selection of fixed stimulus–response habits</td>
<td>510</td>
</tr>
<tr>
<td>10.4</td>
<td>Selection of arbitrary behaviours to obtain goals, action–outcome learning, and emotional learning</td>
<td>510</td>
</tr>
<tr>
<td>10.5</td>
<td>The roles of the prefrontal cortex in decision-making and attention</td>
<td>511</td>
</tr>
<tr>
<td>10.5.1</td>
<td>Prefrontal attentional influences on perceptual processing</td>
<td>512</td>
</tr>
<tr>
<td>10.5.2</td>
<td>Attentional influences on mapping from stimuli to responses</td>
<td>512</td>
</tr>
<tr>
<td>10.5.3</td>
<td>Executive control</td>
<td>513</td>
</tr>
<tr>
<td>10.5.4</td>
<td>Disorders of attention and decision-making</td>
<td>514</td>
</tr>
<tr>
<td>10.6</td>
<td>Neuroeconomics, reward magnitude, expected value, and expected utility</td>
<td>515</td>
</tr>
<tr>
<td>10.6.1</td>
<td>Expected utility ≈ expected value = probability multiplied by reward magnitude</td>
<td>515</td>
</tr>
<tr>
<td>10.6.2</td>
<td>Delay of reward, emotional choice, and rational choice</td>
<td>516</td>
</tr>
<tr>
<td>10.6.3</td>
<td>Reward prediction error, temporal difference error, and choice</td>
<td>518</td>
</tr>
</tbody>
</table>
10.6.4 Reciprocal altruism, strong reciprocity, generosity, and altruistic punishment 519
10.7 Dual routes to action, and decision-making 523
10.8 Apostasis 529

A Introduction to linear algebra for neural networks 531
A.1 Vectors 531
A.1.1 The inner or dot product of two vectors 531
A.1.2 The length of a vector 532
A.1.3 Normalizing the length of a vector 533
A.1.4 The angle between two vectors: the normalized dot product 533
A.1.5 The outer product of two vectors 534
A.1.6 Linear and non-linear systems 535
A.1.7 Linear combinations of vectors, linear independence, and linear separability 536
A.2 Application to understanding simple neural networks 538
A.2.1 Capability and limitations of single-layer networks: linear separability and capacity 538
A.2.2 Non-linear networks: neurons with non-linear activation functions 540
A.2.3 Non-linear networks: neurons with non-linear activations 541

B Neural network models 543
B.1 Introduction 543
B.2 Pattern association memory 543
B.2.1 Architecture and operation 544
B.2.2 A simple model 547
B.2.3 The vector interpretation 549
B.2.4 Properties 550
B.2.5 Prototype extraction, extraction of central tendency, and noise reduction 553
B.2.6 Speed 553
B.2.7 Local learning rule 554
B.2.8 Implications of different types of coding for storage in pattern associators 559
B.3 Autoassociation or attractor memory 560
B.3.1 Architecture and operation 560
B.3.2 Introduction to the analysis of the operation of autoassociation networks 562
B.3.3 Properties 564
B.3.4 Use of autoassociation networks in the brain 570
B.4 Competitive networks, including self-organizing maps 571
B.4.1 Function 571
B.4.2 Architecture and algorithm 572
B.4.3 Properties 573
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.4.4</td>
<td>Utility of competitive networks in information processing by the brain</td>
<td>578</td>
</tr>
<tr>
<td>B.4.5</td>
<td>Guidance of competitive learning</td>
<td>579</td>
</tr>
<tr>
<td>B.4.6</td>
<td>Topographic map formation</td>
<td>582</td>
</tr>
<tr>
<td>B.4.7</td>
<td>Invariance learning by competitive networks</td>
<td>586</td>
</tr>
<tr>
<td>B.4.8</td>
<td>Radial Basis Function networks</td>
<td>588</td>
</tr>
<tr>
<td>B.4.9</td>
<td>Further details of the algorithms used in competitive networks</td>
<td>589</td>
</tr>
<tr>
<td>B.5</td>
<td>Continuous attractor networks</td>
<td>593</td>
</tr>
<tr>
<td>B.5.1</td>
<td>Introduction</td>
<td>593</td>
</tr>
<tr>
<td>B.5.2</td>
<td>The generic model of a continuous attractor network</td>
<td>595</td>
</tr>
<tr>
<td>B.5.3</td>
<td>Learning the synaptic strengths between the neurons that implement a continuous attractor network</td>
<td>595</td>
</tr>
<tr>
<td>B.5.4</td>
<td>The capacity of a continuous attractor network: multiple charts and packets</td>
<td>598</td>
</tr>
<tr>
<td>B.5.5</td>
<td>Continuous attractor models: path integration</td>
<td>598</td>
</tr>
<tr>
<td>B.5.6</td>
<td>Stabilization of the activity packet within the continuous attractor network when the agent is stationary</td>
<td>601</td>
</tr>
<tr>
<td>B.5.7</td>
<td>Continuous attractor networks in two or more dimensions</td>
<td>603</td>
</tr>
<tr>
<td>B.5.8</td>
<td>Mixed continuous and discrete attractor networks</td>
<td>603</td>
</tr>
<tr>
<td>B.6</td>
<td>Network dynamics: the integrate-and-fire approach</td>
<td>604</td>
</tr>
<tr>
<td>B.6.1</td>
<td>From discrete to continuous time</td>
<td>604</td>
</tr>
<tr>
<td>B.6.2</td>
<td>Continuous dynamics with discontinuities</td>
<td>606</td>
</tr>
<tr>
<td>B.6.3</td>
<td>An integrate-and-fire implementation</td>
<td>610</td>
</tr>
<tr>
<td>B.6.4</td>
<td>Simulation of fMRI signals: hemodynamic convolution of synaptic activity</td>
<td>611</td>
</tr>
<tr>
<td>B.6.5</td>
<td>The speed of processing of one-layer attractor networks with integrate-and-fire neurons</td>
<td>613</td>
</tr>
<tr>
<td>B.6.6</td>
<td>The speed of processing of a four-layer hierarchical network with integrate-and-fire attractor dynamics in each layer</td>
<td>616</td>
</tr>
<tr>
<td>B.6.7</td>
<td>Spike response model</td>
<td>619</td>
</tr>
<tr>
<td>B.7</td>
<td>Network dynamics: introduction to the mean-field approach</td>
<td>620</td>
</tr>
<tr>
<td>B.8</td>
<td>Mean-field based neurodynamics</td>
<td>621</td>
</tr>
<tr>
<td>B.8.1</td>
<td>Population activity</td>
<td>622</td>
</tr>
<tr>
<td>B.8.2</td>
<td>A basic computational module based on biased competition</td>
<td>624</td>
</tr>
<tr>
<td>B.8.3</td>
<td>Multimodular neurodynamical architectures</td>
<td>625</td>
</tr>
<tr>
<td>B.9</td>
<td>Interacting attractor networks</td>
<td>627</td>
</tr>
<tr>
<td>B.10</td>
<td>Sequence memory implemented by adaptation in an attractor network</td>
<td>631</td>
</tr>
<tr>
<td>B.11</td>
<td>Error correction networks</td>
<td>631</td>
</tr>
<tr>
<td>B.11.1</td>
<td>Architecture and general description</td>
<td>632</td>
</tr>
<tr>
<td>B.11.2</td>
<td>Generic algorithm for a one-layer error correction network</td>
<td>632</td>
</tr>
<tr>
<td>B.11.3</td>
<td>Capability and limitations of single-layer error-correcting networks</td>
<td>633</td>
</tr>
<tr>
<td>B.11.4</td>
<td>Properties</td>
<td>636</td>
</tr>
<tr>
<td>B.12</td>
<td>Error backpropagation multilayer networks</td>
<td>638</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>B.12.1</td>
<td>Introduction</td>
<td>638</td>
</tr>
<tr>
<td>B.12.2</td>
<td>Architecture and algorithm</td>
<td>639</td>
</tr>
<tr>
<td>B.12.3</td>
<td>Properties of multilayer networks trained by error backpropagation</td>
<td>640</td>
</tr>
<tr>
<td>B.13</td>
<td>Biologically plausible networks</td>
<td>641</td>
</tr>
<tr>
<td>B.14</td>
<td>Contrastive Hebbian learning: the Boltzmann machine</td>
<td>642</td>
</tr>
<tr>
<td>B.15</td>
<td>Reinforcement learning</td>
<td>644</td>
</tr>
<tr>
<td>B.15.1</td>
<td>Associative reward–penalty algorithm of Barto and Sutton</td>
<td>645</td>
</tr>
<tr>
<td>B.15.2</td>
<td>Reward prediction error or delta rule learning, and classical conditioning</td>
<td>646</td>
</tr>
<tr>
<td>B.15.3</td>
<td>Temporal Difference (TD) learning</td>
<td>647</td>
</tr>
<tr>
<td>B.16</td>
<td>Forgetting in associative neural networks and in the brain, and memory reconsolidation</td>
<td>650</td>
</tr>
<tr>
<td>B.17</td>
<td>Brain computation compared to computation on a digital computer</td>
<td>654</td>
</tr>
<tr>
<td>C</td>
<td>Information theory, and neuronal encoding</td>
<td>659</td>
</tr>
<tr>
<td>C.1</td>
<td>Information theory</td>
<td>660</td>
</tr>
<tr>
<td>C.1.1</td>
<td>The information conveyed by definite statements</td>
<td>660</td>
</tr>
<tr>
<td>C.1.2</td>
<td>Information conveyed by probabilistic statements</td>
<td>661</td>
</tr>
<tr>
<td>C.1.3</td>
<td>Information sources, information channels, and information measures</td>
<td>662</td>
</tr>
<tr>
<td>C.1.4</td>
<td>The information carried by a neuronal response and its averages</td>
<td>663</td>
</tr>
<tr>
<td>C.1.5</td>
<td>The information conveyed by continuous variables</td>
<td>666</td>
</tr>
<tr>
<td>C.2</td>
<td>The information carried by neuronal responses</td>
<td>668</td>
</tr>
<tr>
<td>C.2.1</td>
<td>The limited sampling problem</td>
<td>668</td>
</tr>
<tr>
<td>C.2.2</td>
<td>Correction procedures for limited sampling</td>
<td>669</td>
</tr>
<tr>
<td>C.2.3</td>
<td>The information from multiple cells: decoding procedures</td>
<td>670</td>
</tr>
<tr>
<td>C.2.4</td>
<td>Information in the correlations between the spikes of different cells: a decoding approach</td>
<td>674</td>
</tr>
<tr>
<td>C.2.5</td>
<td>Information in the correlations between the spikes of different cells: a second derivative approach</td>
<td>679</td>
</tr>
<tr>
<td>C.3</td>
<td>Information theory results</td>
<td>682</td>
</tr>
<tr>
<td>C.3.1</td>
<td>The sparseness of the distributed encoding used by the brain</td>
<td>683</td>
</tr>
<tr>
<td>C.3.2</td>
<td>The information from single neurons</td>
<td>693</td>
</tr>
<tr>
<td>C.3.3</td>
<td>The information from single neurons: temporal codes versus rate codes within the spike train of a single neuron</td>
<td>697</td>
</tr>
<tr>
<td>C.3.4</td>
<td>The information from single neurons: the speed of information transfer</td>
<td>698</td>
</tr>
<tr>
<td>C.3.5</td>
<td>The information from multiple cells: independent information versus redundancy across cells</td>
<td>709</td>
</tr>
<tr>
<td>C.3.6</td>
<td>Should one neuron be as discriminative as the whole organism, in object encoding systems?</td>
<td>713</td>
</tr>
<tr>
<td>C.3.7</td>
<td>The information from multiple cells: the effects of cross-correlations between cells</td>
<td>715</td>
</tr>
</tbody>
</table>
4 Invariant visual object recognition learning

4.1 Introduction

One of the major problems that is solved by the visual system in the cerebral cortex is the building of a representation of visual information which allows object and face recognition to occur relatively independently of size, contrast, spatial frequency, position on the retina, angle of view, etc. The importance of and requirements for these representations are described in Section 3.5. These invariant representations of objects, provided by the inferior temporal visual cortex (as described in Section 4.2, Appendix C, and Section 2.2.6), are extremely important for the operation of other memory systems in the brain, for if there is an invariant representation, it is possible to learn on a single trial about reward/punishment associations of the object, the place where that object is located, and whether the object has been seen recently, and then to correctly generalize to other views etc. of the same object. The way in which these invariant representations of objects are formed is a major issue in understanding learning and memory, for with this type of learning, we must not only store and retrieve information, but we must solve in addition the major computational problem of how all the different images on the retina (position, size, view etc.) of an object can be mapped to the same representation of that object in the brain. It is this learning process with which we are concerned in this Chapter.

In the first part of this Chapter, we consider the nature of the representations of objects and faces that are found in the inferior temporal visual cortex. Then later in the Chapter we consider the major computational issue of how these invariant representations of objects and faces may be formed by self-organizing learning in the brain.

Before considering invariant object and face recognition in the inferior temporal visual cortex, we note that this builds on the preprocessing performed in V1, the striate visual cortex of primates (described for example in Chapter 2 of Rolls and Deco (2002)). An outline of the overall architecture of the visual pathways is shown in Fig. 4.1. The model to be described in Section 4.5 also builds on what is known about visual processing in extrastriate visual areas (described for example in Chapter 3 of Rolls and Deco (2002)), and the inferior temporal visual cortex (described next in Section 4.2) (Rolls and Deco 2002, Rolls 2007i, Rolls 2008a).

4.2 Invariant representations of faces and objects in the inferior temporal visual cortex

Some of the desirable properties of an object recognition system, which include position, size and view invariant representations, and distributed representations that allow for generalization, are described in Section 3.5. These are properties of the inferior temporal visual cortex, as described next in this Section (4.2).
### 4.2.1 Processing to the inferior temporal cortex in the primate visual system

A schematic diagram to indicate some aspects of the processing involved in object identification from the primary visual cortex, V1, through V2 and V4 to the posterior inferior temporal cortex (TEO) and the anterior inferior temporal cortex (TE) is shown in Fig. 4.2. The approximate location of these visual cortical areas on the brain of a macaque monkey is shown in Figs. 4.3 and 1.9, which also show that TE has a number of different subdivisions. The different TE areas all contain visually responsive neurons, as do many of the areas within the cortex in the superior temporal sulcus (Baylis, Rolls and Leonard 1987). For the purposes of this summary, these areas will be grouped together as the anterior inferior temporal cortex (IT), except where otherwise stated. A description of some of the specializations within this region, and a hypothesis about why there is specialization, is provided in Section 4.2.10. Some of the information processing that takes place through these pathways that must be addressed by computational models is described in the following subsections. A fuller account is provided by Rolls (2000a), Rolls and Deco (2002), Rolls (2007i), Rolls (2007f), and Rolls (2008a).

Many of the studies on neurons in the anterior inferior temporal cortex and cortex in the
superior temporal sulcus have been performed with neurons that respond particularly to faces, because such neurons can be found regularly in recordings in this region, and therefore provide a good population for systematic studies (Rolls 2000a, Rolls and Deco 2002, Rolls 2007i, Rolls and Deco 2006). The face-selective neurons described in this book are found mainly between 7 mm and 3 mm posterior to the sphenoid reference, which in a 3–4 kg macaque corresponds to approximately 11 to 15 mm anterior to the interaural plane (Baylis, Rolls and Leonard 1987, Rolls 2007i, Rolls 2007f). The ‘middle face patch’ of Tsao, Freiwald, Tootell and Livingstone (2006) was at A6, which is probably part of the posterior inferior temporal cortex. In the anterior inferior temporal cortex areas we have investigated, there are separate regions specialized for face identity in areas T Ea and TEM on the ventral lip of the superior temporal sulcus and the adjacent gyrus, and for face expression and movement in the cortex deep in the superior temporal sulcus (Hasselmo, Rolls and Baylis 1989a, Baylis, Rolls and Leonard 1987, Rolls 2007i) (see Sections 4.2.9 and 4.2.10).

Examples of some of the objects encoded by anterior inferior temporal cortex neurons as described in this book (for example in this Chapter, in Sections 2.2.6 and 3.5, and in Appendix C) are shown in Figs. 2.11 and C.7.

4.2.2 Translation invariance and receptive field size

There is convergence from each small part of a region to the succeeding region (or layer in the hierarchy) in such a way that the receptive field sizes of neurons (for example 1 degree near the fovea in V1) become larger by a factor of approximately 2.5 with each succeeding stage. [The typical parafoveal receptive field sizes found would not be inconsistent with the calculated approximations of, for example, 8 degrees in V4, 20 degrees in TEO, and 50 degrees in inferior temporal cortex (Boussaoud, Desimone and Ungerleider 1991) (see Fig. 4.2).] Such zones of convergence would overlap continuously with each other (see Fig. 4.2). This connectivity provides part of the basis for the fact that many neurons in the temporal cortical visual areas respond to a stimulus relatively independently of where it is in their receptive field, and moreover maintain their stimulus selectivity when the stimulus appears in different parts of the visual field (Gross, Desimone, Albright and Schwartz 1985, Tovee, Rolls...
and Azzopardi 1994, Rolls, Aggelopoulos and Zheng 2003a). This is called translation or shift invariance. In addition to having topologically appropriate connections, it is necessary for the connections to have the appropriate synaptic weights to perform the mapping of each set of features, or object, to the same set of neurons in IT. How this could be achieved is addressed in the computational neuroscience models described later in this Chapter and by Wallis and Rolls (1997), Rolls and Deco (2002), Rolls and Stringer (2006), and Stringer, Perry, Rolls and Proske (2006).

### 4.2.3 Reduced translation invariance in natural scenes, and the selection of a rewarded object

Until recently, research on translation invariance considered the case in which there is only one object in the visual field. What happens in a cluttered, natural, environment? Do all objects that can activate an inferior temporal neuron do so whenever they are anywhere within the large receptive fields of inferior temporal neurons (cf. Sato (1989))? If so, the output of the visual system might be confusing for structures that receive inputs from the temporal cortical visual areas. If one of the objects in the visual field was associated with reward, and another with punishment, would the output of the inferior temporal visual cortex to emotion-related brain systems be an amalgam of both stimuli? If so, how would we be able to choose between the stimuli, and have an emotional response to one but not perhaps the other, and select one for action and not the other (see Fig. 4.4)?

In an investigation of this, it was found that the mean firing rate across all cells to a fixated effective face with a non-effective face in the parafovea (centred 8.5 degrees from the fovea)
was 34 spikes/s. On the other hand, the average response to a fixated non-effective face with an effective face in the periphery was 22 spikes/s (Rolls and Tovee 1995a). Thus these cells gave a reliable output about which stimulus is actually present at the fovea, in that their response was larger to a fixated effective face than to a fixated non-effective face, even when there were other parafoveal stimuli effective for the neuron. Thus the neurons provide information biased towards what is present at the fovea, and not equally about what is present anywhere in the visual field. This makes the interface to action simpler, in that what is at the fovea can be interpreted (e.g. by an associative memory) partly independently of the surroundings, and choices and actions can be directed if appropriate to what is at the fovea (cf. Ballard (1993)). These findings are a step towards understanding how the visual system functions in a normal environment (see also Gallant, Connor and Van-Essen (1998), Stringer and Rolls (2000), and Rolls and Deco (2006)).

To investigate further how information is passed from the inferior temporal cortex (IT) to other brain regions to enable stimuli to be selected from natural scenes for action, Rolls, Aggelopoulos and Zheng (2003a) analysed the responses of single and simultaneously recorded IT neurons to stimuli presented in complex natural backgrounds. In one situation, a visual fixation task was performed in which the monkey fixated at different distances from the
effective stimulus. In another situation the monkey had to search for two objects on a screen, and a touch of one object was rewarded with juice, and of another object was punished with saline (see Fig. 4.4 for a schematic overview, and Fig. 4.62 on page 354 for the actual display). In both situations neuronal responses to the effective stimuli for the neurons were compared when the objects were presented in the natural scene or on a plain background. It was found that the overall response of the neuron to objects was sometimes somewhat reduced when they were presented in natural scenes, though the selectivity of the neurons remained. However, the main finding was that the magnitudes of the responses of the neurons typically became much less in the real scene the further the monkey fixated in the scene away from the object (see Figs. 4.5 and 4.63, and Section 4.5.9.1). It is proposed that this reduced translation invariance in natural scenes helps an unambiguous representation of an object which may be the target for action to be passed to the brain regions that receive from the primate inferior temporal visual cortex. It helps with the binding problem, by reducing in natural scenes the effective receptive field of inferior temporal cortex neurons to approximately the size of an object in the scene. The computational utility and basis for this is considered in Section 4.5.9 and by Rolls and Deco (2002), Trappenberg, Rolls and Stringer (2002), Deco and Rolls (2004), Aggelopoulos and Rolls (2005), and Rolls and Deco (2006), and includes an advantage for what is at the fovea because of the large cortical magnification of the fovea, and shunting interactions between representations weighted by how far they are from the fovea.

These findings suggest that the principle of providing strong weight to whatever is close to the fovea is an important principle governing the operation of the inferior temporal visual cortex, and in general of the output of the ventral visual system in natural environments. This principle of operation is very important in interfacing the visual system to action systems, because the effective stimulus in making inferior temporal cortex neurons fire is in natural
scenes usually on or close to the fovea. This means that the spatial coordinates of where the object is in the scene do not have to be represented in the inferior temporal visual cortex, nor passed from it to the action selection system, as the latter can assume that the object making IT neurons fire is close to the fovea in natural scenes. Thus the position in visual space being fixated provides part of the interface between sensory representations of objects and their coordinates as targets for actions in the world. The small receptive fields of IT neurons in natural scenes make this possible. After this, local, egocentric, processing implemented in the dorsal visual processing stream using e.g. stereodisparity may be used to guide action towards objects being fixated (Rolls and Deco 2002).

The reduced receptive field size in complex natural scenes also enables emotions to be selective to just what is being fixated, because this is the information that is transmitted by the firing of IT neurons to structures such as the orbitofrontal cortex and amygdala.

Interestingly, although the size of the receptive fields of inferior temporal cortex neurons become reduced in natural scenes so that neurons in IT respond primarily to the object being fixated, there is nevertheless frequently some asymmetry in the receptive fields (see Section 4.5.10 and Fig. 4.67). This provides a partial solution to how multiple objects and their position in a scene can be captured with a single glance (Aggelopoulos and Rolls 2005).

### 4.2.4 Size and spatial frequency invariance

Some neurons in the inferior temporal visual cortex and cortex in the anterior part of the superior temporal sulcus (IT/STS) respond relatively independently of the size of an effective face stimulus, with a mean size invariance (to a half maximal response) of 12 times (3.5 octaves) (Rolls and Baylis 1986). An example of the responses of an inferior temporal cortex face-selective neuron to faces of different sizes is shown in Fig. 4.6. This is not a property of a simple single-layer network (see Fig. 4.18), nor of neurons in V1, which respond best to small stimuli, with a typical size-invariance of 1.5 octaves. Also, the neurons typically responded to a face when the information in it had been reduced from 3D to a 2D representation in grey on a monitor, with a response that was on average 0.5 of that to a real face.

Another transform over which recognition is relatively invariant is spatial frequency. For example, a face can be identified when it is blurred (when it contains only low spatial frequencies), and when it is high-pass spatial frequency filtered (when it looks like a line drawing). If the face images to which these neurons respond are low-pass filtered in the
spatial frequency domain (so that they are blurred), then many of the neurons still respond when the images contain frequencies only up to 8 cycles per face. Similarly, the neurons still respond to high-pass filtered images (with only high spatial frequency edge information) when frequencies down to only 8 cycles per face are included (Rolls, Baylis and Leonard 1985).

Face recognition shows similar invariance with respect to spatial frequency (see Rolls, Baylis and Leonard (1985)). Further analysis of these neurons with narrow (octave) bandpass spatial frequency filtered face stimuli shows that the responses of these neurons to an unfiltered face cannot be predicted from a linear combination of their responses to the narrow band stimuli (Rolls, Baylis and Hasselmo 1987). This lack of linearity of these neurons, and their responsiveness to a wide range of spatial frequencies (see also their broad critical band masking (Rolls 2008a)), indicate that in at least this part of the primate visual system recognition does not occur using Fourier analysis of the spatial frequency components of images.

The utility of this representation for memory systems in the brain is that the output of the visual system will represent an object invariantly with respect to position on the retina, size, etc, and this simplifies the functionality required of the (multiple) memory systems, which need then simply associate the object representation with reward (orbitofrontal cortex and amygdala), associate it with position in the environment (hippocampus), recognise it as familiar (perirhinal cortex), associate it with a motor response in a habit memory (basal ganglia), etc. (Rolls 2007g). The associations can be relatively simple, involving for example Hebbian associativity (see chapters throughout this book, and Appendix B).

Some neurons in the temporal cortical visual areas actually represent the absolute size of objects such as faces independently of viewing distance (Rolls and Baylis 1986). This could be called neurophysiological size constancy. The utility of this representation by a small population of neurons is that the absolute size of an object is a useful feature to use as an input to neurons that perform object recognition. Faces only come in certain sizes.

### 4.2.5 Combinations of features in the correct spatial configuration

Many neurons in this processing stream respond to combinations of features (including objects), but not to single features presented alone, and the features must have the correct spatial arrangement. This has been shown, for example, with faces, for which it has been shown by masking out or presenting parts of the face (for example eyes, mouth, or hair) in isolation, or by jumbling the features in faces, that some cells in the cortex in IT/STS respond only if two or more features are present, and are in the correct spatial arrangement (Perrett, Rolls and Caan 1982, Rolls, Tovee, Purcell, Stewart and Azzopardi 1994b). Fig. 4.7 shows examples of four neurons, the top one of which responds only if all the features are present, and the others of which respond not only to the full face, but also to one or more features. Corresponding evidence has been found for non-face cells. For example, Tanaka, Saito, Fukada and Moriya (1990) showed that some posterior inferior temporal cortex neurons might only respond to the combination of an edge and a small circle if they were in the correct spatial relationship to each other.

Evidence consistent with the suggestion that neurons are responding to combinations of a few variables represented at the preceding stage of cortical processing is that some neurons in V2 and V4 respond to end-stopped lines, to tongues flanked by inhibitory subregions, to combinations of lines, or to combinations of colours (see Rolls and Deco (2002), Hegde and Van Essen (2000) and Ito and Komatsu (2004)). Neurons that respond to combinations of features but not to single features indicate that the system is non-linear (Elliffe, Rolls and Stringer 2002).

The fact that some temporal cortex neurons respond only to objects or faces consisting of a set of features only if the whole combination of features is present and they are in the
correct spatial arrangement with respect to each other (not jumbled) is part of the evidence that some inferior temporal cortex neurons are tuned to respond to objects or faces.

If a task requires in addition to the normal representation of objects by the inferior temporal cortex, the ability to discriminate between stimuli composed of highly overlapping feature conjunctions in a low-dimensional feature space, then the perirhinal cortex may contribute to this type of discrimination, as described in Section 2.2.6.

4.2.6 A view-invariant representation

For recognizing and learning about objects (including faces), it is important that an output of the visual system should be not only translation- and size-invariant, but also relatively view-invariant. In an investigation of whether there are such neurons, we found that some
temporal cortical neurons reliably responded differently to the faces of two different individuals independently of viewing angle (Hasselmo, Rolls, Baylis and Nalwa 1989b) (see example in Fig. 4.8 upper), although in most cases (16/18 neurons) the response was not perfectly view-independent. Mixed together in the same cortical regions there are neurons with view-dependent responses (for example Hasselmo, Rolls, Baylis and Nalwa (1989b) and Rolls and Tovee (1995b) (see example in Fig. 4.8 lower)). Such neurons might respond, for example, to a view of a profile of a monkey but not to a full-face view of the same monkey (Perrett et al. 1985b, Hasselmo et al. 1989b).

These findings of view-dependent, partially view-independent, and view-independent representations in the same cortical regions are consistent with the hypothesis discussed below that view-independent representations are being built in these regions by associating together the outputs of neurons that have different view-dependent responses to the same individual. These findings also provide evidence that one output of the visual system includes representations of what is being seen, in a view-independent way that would be useful for object recognition and for learning associations about objects; and that another output is a
Fig. 4.9 Object-centred encoding: Neuron responding to ventral flexion (V) of the head independently of view even when the head was inverted. Ventral flexion is tilting the head from the full frontal view forwards to face down at 45 degrees. Dorsal flexion is tilting the head until it is looking 45 degrees up. The response is shown as a change in firing rate in spikes/s ± sem from the spontaneous firing rate. (After Hasselmo, Rolls, Baylis and Nalwa 1989b.)

View-based representation that would be useful in social interactions to determine whether another individual is looking at one, and for selecting details of motor responses, for which the orientation of the object with respect to the viewer is required (Rolls and Deco 2002).

Further evidence that some neurons in the temporal cortical visual areas have object-based rather than view-based responses comes from a study of a population of neurons that responds to moving faces (Hasselmo, Rolls, Baylis and Nalwa 1989b). For example, four neurons responded vigorously to a head undergoing ventral flexion, irrespective of whether the view of the head was full face, of either profile, or even of the back of the head. These different views could only be specified as equivalent in object-based coordinates. Further, the movement specificity was maintained across inversion, with neurons responding for example to ventral flexion of the head irrespective of whether the head was upright or inverted (see example in Fig. 4.9). In this procedure, retinally encoded or viewer-centered movement vectors are reversed, but the object-based description remains the same.

Also consistent with object-based encoding is the finding of a small number of neurons that respond to images of faces of a given absolute size, irrespective of the retinal image size or distance (Rolls and Baylis 1986).

Neurons with view invariant responses to objects seen naturally by macaques have also been described (Booth and Rolls 1998). The stimuli were presented for 0.5 s on a colour video monitor while the monkey performed a visual fixation task. The stimuli were images of 10 real plastic objects that had been in the monkey’s cage for several weeks, to enable him to build view invariant representations of the objects. Control stimuli were views of objects that had never been seen as real objects. The neurons analyzed were in the TE cortex in and close to the ventral lip of the anterior part of the superior temporal sulcus. Many neurons were found that responded to some views of some objects. However, for a smaller number of neurons, the responses occurred only to a subset of the objects (using ensemble encoding), irrespective of the viewing angle. Moreover, the firing of a neuron on any one trial, taken at random and irrespective of the particular view of any one object, provided information about
which object had been seen, and this information increased approximately linearly with the number of neurons in the sample (see Fig. 4.10). This is strong quantitative evidence that some neurons in the inferior temporal cortex provide an invariant representation of objects. Moreover, the results of Booth and Rolls (1998) show that the information is available in the firing rates, and has all the desirable properties of distributed representations described above, including exponentially high coding capacity, and rapid speed of readout of the information.

Further evidence consistent with these findings is that some studies have shown that the responses of some visual neurons in the inferior temporal cortex do not depend on the presence or absence of critical features for maximal activation (Perrett, Rolls and Caan 1982, Tanaka 1993, Tanaka 1996). For example, the responses of neuron 4 in Fig. 4.7 responded to several of the features in a face when these features were presented alone (Perrett, Rolls and Caan 1982). In another example, Mikami, Nakamura and Kubota (1994) showed that some TE cells respond to partial views of the same laboratory instrument(s), even when these partial views contain different features. Such functionality is important for object recognition when part of an object is occluded, by for example another object. In a different approach, Logothetis, Pauls, Bulthoff and Poggio (1994) have reported that in monkeys extensively trained (over thousands of trials) to treat different views of computer generated wire-frame 'objects' as the same, a small population of neurons in the inferior temporal cortex did respond to different views of the same wire-frame object (see also Logothetis and Sheinberg (1996)). However, extensive training is not necessary for invariant representations to be formed, and indeed no explicit training in invariant object recognition was given in the experiment by Booth and Rolls (1998), as Rolls' hypothesis (1992a) is that view invariant representations can be learned by associating together the different views of objects as they are moved and inspected naturally in a period that may be in the order of a few seconds. Evidence for this is described in Section 4.2.7.
4.2.7 Learning of new representations in the temporal cortical visual areas

To investigate the idea that visual experience might guide the formation of the responsiveness of neurons so that they provide an economical and ensemble-encoded representation of items actually present in the environment, the responses of inferior temporal cortex face-selective neurons have been analyzed while a set of new faces were shown. Some of the neurons studied in this way altered the relative degree to which they responded to the different members of the set of novel faces over the first few (1–2) presentations of the set (Rolls, Baylis, Hasselmo and Nalwa 1989a) (see examples in Fig. 4.11). If in a different experiment a single novel face was introduced when the responses of a neuron to a set of familiar faces were being recorded, the responses to the set of familiar faces were not disrupted, while the responses to the novel face became stable within a few presentations. Alteration of the tuning of individual neurons in this way may result in a good discrimination over the population as a whole of the faces known to the monkey. This evidence is consistent with the categorization being performed by self-organizing competitive neuronal networks, as described in Section B.4 and elsewhere (Rolls and Treves 1998).
Invariant representations of faces and objects in the inferior temporal visual cortex

Fig. 4.12 Images of the type used to investigate rapid learning in the neurophysiological experiments of Tovee, Rolls and Ramachandran (1996) and the PET imaging study of Dolan, Fink, Rolls et al. (1997). When the black and white (two-tone) images at the left are shown, the objects or faces are not generally recognized. After the full grey scale images (middle) have been shown for a few seconds, humans and inferior temporal cortex neurons respond to the faces or objects in the black and white images (right).

Further evidence that these neurons can learn new representations very rapidly comes from an experiment in which binarized black and white (two-tone) images of faces that blended with the background were used (see Fig. 4.12, left). These did not activate face-selective neurons. Full grey-scale images of the same photographs were then shown for ten 0.5 s presentations (Fig. 4.12, middle). In a number of cases, if the neuron happened to be responsive to that face, when the binarized version of the same face was shown next (Fig. 4.12, right), the neurons responded to it (Tovee, Rolls and Ramachandran 1996). This is a direct parallel to the same phenomenon that is observed psychophysically, and provides dramatic evidence that these neurons are influenced by only a very few seconds (in this case 5 s) of experience with a visual stimulus. We have shown a neural correlate of this effect using similar stimuli and a similar paradigm in a PET (positron emission tomography) neuroimaging study in humans, with a region showing an effect of the learning found for faces in the right temporal lobe, and for objects in the left temporal lobe (Dolan, Fink, Rolls, Booth, Holmes, Frackowiak and Friston 1997).

Such rapid learning of representations of new objects appears to be a major type of learning in which the temporal cortical areas are involved. Ways in which this learning could occur are considered later in this Chapter. In addition, some of these neurons may be involved in a short term memory for whether a particular familiar visual stimulus (such as a face) has been seen recently. The evidence for this is that some of these neurons respond differently to recently seen stimuli in short term visual memory tasks (Baylis and Rolls 1987, Miller and Desimone 1994, Xiang and Brown 1998). In the inferior temporal visual cortex proper, neurons respond more to novel than to familiar stimuli, but treat the stimuli as novel if more than one other stimuli intervene between the first (novel) and second (familiar) presentations of a particular stimulus (Baylis and Rolls 1987). More ventrally, in what is in or close to the
perirhinal cortex, these memory spans may hold for several intervening stimuli in the same task (Xiang and Brown 1998) (see Section 2.2.6). Some neurons in these areas respond more when a sample stimulus reappears in a delayed match to sample task with intervening stimuli (Miller and Desimone 1994), and the basis for this using a short term memory implemented in the prefrontal cortex is described in Chapter 5. Neurons in the more ventral (perirhinal) cortical area respond during the delay in a match to sample task with a delay between the sample stimulus and the to-be-matched stimulus (Miyashita 1993, Renart, Parga and Rolls 2000) (see Section 2.2.6).

### 4.2.8 Distributed encoding

An important question for understanding brain function is whether a particular object (or face) is represented in the brain by the firing of one or a few gnostic (or ‘grandmother’) cells (Barlow 1972), or whether instead the firing of a group or ensemble of cells each with somewhat different responsiveness provides the representation. Advantages of distributed codes (see Section 1.6, Appendix B, Appendix C, Rolls and Treves (1998), and Rolls and Deco (2002)) include generalization and graceful degradation (fault tolerance), and a potentially very high capacity in the number of stimuli that can be represented (that is exponential growth of capacity with the number of neurons in the representation). If the ensemble encoding is sparse, this provides a good input to an associative memory, for then large numbers of stimuli can be stored (see Appendix B of this book, Chapters 2 and 3 of Rolls and Treves (1998), and Chapter 7 of Rolls and Deco (2002)). We have shown that in the inferior temporal visual cortex and cortex in the anterior part of the superior temporal sulcus (IT/STS), responses of a group of neurons, but not of a single neuron, provide evidence on which face was shown. We showed, for example, that these neurons typically respond with a graded set of firing to different faces, with firing rates from 120 spikes/s to the most effective face, to no response at all to a number of the least effective faces (Baylis, Rolls and Leonard 1985, Rolls and Tovee 1995b, Rolls and Deco 2002). In fact, the firing rate probability distribution of a single neuron to a set of stimuli is approximately exponential (Rolls and Tovee 1995b, Treves, Panzeri, Rolls, Booth and Wakeman 1999, Rolls and Deco 2002, Baddeley, Abbott, Booth, Sengpiel, Freeman, Wakeman and Rolls 1997, Franco, Rolls, Aggelopoulos and Jerez 2007).

To provide examples, Fig. 4.13 shows typical firing rate changes of a single neuron on different trials to each of several different faces. This makes it clear that from the firing rate on any one trial, information is available about which stimulus was shown, and that the firing rate is graded, with a different firing rate response of the neuron to each stimulus.

The distributed nature of the encoding typical for neurons in the inferior temporal visual cortex is illustrated in Fig. 4.14, which shows that temporal cortical neurons typically responded to several members of a set of five faces, with each neuron having a different profile of responses to each face (Baylis, Rolls and Leonard 1985). It would be difficult for most of these single cells to tell which of even five faces, let alone which of hundreds of faces, had been seen. Yet across a population of such neurons, much information about the particular face that has been seen is provided, as shown below.

The single neuron selectivity or sparseness $\alpha^s$ of the activity of inferior temporal cortex neurons was 0.65 over a set of 68 stimuli including 23 faces and 45 non-face natural scenes, and a measure called the response sparseness $\alpha_r^s$ of the representation, in which the spontaneous rate was subtracted from the firing rate to each stimulus so that the responses of the neuron were being assessed, was 0.38 across the same set of stimuli (Rolls and Tovee 1995b). [For the definition of sparseness see Section C.3.1. For binary neurons (firing for example either at a high rate or not at all), the single neuron sparseness is the proportion of stimuli that a single neuron responds to. These definitions, and what is found in cortical neuronal representations,
are described further in Sections 1.6, C.3.1 and C.3.2, by Rolls and Deco (2002), and by
Franco, Rolls, Aggelopoulos and Jerez (2007).]

It has been possible to apply information theory to show that each neuron conveys on
average approximately 0.4 bits of information about which face in a set of 20 faces has
been seen (Tovee and Rolls 1995, Tovee, Rolls, Treves and Bellis 1993, Rolls, Treves, Tovee
and Panzeri 1997d). If a neuron responded to only one of the faces in the set of 20, then it
could convey (if noiseless) 4.6 bits of information about one of the faces (when that face was
shown). If, at the other extreme, it responded to half the faces in the set, it would convey
1 bit of information about which face had been seen on any one trial. In fact, the average
Fig. 4.14 Responses of four different temporal cortex visual neurons to a set of five faces (A–E), and, for comparison, to a wide range of non-face objects and foods. F–J are non-face stimuli. The means and standard errors of the responses computed over 8–10 trials are shown. (From Baylis, Rolls and Leonard 1985.)

maximum information about the best stimulus was 1.8 bits of information. This provides good evidence not only that the representation is distributed, but also that it is a sufficiently reliable
Fig. 4.15 (a) The values for the average information available in the responses of different numbers of these neurons on each trial, about which of a set of 20 face stimuli has been shown. The decoding method was Dot Product (DP, ×) or Probability Estimation (PE, +). The full line indicates the amount of information expected from populations of increasing size, when assuming random correlations within the constraint given by the ceiling (the information in the stimulus set, $I = 4.32$ bits). (b) The percent correct for the corresponding data to those shown in (a). The measurement period was 500 ms. (After Rolls, Treves and Tovee 1997b.)

An important result is that when the information available from a population of neurons about which of 20 faces has been seen is considered, the information increases approximately linearly as the number of cells in the population increases from 1 to 14 (Rolls, Treves and Tovee 1997b, Abbott, Rolls and Tovee 1996) (see Fig. 4.15 and Section C.3.5). Remembering that the information in bits is a logarithmic measure, this shows that the representational capacity of this population of cells increases exponentially (see Fig. 4.16). This is the case both when an optimal, probability estimation, form of decoding of the activity of the neuronal population is used, and also when the neurally plausible dot product type of decoding is used (Fig. 4.15). (The dot product decoding assumes that what reads out the information from the population activity vector is a neuron or a set of neurons that operates just by forming the dot product of the input population vector and its synaptic weight vector – see Rolls, Treves and Tovee (1997b), and Appendix B.) By simulation of further neurons and further stimuli, we have shown that the capacity grows very impressively, approximately as shown in Fig. 4.16 (Abbott, Rolls and Tovee 1996). The result has been replicated with simultaneously recorded neurons (Rolls, Franco, Aggelopoulos and Reece 2003b, Rolls, Aggelopoulos, Franco and Treves 2004) (see further Section C.3.5). This result is exactly what would be hoped for from a distributed representation. This result is not what would be expected for local encoding, for which the number of stimuli that could be encoded would increase linearly with the number of cells. (Even if the grandmother cells were noisy, adding more replicates to increase reliability..."
would not lead to more than a linear increase in the number of stimuli that can be encoded as a function of the number of cells.) Moreover, the encoding in the inferior temporal visual cortex about objects remains based on the spike count from each neuron, and not on the relative time of firing of each neuron or stimulus-dependent synchronization, when analysed with simultaneous single neuron recording (Rolls, Franco, Aggelopoulos and Reece 2003b, Rolls, Aggelopoulos, Franco and Treves 2004, Franco, Rolls, Aggelopoulos and Treves 2004) even in natural scenes when an attentional task is being performed (Aggelopoulos, Franco and Rolls 2005). Further, much of the information is available in short times of e.g. 20 or 50 ms (Tovee and Rolls 1995, Rolls, Franco, Aggelopoulos and Jerez 2006b), so that the receiving neuron does not need to integrate over a long time period to estimate a firing rate.

These findings provide very firm evidence that the encoding built at the end of the visual system is distributed, and that part of the power of this representation is that by receiving inputs from relatively small numbers of such neurons, neurons at the next stage of processing (for example in memory structures such as the hippocampus, amygdala, and orbitofrontal cortex) would obtain information about which of a very great number of objects or faces had been shown.

In this sense, the inferior temporal visual cortex provides a representation of objects and faces, in which information about which object or face is shown is made explicit in the firing of the neurons in such a way that the information can be read off very simply by memory systems such as the orbitofrontal cortex, amygdala, and perirhinal cortex / hippocampal systems. The information can be read off using dot product decoding, that is by using a synaptically weighted sum of inputs from inferior temporal cortex neurons (see further Appendix C and Section 2.2.6). Examples of some of the types of objects and faces that are encoded in this way by anterior inferior temporal cortex neurons are shown in Figs. 2.11 and C.7.

This representational capacity of neuronal populations has fundamental implications for the connectivity of the brain, for it shows that neurons need not have hundreds of thousands or millions of inputs to have available to them information about what is represented in another population of cells, but that instead the real numbers of perhaps 8,000–10,000 synapses per neuron would be adequate for them to receive considerable information from the several different sources between which this set of synapses is allocated.

It may be noted that it is unlikely that there are further processing areas beyond those
described where ensemble coding changes into grandmother cell encoding. Anatomically, there does not appear to be a whole further set of visual processing areas present in the brain; and outputs from the temporal lobe visual areas such as those described are taken directly to limbic and related regions such as the amygdala and orbitofrontal cortex, and via the perirhinal and entorhinal cortex to the hippocampus (see Chapter 2, Rolls (2000a), Rolls and Deco (2002), and Rolls and Stringer (2005)). Indeed, tracing this pathway onwards, we have found a population of neurons with face-selective responses in the amygdala, and in the majority of these neurons, different responses occur to different faces, with ensemble (not local) coding still being present (Leonard, Rolls, Wilson and Baylis 1985). The amygdala, in turn, projects to another structure that may be important in other behavioural responses to faces, the ventral striatum, and comparable neurons have also been found in the ventral striatum (Williams, Rolls, Leonard and Stern 1993). We have also recorded from face-responding neurons in the part of the orbitofrontal cortex that receives from the IT/STS cortex, and have found that the encoding there is also not local but is distributed (Rolls, Critchley, Browning and Inoue 2006a).

4.2.9 Face expression, gesture, and view represented in a population of neurons in the cortex in the superior temporal sulcus

In addition to the population of neurons that code for face identity, which tend to have object-based representations and are in areas TEa and TEM on the ventral bank of the superior temporal sulcus, there is a separate population in the cortex in the superior temporal sulcus (e.g. area TPO) that conveys information about facial expression (Hasselmo, Rolls and Baylis 1989a) (see e.g. Fig. 4.17). Some of the neurons in this region tend to have view-based representations (so that information is conveyed for example about whether the face is looking at one, or is looking away), and might respond to moving faces, and to facial gesture (Hasselmo, Rolls, Baylis and Nalwa 1989b).

Thus information in cortical areas that project to the amygdala and orbitofrontal cortex is about face identity, and about face expression and gesture. Both types of information are important in social and emotional responses to other primates (including humans), which must be based on who the individual is as well as on the face expression or gesture being made.

One output from the amygdala for this information is probably via the ventral striatum, for a small population of neurons has been found in the ventral striatum with responses selective for faces (Rolls and Williams 1987a, Williams, Rolls, Leonard and Stern 1993).

4.2.10 Specialized regions in the temporal cortical visual areas

As we have just seen, some neurons respond to face identity, and others to face expression (Hasselmo, Rolls and Baylis 1989a). The neurons responsive to expression were found primarily in the cortex in the superior temporal sulcus, while the neurons responsive to identity were found in the inferior temporal gyrus.

A further way in which some of these neurons in the cortex in the superior temporal sulcus may be involved in social interactions is that some of them respond to gestures, for example to a face undergoing ventral flexion (Perrett, Smith, Potter, Mistlin, Head, Milner and Jeeves 1985b, Hasselmo, Rolls, Baylis and Nalwa 1989b). The interpretation of these neurons as being useful for social interactions is that in some cases these neurons respond not only to ventral head flexion, but also to the eyes lowering and the eyelids closing (Hasselmo, Rolls, Baylis and Nalwa 1989b). These movements (turning the head away, breaking eye contact, and eyelid lowering) often occur together when a monkey is breaking social contact with another, and neurons that respond to these components could be built by associative synaptic
There is a population of neurons in the cortex in the superior temporal sulcus with responses tuned to respond differently to different face expressions. The cells in the two left panels did not discriminate between individuals (faces MM, FF, and MM), but did discriminate between different expressions on the faces of those individuals (C, calm expression; t, mild threat; T, strong threat). In contrast, the cells in the right two panels responded differently to different individuals, and did not discriminate between different expressions. The neurons that discriminated between expressions were found mainly in the cortex in the fundus of the superior temporal sulcus; the neurons that discriminated between identity were in contrast found mainly in the cortex in the lateral part of the ventral lip of the superior temporal sulcus (areas TEa and TEm). (From Hasselmo, Rolls and Baylis 1989a.)

It is also important when decoding facial expression to retain some information about the direction of the head relative to the observer, for this is very important in determining whether a threat is being made in your direction. The presence of view–dependent, head and body gesture (Hasselmo, Rolls, Baylis and Nalwa 1989b), and eye gaze (Perrett, Smith, Potter, Mistlin, Head, Milner and Jeeves 1985b), representations in some of these cortical regions where face expression is represented is consistent with this requirement. In contrast, the TE areas (more ventral, mainly in the macaque inferior temporal gyrus), in which neurons tuned to face identity (Hasselmo, Rolls and Baylis 1989a) and with view–independent responses (Hasselmo, Rolls, Baylis and Nalwa 1989b) are more likely to be found, may be more related to an object–based representation of identity. Of course, for appropriate social and emotional responses, both types of subsystem would be important, for it is necessary to know both the direction of a social gesture, and the identity of the individual, in order to make the correct social or emotional response.

Further evidence for specialization of function in the different architectonically defined
areas of the temporal cortex (Seltzer and Pandya 1978) (see Fig. 4.3) was found by Baylis, Rolls and Leonard (1987). Areas TPO, PGa and IPa are multimodal, with neurons that respond to visual, auditory and/or somatosensory inputs. The more ventral areas in the inferior temporal gyrus (areas TE3, TE2, TE1, TEa and TEM) are primarily unimodal visual areas. Areas in the cortex in the anterior and dorsal part of the superior temporal sulcus (e.g. TPO, IPa and IPg) have neurons specialized for the analysis of moving visual stimuli. Neurons responsive primarily to faces are found more frequently in areas TPO, TEa and TEM, where they comprise approximately 20% of the visual neurons responsive to stationary stimuli, in contrast to the other temporal cortical areas in which they comprise 4–10%. The stimuli that activate other cells in these TE regions include simple visual patterns such as gratings, and combinations of simple stimulus features (Gross, Desimone, Albright and Schwartz 1985, Baylis, Rolls and Leonard 1987, Tanaka, Saito, Fukada and Moriya 1990). If patches are identified by fMRI, the proportion of neurons tuned to for example faces may be high (Tsao et al. 2006). Due to the fact that face-selective neurons have a wide distribution (Baylis, Rolls and Leonard 1987), and also occur in different patches (Tsao et al. 2006), it might be expected that only large lesions, or lesions that interrupt outputs of these visual areas, would produce readily apparent face-processing deficits.

Another specialization is that areas TEa and TEM, which receive inter alia from the cortex in the intraparietal sulcus, have neurons that are tuned to binocular disparity, so that information derived from stereopsis about the 3D structure of objects is represented in the inferior temporal cortical visual areas (Janssen, Vogels and Orban 1999, Janssen, Vogels and Orban 2000). Interestingly, these neurons respond only when the black and white regions in the two eyes are correlated (that is, when a white patch in one eye corresponds to a white patch in the other eye) (Janssen, Vogels, Liu and Orban 2003). This corresponds to what we see perceptually and consciously. In contrast, in V1, MT and MST, depth-tuned neurons can respond to anticorrelated depth images, that is where white in one eye corresponds to black in the other (Parker, Cumming and Dodd 2000, DeAngelis, Cumming and Newsome 2000, Parker 2007), and this may be suitable for eye movement control, but much less for shape and object discrimination and identification. It may be expected that other depth cues, such as perspective, surface shading, and occlusion, affect the response properties of some neurons in inferior temporal cortex visual areas. Binocular disparity, and information from these other depth cues, may be used to compute the absolute size of objects, which is represented independently of the distance of the object for a small proportion of inferior temporal cortex neurons (Rolls and Baylis 1986). Knowing the absolute size of an object is useful evidence to include in the identification of an object. Although cues from binocular disparity can thus drive some temporal cortex visual neurons, and there is a small proportion of inferior temporal neurons that respond better to real faces and objects than to 2D representations on a monitor, it is found that the majority of TE neurons respond as well or almost as well to 2D images on a video monitor as to real faces or objects (Perrett, Rolls and Caan 1982). Moreover, the tuning of inferior temporal cortex neurons to images of faces or objects on a video monitor is similar to that to real objects (personal observations of E. T. Rolls).

Neuroimaging data, while not being able to address the details of what is encoded in a brain area or of how it is encoded, does provide evidence consistent with the neurophysiology that there are different face processing systems in the human brain, and different processing subsystems for objects, moving objects, and scenes (Spiridon, Fischl and Kanwisher 2006, Grill-Spector and Malach 2004, Epstein and Kanwisher 1998). For example, Kanwisher, McDermott and Chun (1997) and Ishai, Ungerleider, Martin, Schouten and Haxby (1999) have shown activation by faces of an area in the fusiform gyrus; Hoffman and Haxby (2000) have shown that distinct areas are activated by eye gaze and face identity; Dolan, Fink,
Rolls, Booth, Holmes, Frackowiak and Friston (1997) have shown that a fusiform gyrus area becomes activated after humans learn to identify faces in complex scenes; and the amygdala (Morris, Fritch, Perrett, Rowland, Young, Calder and Dolan 1996a) and orbitofrontal cortex (Blair, Morris, Frith, Perrett and Dolan 1999) may become activated particularly by certain face expressions. Consistent with the neurophysiology described below and by Baylis, Rolls and Leonard (1987), human fMRI studies are now able to detect small patches of non-face and other types of responsiveness in areas that appeared to respond primarily to one category such as faces (Grill-Spector, Sayres and Ress 2006, Haxby 2006).

Different inferior temporal cortex neurons in macaques not only provide different types of information about different aspects of faces, as just described, but also respond differently to different categories of visual stimuli, with for example some neurons conveying information primarily about faces, and others about objects (Rolls and Tovee 1995b, Rolls, Treves, Tovee and Panzeri 1997d, Booth and Rolls 1998) (see Appendix C). In fact, when recording in the inferior temporal cortex, one finds small clustered groups of neurons, with neurons within a cluster responding to somewhat similar attributes of stimuli, and different clusters responding to different categories or types of visual stimuli. For example, within the set of neurons responding to moving faces, there are some neuronal clusters that respond to for example ventral and dorsal flexion of the head on the body; others that respond to axial rotation of the head; others that respond to the sight of mouth movements; and others that respond to the gaze direction in a face being viewed (Hasselmo, Rolls, Baylis and Nalwa 1989b, Perrett, Smith, Potter, Mistlin, Head, Milner and Jeeves 1985b). Within the domain of objects, some neuronal clusters respond based on the shapes of the object (e.g. responding to elongated and not round objects), others respond based on the texture of the object being viewed, and others are sensitive to what the object is doing (Hasselmo, Rolls, Baylis and Nalwa 1989b, Perrett, Smith, Mistlin, Chitty, Head, Potter, Broennimann, Milner and Jeeves 1985a). Within a cluster, each neuron is tuned differently, and may combine at least partly with information predominant in other clusters, so that at least some of the neurons in a cluster can become quite selective between different objects, with the preference for individual objects showing the usual exponential firing rate distribution (see Figs. C.4 and C.6). Using optical imaging, Tanaka and colleagues (Tanaka 1996) have seen comparable evidence for different localized clusters of neuronal activation produced by different types or categories of stimuli.

The principle that Rolls proposes underlies this local clustering is the representation of the high dimensional space of objects and their features into a two-dimensional cortical sheet using the local self-organizing mapping principles described in Section B.4.6. In this situation, these principles produce maps that have many fractures, but nevertheless include local clusters of similar neurons. Exactly this clustering can be produced in the model of visual object recognition, VisNet, described in Section 4.5 by replacing the lateral inhibition filter shown in Fig. 4.21 with a difference of Gaussians filter of the type illustrated in Fig. B.20. Such a difference of Gaussians filter would reflect the effects of short range excitatory (recurrent collateral) connections between cortical pyramidal cells, and longer range inhibition produced through inhibitory interneurons. These local clusters minimize the wiring length between neurons, if the computations involve exchange of information within neurons of a similar general type, as will usually be the case (see Section B.4.6). Minimizing wiring length is crucial in keeping brain size relatively low (see Section B.4.6). Another useful property of such self-organizing maps is that they encourage distributed representations in which semi-continuous similarity functions are mapped. This is potentially useful in building representations that then generalize usefully when new stimuli are shown between representations that are already set up continuously in the map.

Another fundamental contribution and key aspect of neocortical design facilitated by the short-range recurrent excitatory connections is that the attractors that are formed are local.
Approaches to invariant object recognition

This is fundamentally important, for if the neocortex had long-range connectivity, it would tend to be able to form only as many attractor states as there are connections per neocortical neuron, which would be a severe limit on cortical memory capacity (O’Kane and Treves 1992) (see Sections 1.10.4 and B.9).

It is consistent with this general conceptual background that Krieman, Koch and Fried (2000) have described some neurons in the human temporal lobe that seem to respond to categories of object. This is consistent with the principles just described, although in humans backprojections from language or other cognitive areas concerned for example with tool use might also influence the categories represented in high order cortical areas, as described in Sections B.4.5 and 1.11, and by Farah, Meyer and McMullen (1996) and Farah (2000).

4.3 Approaches to invariant object recognition

A goal of this book is to provide some of the foundations for understanding memory systems in the brain, and how they could function at the computational and neuronal network level. This in turn is fundamental to understanding very many aspects of brain function, for the brain performs many of its functions by what are memory-related operations that involve synaptic modification (see Appendix B). Some of the ways in which the visual system may produce the distributed invariant representations of objects needed for inputs to the emotion-learning systems and other memory systems described in this book are described at the computational level in the rest of this Chapter and by Rolls and Deco (2002) and Rolls and Stringer (2006), and include a hierarchical feed-forward series of competitive networks using convergence from stage to stage; and the use of a modified Hebb synaptic learning rule that incorporates a short-term memory trace of previous neuronal activity to help learn the invariant properties of objects from the temporo-spatial statistics produced by the normal viewing of objects (Rolls and Deco 2002, Wallis and Rolls 1997, Rolls and Milward 2000, Stringer and Rolls 2000, Rolls and Stringer 2001a, Elliffe, Rolls and Stringer 2002, Stringer and Rolls 2002, Deco and Rolls 2004, Rolls and Stringer 2006).

We start by emphasizing that generalization to different positions, sizes, views etc. of an object is not a simple property of one-layer neural networks. Although neural networks do generalize well, the type of generalization they show naturally is to vectors which have a high dot product or correlation with what they have already learned. To make this clear, Fig. 4.18 is a reminder that the activation $h_i$ of each neuron is computed as

$$h_i = \sum_j x_j w_{ij}$$

where the sum is over the $C$ input axons, indexed by $j$. Now consider translation (or shift) of the input pattern vector by one position. The dot product will now drop to a low level, and the neuron will not respond, even though it is the same pattern, just shifted by one location. This makes the point that special processes are needed to compute invariant representations.

Network approaches to such invariant pattern recognition are described in this chapter, and the nature of the problem to be solved is described further in Section B.4.7 and Fig. B.23 on page 587. Once an invariant representation has been computed by a sensory system, it is in a form that is suitable for presentation to a pattern association or autoassociation neural network (see Appendix B).

A number of different computational approaches that have been taken both in artificial vision systems and as suggestions for how the brain performs invariant object recognition are described in this section. This places in context the approach that appears to be taken in the brain and that forms the basis for VisNet described in Section 4.5.
Fig. 4.18 A neuron that computes a dot product of the input pattern with its synaptic weight vector generalizes well to other patterns based on their similarity measured in terms of dot product or correlation, but shows no translation (or size, etc.) invariance.

4.3.1 Feature spaces

One very simple possibility for performing object classification is based on feature spaces, which amount to lists of (the extent to which) different features are present in a particular object. The features might consist of textures, colours, areas, ratios of length to width, etc. The spatial arrangement of the features is not taken into account. If $n$ different properties are used to characterize an object, each viewed object is represented by a set of $n$ real numbers. It then becomes possible to represent an object by a point $R^n$ in an $n$-dimensional space (where $R$ is the resolution of the real numbers used). Such schemes have been investigated (Tou and Gonzalez 1974, Gibson 1950, Gibson 1979, Bolles and Cain 1982, Mundy and Zisserman 1992, Selfridge 1959, Mel 1997), but, because the relative positions of the different parts are not implemented in the object recognition scheme, are not sensitive to spatial jumbling of the features. For example, if the features consisted of nose, mouth, and eyes, such a system would respond to faces with jumbled arrangements of the eyes, nose, and mouth, which does not match human vision, nor the responses of macaque inferior temporal cortex neurons, which are sensitive to the spatial arrangement of the features in a face (Rolls, Tovee, Purcell, Stewart and Azzopardi 1994b). Similarly, such an object recognition system might not distinguish a normal car from a car with the back wheels removed and placed on the roof. Such systems do not therefore perform shape recognition (where shape implies something about the spatial arrangement of features within an object, see further Ullman (1996)), and something more is needed, and is implemented in the primate visual system. However, I note that the features that are present in objects, e.g. a furry texture, are useful to incorporate in object recognition systems, and the brain may well use, and the model VisNet in principle can use, evidence from which features are present in an object as part of the evidence for identification of a particular object. I note that the features might consist also of for example the pattern of movement that is characteristic of a particular object (such as a buzzing fly), and might use this as part of the input to final object identification.

The capacity to use shape in invariant object recognition is fundamental to primate vision, but may not be used or fully implemented on the visual systems of some other animals with less developed visual systems. For example, pigeons may correctly identify pictures containing people, a particular person, trees, pigeons etc, but may fail to distinguish a figure from a
Fig. 4.19 A 3D structural description of an object based on generalized cone parts. Each box corresponds to a 3D model, with its model axis on the left side of the box and the arrangement of its component axes on the right. In addition, some component axes have 3D models associated with them, as indicated by the way the boxes overlap. (After Marr and Nishihara 1978.)

scrambled version of a figure (Herrnstein 1984, Cerella 1986). Thus their object recognition may be based more on a collection of parts than on a direct comparison of complete figures in which the relative positions of the parts are important. Even if the details of the conclusions reached from this research are revised (Wasserman, Kirkpatrick-Steger and Biederman 1998), it nevertheless does appear that at least some birds may use computationally simpler methods than those needed for invariant shape recognition. For example, it may be that when some birds are trained to discriminate between images in a large set of pictures, they tend to rely on some chance detail of each picture (such as a spot appearing by mistake on the picture), rather than on recognition of the shapes of the object in the picture (Watanabe, Lea and Dittrich 1993).

4.3.2 Structural descriptions and syntactic pattern recognition

A second approach to object recognition is to decompose the object or image into parts, and to then produce a structural description of the relations between the parts. The underlying assumption is that it is easier to capture object invariances at a level where parts have been identified. This is the type of scheme for which Marr and Nishihara (1978) and Marr (1982) opted. The particular scheme (Binford 1981) they adopted consists of generalized cones, series of which can be linked together to form structural descriptions of some, especially animate, stimuli (see Fig. 4.19). Such schemes assume that there is a 3D internal model (structural description) of each object. Perception of the object consists of parsing or segmenting the scene into objects, and then into parts, then producing a structural description of the object, and then testing whether this structural description matches that of any known object stored in the system. Other examples of structural description schemes include those of Winston (1975), Sutherland (1968), and Milner (1974). The relations in the structural description may need to be quite complicated, for example ‘connected together’, ‘inside of’, ‘larger than’ etc.

Perhaps the most developed model of this type is the recognition by components (RBC) model of Biederman (1987), implemented in a computational model by Hummel and Biederman (1992). His small set (less than 50) of primitive parts named ‘geons’ include simple 3D
shapes such as boxes, cylinders and wedges. Objects are described by a syntactically linked list of the relations between each of the geons of which they are composed. Describing a table in this way (as a flat top supported by three or four legs) seems quite economical. Other schemes use 2D surface patches as their primitives (Dane and Bajcsy 1982, Brady, Ponce, Yuille and Asada 1985, Faugeras and Hebert 1986, Faugeras 1993). When 3D objects are being recognized, the implication is that the structural description is a 3D description. This is in contrast to feature hierarchical systems, in which recognition of a 3D object from any view might be accomplished by storing a set of associated 2D views (see below, Section 4.3.5).

There are a number of difficulties with schemes based on structural descriptions, some general, and some with particular reference to the potential difficulty of their implementation in the brain. First, it is not always easy to decompose the object into its separate parts, which must be performed before the structural description can be produced. For example, it may be difficult to produce a structural description of a cat curled up asleep from separately identifiable parts. Identification of each of the parts is also frequently very difficult when 3D objects are seen from different viewing angles, as key parts may be invisible or highly distorted. This is particularly likely to be difficult in 3D shape perception. It appears that being committed to producing a correct description of the parts before other processes can operate is making too strong a commitment early on in the recognition process.

A second difficulty is that many objects or animals that can be correctly recognized have rather similar structural descriptions. For example, the structural description of many four-legged animals is rather similar. Rather more than a structural description seems necessary to identify many objects and animals.

A third difficulty, which applies especially to biological systems, is the difficulty of implementing the syntax needed to hold the structural description as a 3D model of the object, of producing a syntactic structural description on the fly (in real time, and with potentially great flexibility of the possible arrangement of the parts), and of matching the syntactic description of the object in the image to all the stored representations in order to find a match. An example of a structural description for a limb might be body > thigh > shin > foot > toes. In this description > means ‘is linked to’, and this link must be between the correct pair of descriptors. If we had just a set of parts, without the syntactic or relational linking, then there would be no way of knowing whether the toes are attached to the foot or to the body. In fact, worse than this, there would be no evidence about what was related to what, just a set of parts. Such syntactical relations are difficult to implement in neuronal networks, because if the representations of all the features or parts just mentioned were active simultaneously, how would the spatial relations between the features also be encoded? (How would it be apparent just from the firing of neurons that the toes were linked to the rest of the foot but not to the body?) It would be extremely difficult to implement this ‘on the fly’ syntactic binding in a biologically plausible network (though cf. Hummel and Biederman (1992)), and the only suggested mechanism for flexible syntactic binding, temporal synchronization of the firing of different neurons, is not well supported as a quantitatively important mechanisms for information encoding in the ventral visual system, and would have major difficulties in implementing correct, relational, syntactic binding (see Sections 4.5.5.1 and C.3.7).

A fourth difficulty of the structural description approach is that segmentation into objects must occur effectively before object recognition, so that the linked structural description list can be of one object. Given the difficulty of segmenting objects in typical natural cluttered scenes (Ullman 1996), and the compounding problem of overlap of parts of objects by other objects, segmentation as a first necessary stage of object recognition adds another major difficulty for structural description approaches.

A fifth difficulty is that metric information, such as the relative size of the parts that are linked syntactically, needs to be specified in the structural description (Stankiewicz and
It is because of these difficulties that even in artificial vision systems implemented on computers, where almost unlimited syntactic binding can easily be implemented, the structural description approach to object recognition has not yet succeeded in producing a scheme which actually works in more than an environment in which the types of objects are limited, and the world is far from the natural world, consisting for example of 2D scenes (Mundy and Zisserman 1992).

Although object recognition in the brain is unlikely to be based on the structural description approach, for the reasons given above, and the fact that the evidence described in this chapter supports a feature hierarchy rather than the structural description implementation in the brain, it is certainly the case that humans can provide verbal, syntactic, descriptions of objects in terms of the relations of their parts, and that this is often a useful type of description. Humans may therefore it is suggested supplement a feature hierarchical object recognition system built into their ventral visual system with the additional ability to use the type of syntax that is necessary for language to provide another level of description of objects. This is of course useful in, for example, engineering applications.

### 4.3.3 Template matching and the alignment approach

Another approach is template matching, comparing the image on the retina with a stored image or picture of an object. This is conceptually simple, but there are in practice major problems. One major problem is how to align the image on the retina with the stored images, so that all possible images on the retina can be compared with the stored template or templates of each object.

The basic idea of the alignment approach (Ullman 1996) is to compensate for the transformations separating the viewed object and the corresponding stored model, and then compare them. For example, the image and the stored model may be similar, except for a difference in size. Scaling one of them will remove this discrepancy and improve the match between them. For a 2D world, the possible transforms are translation (shift), scaling, and rotation. Given for example an input letter of the alphabet to recognize, the system might, after segmentation (itself a very difficult process if performed independently of (prior to) object recognition), compensate for translation by computing the centre of mass of the object, and shifting the character to a ‘canonical location’. Scale might be compensated for by calculating the convex hull (the smallest envelope surrounding the object), and then scaling the image. Of course how the shift and scaling would be accomplished is itself a difficult point – easy to perform on a computer using matrix multiplication as in simple computer graphics, but not the sort of computation that could be performed easily or accurately by any biologically plausible network. Compensating for rotation is even more difficult (Ullman 1996). All this has to happen before the segmented canonical representation of the object is compared to the stored object templates with the same canonical representation. The system of course becomes vastly more complicated when the recognition must be performed of 3D objects seen in a 3D world, for now the particular view of an object after segmentation must be placed into a canonical form, regardless of which view, or how much of any view, may be seen in a natural scene with occluding contours. However, this process is helped, at least in computers that can perform high precision matrix multiplication, by the fact that (for many continuous transforms such as 3D rotation, translation, and scaling) all the possible views of an object transforming in 3D space can be expressed as the linear combination of other views of the same object (see Chapter 5 of Ullman (1996); Koenderink and van Doorn (1991); and Koenderink (1990)).

This alignment approach is the main theme of the book by Ullman (1996), and there are a number of computer implementations (Lowe 1985, Grimson 1990, Huttenlocher and
Ullman 1990, Shashua 1995). However, as noted above, it seems unlikely that the brain is able to perform the high precision calculations needed to perform the transforms required to align any view of a 3D object with some canonical template representation. For this reason, and because the approach also relies on segmentation of the object in the scene before the template alignment algorithms can start, and because key features may need to be correctly identified to be used in the alignment (Edelman 1999), this approach is not considered further here.

We may note here in passing that some animals with a less computationally developed visual system appear to attempt to solve the alignment problem by actively moving their heads or eyes to see what template fits, rather than starting with an image on the eye and attempting to transform it into canonical coordinates. This ‘active vision’ approach used for example by some invertebrates has been described by Land (1999) and Land and Collett (1997).

4.3.4 Invertible networks that can reconstruct their inputs

Hinton, Dayan, Frey and Neal (1995) and Hinton and Ghahramani (1997) have argued that cortical computation is invertible, so that for example the forward transform of visual information from V1 to higher areas loses no information, and there can be a backward transform from the higher areas to V1. A comparison of the reconstructed representation in V1 with the actual image from the world might in principle be used to correct all the synaptic weights between the two (in both the forward and the reverse directions), in such a way that there are no errors in the transform. This suggested reconstruction scheme would seem to involve non-local synaptic weight correction (though see Section B.14 for a suggested, although still biologically implausible, neural implementation), or other biologically implausible operations. The scheme also does not seem to provide an account for why or how the responses of inferior temporal cortex neurons become the way they are (providing information about which object is seen relatively independently of position on the retina, size, or view). The whole forward transform performed in the brain seems to lose much of the information about the size, position and view of the object, as it is evidence about which object is present invariant of its size, view, etc. that is useful to the stages of processing about objects that follow (as described in Section 3.5 but see Section 4.5.10). Because of these difficulties, and because the backprojections are needed for processes such as recall (see Sections 1.11 and 2.3.7), this approach is not considered further in this book.

In the context of recall, if the visual system were to perform a reconstruction in V1 of a visual scene from what is represented in the inferior temporal visual cortex, then it might be supposed that remembered visual scenes might be as information-rich (and subjectively as full of rich detail) as seeing the real thing. This is not the case for most humans, and indeed this point suggests that at least what reaches consciousness from the inferior temporal visual cortex (which is activated during the recall of visual memories) is the identity of the object (as made explicit in the firing rate of the neurons), and not the low-level details of the exact place, size, and view of the object in the recalled scene, even though, according to the reconstruction argument, that information should be present in the inferior temporal visual cortex.

4.3.5 Feature hierarchies and 2D view-based object recognition

Another approach, and one that is much closer to what appears to be present in the primate ventral visual system (Rolls and Deco 2002, Rolls 2007i, Rolls 2008a, Wurtz and Kandel 2000a), is a feature hierarchy system (see Fig. 4.20).

In this approach, the system starts with some low-level description of the visual scene, in terms for example of oriented straight line segments of the type that are represented in
Fig. 4.20 The feature hierarchy approach to object recognition. The inputs may be neurons tuned to oriented straight line segments. In early intermediate layers neurons respond to a combination of these inputs in the correct spatial position with respect to each other. In further intermediate layers, of which there may be several, neurons respond with some invariance to the feature combinations represented early, and form higher order feature combinations. Finally, in the top layer, neurons respond to combinations of what is represented in the preceding intermediate layer, and thus provide evidence about objects in a position (and scale and even view) invariant way. Convergence through the network is designed to provide top layer neurons with information from across the entire input retina, as part of the solution to translation invariance, and other types of invariance are treated similarly.

A number of problems need to be solved for such feature hierarchy visual systems to provide a useful model of object recognition in the primate visual system.

First, some way needs to be found to keep the number of feature combination neurons realistic at each stage, without undergoing a combinatorial explosion. If a separate feature
combination neuron was needed to code for every possible combination of \( n \) types of feature each with a resolution of 2 levels (binary encoding) in the preceding stage, then \( 2^n \) neurons would be needed. The suggestion that is made in Section 4.4 is that by forming neurons that respond to low-order combinations of features (neurons that respond to just say 2–4 features from the preceding stage), the number of actual feature analysing neurons can be kept within reasonable numbers. By reasonable we mean the number of neurons actually found at any one stage of the visual system, which, for V4 might be in the order of \( 60 \times 10^6 \) neurons (assuming a volume for macaque V4 of approximately 2,000 \( \text{mm}^3 \), and a cell density of 20,000–40,000 neurons per \( \text{mm}^3 \), see Table 1.1). This is certainly a large number; but the fact that a large number of neurons is present at each stage of the primate visual system is in fact consistent with the hypothesis that feature combination neurons are part of the way in which the brain solves object recognition. A factor which also helps to keep the number of neurons under control is the statistics of the visual world, which contain great redundancies. The world is not random, and indeed the statistics of natural images are such that many regularities are present (Field 1994), and not every possible combination of pixels on the retina needs to be separately encoded. A third factor which helps to keep the number of connections required onto each neuron under control is that in a multilayer hierarchy each neuron can be set up to receive connections from only a small region of the preceding layer. Thus an individual neuron does not need to have connections from all the neurons in the preceding layer. Over multiple layers, the required convergence can be produced so that the same neurons in the top layer can be activated by an image of an effective object anywhere on the retina (see Fig. 4.2).

A second problem of feature hierarchy approaches is how to map all the different possible images of an individual object through to the same set of neurons in the top layer by modifying the synaptic connections (see Fig. 4.2). The solution discussed in Sections 4.4, 4.5.1.1 and 4.5.4 is the use of a synaptic modification rule with a short-term memory trace of the previous activity of the neuron, to enable it to learn to respond to the now transformed version of what was seen very recently, which, given the statistics of looking at the visual world, will probably be an input from the same object.

A third problem of feature hierarchy approaches is how they can learn in just a few seconds of inspection of an object to recognize it in different transforms, for example in different positions on the retina in which it may never have been presented during training. A solution to this problem is provided in Section 4.5.5, in which it is shown that this can be a natural property of feature hierarchy object recognition systems, if they are trained first for all locations on the intermediate level feature combinations of which new objects will simply be a new combination, and therefore requiring learning only in the upper layers of the hierarchy.

A fourth potential problem of feature hierarchy systems is that when solving translation invariance they need to respond to the same local spatial arrangement of features (which are needed to specify the object), but to ignore the global position of the whole object. It is shown in Section 4.5.5 that feature hierarchy systems can solve this problem by forming feature combination neurons at an early stage of processing (e.g. V1 or V2 in the brain) that respond with high spatial precision to the local arrangement of features. Such neurons would respond differently for example to L, +, and T if they receive inputs from two line-responding neurons. It is shown in Section 4.5.5 that at later layers of the hierarchy, where some of the intermediate level feature combination neurons are starting to show translation invariance, then correct object recognition may still occur because only one object contains just those sets of intermediate level neurons in which the spatial representation of the features is inherent in the encoding.

The type of representation developed in a hierarchical object recognition system, in the brain, and by VisNet as described in the rest of this chapter would be suitable for recognition of an object, and for linking associative memories to objects, but would be less good for
making actions in 3D space to particular parts of, or inside, objects, as the 3D coordinates of each part of the object would not be explicitly available. It is therefore proposed that visual fixation is used to locate in foveal vision part of an object to which movements must be made, and that local disparity and other measurements of depth (made explicit in the dorsal visual system) then provide sufficient information for the motor system to make actions relative to the small part of space in which a local, view-dependent, representation of depth would be provided (cf. Ballard (1990)).

One advantage of feature hierarchy systems is that they can operate fast (see Sections B.4 and B.6.6).

A second advantage is that the feature analyzers can be built out of the rather simple competitive networks described in Section B.4 which use a local learning rule, and have no external teacher, so that they are rather biologically plausible. Another advantage is that, once trained on subset features common to most objects, the system can then learn new objects quickly.

A related third advantage is that, if implemented with competitive nets as in the case of VisNet (see Section 4.5), then neurons are allocated by self-organization to represent just the features present in the natural statistics of real images (cf. Field (1994)), and not every possible feature that could be constructed by random combinations of pixels on the retina.

A related fourth advantage of feature hierarchy networks is that because they can utilize competitive networks, they can still produce the best guess at what is in the image under non-ideal conditions, when only parts of objects are visible because for example of occlusion by other objects, etc. The reasons for this are that competitive networks assess the evidence for the presence of certain ‘features’ to which they are tuned using a dot product operation on their inputs, so that they are inherently tolerant of missing input evidence; and reach a state that reflects the best hypothesis or hypotheses (with soft competition) given the whole set of inputs, because there are competitive interactions between the different neurons (see Section B.4).

A fifth advantage of a feature hierarchy system is that, as shown in Section 4.5.6, the system does not need to perform segmentation into objects as part of pre-processing, nor does it need to be able to identify parts of an object, and can also operate in cluttered scenes in which the object may be partially obscured. The reason for this is that once trained on objects, the system then operates somewhat like an associative memory, mapping the image properties forward onto whatever it has learned about before, and then by competition selecting just the most likely output to be activated. Indeed, the feature hierarchy approach provides a mechanism by which processing at the object recognition level could feed back using backprojections to early cortical areas to provide top-down guidance to assist segmentation. Although backprojections are not built into VisNet2 (Rolls and Milward 2000), they have been added when attentional top-down processing must be incorporated (Deco and Rolls 2004), are present in the brain, and are incorporated into the models described in Chapter 6.

A sixth advantage of feature hierarchy systems is that they can naturally utilize features in the images of objects which are not strictly part of a shape description scheme, such as the fact that different objects have different textures, colours etc. Feature hierarchy systems, because they utilize whatever is represented at earlier stages in forming feature combination neurons at the next stage, naturally incorporate such ‘feature list’ evidence into their analysis, and have the advantages of that approach (see Section 4.3.1 and also Mel (1997)). Indeed, the feature space approach can utilize a hybrid representation, some of whose dimensions may be discrete and defined in structural terms, while other dimensions may be continuous and defined in terms of metric details, and others may be concerned with non-shape properties such as texture and colour (cf. Edelman (1999)).

A seventh advantage of feature hierarchy systems is that they do not need to utilize ‘on
the fly’ or run-time arbitrary binding of features. Instead, the spatial syntax is effectively hardwired into the system when it is trained, in that the feature combination neurons have learned to respond to their set of features when they are in a given spatial arrangement on the retina.

An eighth advantage of feature hierarchy systems is that they can self-organize (given the right functional architecture, trace synaptic learning rule, and the temporal statistics of the normal visual input from the world), with no need for an external teacher to specify that the neurons must learn to respond to objects. The correct, object, representation self-organizes itself given rather economically specified genetic rules for building the network (cf. Rolls and Stringer (2000)).

Ninth, it is also noted that hierarchical visual systems may recognize 3D objects based on a limited set of 2D views of objects, and that the same architectural rules just stated and implemented in VisNet will correctly associate together the different views of an object. It is part of the concept (see below), and consistent with neurophysiological data (Tanaka 1996), that the neurons in the upper layers will generalize correctly within a view (see Section 4.5.7).

After the immediately following description of early models of a feature hierarchy approach implemented in the Cognitron and Neocognitron, we turn for the remainder of this chapter to analyses of how a feature hierarchy approach to invariant visual object recognition might be implemented in the brain, and how key computational issues could be solved by such a system. The analyses are developed and tested with a model, VisNet, which will shortly be described. Much of the data we have on the operation of the high order visual cortical areas (see Section 4.2, Rolls and Deco (2002) and Rolls (2007i)) suggest that they implement a feature hierarchy approach to visual object recognition, as is made evident in the remainder of this chapter.

4.3.5.1 The Cognitron and Neocognitron

An early computational model of a hierarchical feature-based approach to object recognition, joining other early discussions of this approach (Selfridge 1959, Sutherland 1968, Barlow 1972, Milner 1974), was proposed by Fukushima (1975, 1980, 1989, 1991). His model used two types of cell within each layer to approach the problem of invariant representations. In each layer, a set of ‘simple cells’, with defined position, orientation, etc. sensitivity for the stimuli to which they responded, was followed by a set of ‘complex cells’, which generalized a little over position, orientation, etc. This simple cell – complex cell pairing within each layer provided some invariance. When a neuron in the network using competitive learning with its stimulus set, which was typically letters on a 16 × 16 pixel array, learned that a particular feature combination had occurred, that type of feature analyzer was replicated in a non-local manner throughout the layer, to provide further translation invariance. Invariant representations were thus learned in a different way from VisNet. Up to eight layers were used. The network could learn to differentiate letters, even with some translation, scaling, or distortion. Although internally it is organized and learns very differently to VisNet, it is an independent example of the fact that useful invariant pattern recognition can be performed by multilayer hierarchical networks. A major biological implausibility of the system is that once one neuron within a layer learned, other similar neurons were set up throughout the layer by a non-local process. A second biological limitation was that no learning rule or self-organizing process was specified as to how the complex cells can provide translation invariant representations of simple cell responses – this was simply handwired. Solutions to both these issues are provided by VisNet.
4.4 Hypotheses about the computational mechanisms in the visual cortex for object recognition

The neurophysiological findings described in Section 4.2, and wider considerations on the possible computational properties of the cerebral cortex (Rolls 1989b, Rolls 1989f, Rolls 1992a, Rolls 1994a, Rolls 1995b, Rolls 1997c, Rolls 2000a, Rolls and Treves 1998, Rolls and Deco 2002), lead to the following outline working hypotheses on object recognition by visual cortical mechanisms (see Rolls (1992a)). The principles underlying the processing of faces and other objects may be similar, but more neurons may become allocated to represent different aspects of faces because of the need to recognize the faces of many different individuals, that is to identify many individuals within the category faces.

Cortical visual processing for object recognition is considered to be organized as a set of hierarchically connected cortical regions consisting at least of V1, V2, V4, posterior inferior temporal cortex (TEO), inferior temporal cortex (e.g. TE3, TEa and TEm), and anterior temporal cortical areas (e.g. TE2 and TE1). (This stream of processing has many connections with a set of cortical areas in the anterior part of the superior temporal sulcus, including area TPO.) There is convergence from each small part of a region to the succeeding region (or layer in the hierarchy) in such a way that the receptive field sizes of neurons (e.g. 1 degree near the fovea in V1) become larger by a factor of approximately 2.5 with each succeeding stage (and the typical parafoveal receptive field sizes found would not be inconsistent with the calculated approximations of e.g. 8 degrees in V4, 20 degrees in TEO, and 50 degrees in the inferior temporal cortex (Boussaoud, Desimone and Ungerleider 1991)) (see Fig. 4.2 on page 264). Such zones of convergence would overlap continuously with each other (see Fig. 4.2). This connectivity would be part of the architecture by which translation invariant representations are computed.

Each layer is considered to act partly as a set of local self-organizing competitive neuronal networks with overlapping inputs. (The region within which competition would be implemented would depend on the spatial properties of inhibitory interneurons, and might operate over distances of 1–2 mm in the cortex.) These competitive nets operate by a single set of forward inputs leading to (typically non-linear, e.g. sigmoid) activation of output neurons; of competition between the output neurons mediated by a set of feedback inhibitory interneurons which receive from many of the principal (in the cortex, pyramidal) cells in the net and project back (via inhibitory interneurons) to many of the principal cells and serve to decrease the firing rates of the less active neurons relative to the rates of the more active neurons; and then of synaptic modification by a modified Hebb rule, such that synapses to strongly activated output neurons from active input axons strengthen, and from inactive input axons weaken (see Section B.4). A biologically plausible form of this learning rule that operates well in such networks is

$$\delta w_{ij} = \alpha y_i (x_j - w_{ij})$$  \hspace{1cm} (4.2)

where \(\alpha\) is a learning rate constant, and \(x_j\) and \(w_{ij}\) are in appropriate units (see Section B.4). Such competitive networks operate to detect correlations between the activity of the input neurons, and to allocate output neurons to respond to each cluster of such correlated inputs. These networks thus act as categorizers. In relation to visual information processing, they would remove redundancy from the input representation, and would develop low entropy representations of the information (cf. Barlow (1985), Barlow, Kaushal and Mitchison (1989)). Such competitive nets are biologically plausible, in that they utilize Hebb-modifiable forward excitatory connections, with competitive inhibition mediated by cortical inhibitory neurons. The competitive scheme I suggest would not result in the formation of ‘winner-take-all’ or ‘grandmother’ cells, but would instead result in a small ensemble of active neurons.
representing each input (Rolls 1989b, Rolls 1989f, Rolls and Treves 1998) (see Section B.4). The scheme has the advantages that the output neurons learn better to distribute themselves between the input patterns (cf. Bennett (1990)), and that the sparse representations formed have utility in maximizing the number of memories that can be stored when, towards the end of the visual system, the visual representation of objects is interfaced to associative memory (Rolls 1989b, Rolls 1989f, Rolls and Treves 1998)\textsuperscript{12}.

Translation invariance would be computed in such a system by utilizing competitive learning to detect regularities in inputs when real objects are translated in the physical world. The hypothesis is that because objects have continuous properties in space and time in the world, an object at one place on the retina might activate feature analyzers at the next stage of cortical processing, and when the object was translated to a nearby position, because this would occur in a short period (e.g. 0.5 s), the membrane of the postsynaptic neuron would still be in its ‘Hebb-modifiable’ state (caused for example by calcium entry as a result of the voltage-dependent activation of NMDA receptors), and the presynaptic afferents activated with the object in its new position would thus become strengthened on the still-activated postsynaptic neuron. It is suggested that the short temporal window (e.g. 0.5 s) of Hebb-modifiability helps neurons to learn the statistics of objects moving in the physical world, and at the same time to form different representations of different feature combinations or objects, as these are physically discontinuous and present less regular correlations to the visual system. Földiák (1991) has proposed computing an average activation of the postsynaptic neuron to assist with the same problem. One idea here is that the temporal properties of the biologically implemented learning mechanism are such that it is well suited to detecting the relevant continuities in the world of real objects. Another suggestion is that a memory trace for what has been seen in the last 300 ms appears to be implemented by a mechanism as simple as continued firing of inferior temporal neurons after the stimulus has disappeared, as has been found in masking experiments (Rolls and Tovee 1994, Rolls, Tovee, Purcell, Stewart and Azzopardi 1994b, Rolls, Tovee and Panzeri 1999b, Rolls 2003).

I also suggest that other invariances, for example size, spatial frequency, and rotation invariance, could be learned by a comparable process. (Early processing in V1 which enables different neurons to represent inputs at different spatial scales would allow combinations of the outputs of such neurons to be formed at later stages. Scale invariance would then result from detecting at a later stage which neurons are almost conjunctively active as the size of an object alters.) It is suggested that this process takes place at each stage of the multiple-layer cortical processing hierarchy, so that invariances are learned first over small regions of space, and then over successively larger regions. This limits the size of the connection space within which correlations must be sought.

Increasing complexity of representations could also be built in such a multiple layer hierarchy by similar mechanisms. At each stage or layer the self-organizing competitive nets would result in combinations of inputs becoming the effective stimuli for neurons. In order to avoid the combinatorial explosion, it is proposed, following Feldman (1985), that low-order

\textsuperscript{12}In that each neuron has graded responses centred about an optimal input, the proposal has some of the advantages with respect to hypersurface reconstruction described by Poggio and Girosi (1990a). However, the system I propose learns differently, in that instead of using perhaps non-biologically plausible algorithms to optimally locate the centres of the receptive fields of the neurons, the neurons use graded competition to spread themselves throughout the input space, depending on the statistics of the inputs received, and perhaps with some guidance from backprojections (see below). In addition, the competitive nets I propose use as a distance function the dot product between the input vector to a neuron and its synaptic weight vector, whereas radial basis function networks use a Gaussian measure of distance (see Section B.4.8). Both systems benefit from the finite width of the response region of each neuron which tapers from a maximum, and is important for enabling the system to generalize smoothly from the examples with which it has learned (cf. Poggio and Girosi (1990b), Poggio and Girosi (1990a)), to help the system to respond for example with the correct invariances as described below.
combinations of inputs would be what is learned by each neuron. (Each input would not be represented by activity in a single input axon, but instead by activity in a set of active input axons.) Evidence consistent with this suggestion that neurons are responding to combinations of a few variables represented at the preceding stage of cortical processing is that some neurons in V1 respond to combinations of bars or edges (see Section 2.5 of Rolls and Deco (2002); Sillito, Grieve, Jones, Cudeiro and Davis (1995); and Shevelev, Novikova, Lazareva, Tikhomirov and Sharaev (1995)); V2 and V4 respond to end-stopped lines, to angles formed by a combination of lines, to tongues flanked by inhibitory subregions, or to combinations of colours (see Hegde and Van Essen (2000), Ito and Komatsu (2004) and Chapter 3 of Rolls and Deco (2002)); in posterior inferior temporal cortex to stimuli which may require two or more simple features to be present (Tanaka, Saito, Fukada and Moriya 1990); and in the temporal cortical face processing areas to images that require the presence of several features in a face (such as eyes, hair, and mouth) in order to respond (Perrett, Rolls and Caan 1982, Yamane, Kaji and Kawano 1988) (see Chapter 5 of Rolls and Deco (2002), Rolls (2007i), and Fig. 4.7). (Precursor cells to face-responsive neurons might, it is suggested, respond to combinations of the outputs of the neurons in V1 that are activated by faces, and might be found in areas such as V4.) It is an important part of this suggestion that some local spatial information would be inherent in the features which were being combined. For example, cells might not respond to the combination of an edge and a small circle unless they were in the correct spatial relation to each other. (This is in fact consistent with the data of Tanaka, Saito, Fukada and Moriya (1990), and with our data on face neurons, in that some faces neurons require the face features to be in the correct spatial configuration, and not jumbled, Rolls, Tovee, Purcell, Stewart and Azzopardi (1994b).) The local spatial information in the features being combined would ensure that the representation at the next level would contain some information about the (local) arrangement of features. Further low-order combinations of such neurons at the next stage would include sufficient local spatial information so that an arbitrary spatial arrangement of the same features would not activate the same neuron, and this is the proposed, and limited, solution which this mechanism would provide for the feature binding problem (Elliffe, Rolls and Stringer 2002) (cf. Malsburg (1990)). By this stage of processing a view-dependent representation of objects suitable for view-dependent processes such as behavioural responses to face expression and gesture would be available.

It is suggested that view-independent representations could be formed by the same type of computation, operating to combine a limited set of views of objects. The plausibility of providing view-independent recognition of objects by combining a set of different views of objects has been proposed by a number of investigators (Koenderink and Van Doorn 1979, Poggio and Edelman 1990, Logothetis, Pauls, Bülthoff and Poggio 1994, Ullman 1996). Consistent with the suggestion that the view-independent representations are formed by combining view-dependent representations in the primate visual system, is the fact that in the temporal cortical areas, neurons with view-independent representations of faces are present in the same cortical areas as neurons with view-dependent representations (from which the view-independent neurons could receive inputs) (Hasselmo, Rolls, Baylis and Dalwa 1989b, Perrett, Smith, Potter, Mistlin, Head, Milner and Jeeves 1985b, Booth and Rolls 1998). This solution to ‘object-based’ representations is very different from that traditionally proposed for artificial vision systems, in which the coordinates in 3D space of objects are stored in a database, and general-purpose algorithms operate on these to perform transforms such as translation, rotation, and scale change in 3D space (e.g. Marr (1982)). In the present, much more limited but more biologically plausible scheme, the representation would be suitable for recognition of an object, and for linking associative memories to objects, but would be less good for making actions in 3D space to particular parts of, or inside, objects, as the 3D coordinates of each part of the object would not be explicitly available. It is therefore proposed that visual
fixation is used to locate in foveal vision part of an object to which movements must be made, and that local disparity and other measurements of depth then provide sufficient information for the motor system to make actions relative to the small part of space in which a local, view-dependent, representation of depth would be provided (cf. Ballard (1990)).

The computational processes proposed above operate by an unsupervised learning mechanism, which utilizes statistical regularities in the physical environment to enable representations to be built. In some cases it may be advantageous to utilize some form of mild teaching input to the visual system, to enable it to learn for example that rather similar visual inputs have very different consequences in the world, so that different representations of them should be built. In other cases, it might be helpful to bring representations together, if they have identical consequences, in order to use storage capacity efficiently. It is proposed elsewhere (Rolls 1989b, Rolls 1989f, Rolls and Treves 1998) (see Section 1.11) that the backprojections from each adjacent cortical region in the hierarchy (and from the amygdala and hippocampus to higher regions of the visual system) play such a role by providing guidance to the competitive networks suggested above to be important in each cortical area. This guidance, and also the capability for recall, are it is suggested implemented by Hebb-modifiable connections from the backprojecting neurons to the principal (pyramidal) neurons of the competitive networks in the preceding stages (Rolls 1989b, Rolls 1989f, Rolls and Treves 1998) (see Section B.4).

The computational processes outlined above use sparse distributed coding with relatively finely tuned neurons with a graded response region centred about an optimal response achieved when the input stimulus matches the synaptic weight vector on a neuron. The distributed nature of the coding but with fine tuning would help to limit the combinatorial explosion, to keep the number of neurons within the biological range. The graded response region would be crucial in enabling the system to generalize correctly to solve for example the invariances. However, such a system would need many neurons, each with considerable learning capacity, to solve visual perception in this way. This is fully consistent with the large number of neurons in the visual system, and with the large number of, probably modifiable, synapses on each neuron (e.g. 5,000). Further, the fact that many neurons are tuned in different ways to faces is consistent with the fact that in such a computational system, many neurons would need to be sensitive (in different ways) to faces, in order to allow recognition of many individual faces when all share a number of common properties.

4.5 The feature hierarchy approach to invariant object recognition: computational issues

The feature hierarchy approach to invariant object recognition was introduced in Section 4.3.5, and advantages and disadvantages of it were discussed. Hypotheses about how object recognition could be implemented in the brain which are consistent with much of the neurophysiology discussed in Section 4.2 and by Rolls and Deco (2002) and Rolls (2007i) were set out in Section 4.4. These hypotheses effectively incorporate a feature hierarchy system while encompassing much of the neurophysiological evidence. In this Section (4.5), we consider the computational issues that arise in such feature hierarchy systems, and in the brain systems that implement visual object recognition. The issues are considered with the help of a particular model, VisNet, which requires precise specification of the hypotheses, and at the same time enables them to be explored and tested numerically and quantitatively. However, we emphasize that the issues to be covered in Section 4.5 are key and major computational issues for architectures of this feature hierarchical type (Rolls and Stringer 2006), and are very relevant to understanding how invariant object recognition is implemented in the brain.
Computational issues in feature hierarchies

VisNet is a model of invariant object recognition based on Rolls’ (1992a) hypotheses. It is a computer simulation that allows hypotheses to be tested and developed about how multilayer hierarchical networks of the type believed to be implemented in the visual cortical pathways operate. The architecture captures a number of aspects of the architecture of the visual cortical pathways, and is described next. The model of course, as with all models, requires precise specification of what is to be implemented, and at the same time involves specified simplifications of the real architecture, as investigations of the fundamental aspects of the information processing being performed are more tractable in a simplified and at the same time quantitatively specified model. First the architecture of the model is described, and this is followed by descriptions of key issues in such multilayer feature hierarchical models, such as the issue of feature binding, the optimal form of training rule for the whole system to self-organize, the operation of the network in natural environments and when objects are partly occluded, how outputs about individual objects can be read out from the network, and the capacity of the system.

4.5.1 The architecture of VisNet

Fundamental elements of Rolls’ (1992a) theory for how cortical networks might implement invariant object recognition are described in Section 4.4. They provide the basis for the design of VisNet, and can be summarized as:

- A series of competitive networks, organized in hierarchical layers, exhibiting mutual inhibition over a short range within each layer. These networks allow combinations of features or inputs occurring in a given spatial arrangement to be learned by neurons, ensuring that higher order spatial properties of the input stimuli are represented in the network.
- A convergent series of connections from a localized population of cells in preceding layers to each cell of the following layer, thus allowing the receptive field size of cells to increase through the visual processing areas or layers.
- A modified Hebb-like learning rule incorporating a temporal trace of each cell’s previous activity, which, it is suggested, will enable the neurons to learn transform invariances.

The first two elements of Rolls’ theory are used to constrain the general architecture of a network model, VisNet, of the processes just described that is intended to learn invariant representations of objects. The simulation results described in this chapter using VisNet show that invariant representations can be learned by the architecture. It is moreover shown that successful learning depends crucially on the use of the modified Hebb rule. The general architecture simulated in VisNet, and the way in which it allows natural images to be used as stimuli, has been chosen to enable some comparisons of neuronal responses in the network and in the brain to similar stimuli to be made.

4.5.1.1 The trace rule

The learning rule implemented in the VisNet simulations utilizes the spatio-temporal constraints placed upon the behaviour of ‘real-world’ objects to learn about natural object transformations. By presenting consistent sequences of transforming objects the cells in the network can learn to respond to the same object through all of its naturally transformed states, as described by Foldiák (1991), Rolls (1992a), Wallis, Rolls and Foldiák (1993), and Wallis and Rolls (1997). The learning rule incorporates a decaying trace of previous cell activity and is henceforth referred to simply as the ‘trace’ learning rule. The learning paradigm we describe here is intended in principle to enable learning of any of the transforms tolerated by inferior temporal cortex neurons, including position, size, view, lighting, and spatial

To clarify the reasoning behind this point, consider the situation in which a single neuron is strongly activated by a stimulus forming part of a real world object. The trace of this neuron’s activation will then gradually decay over a time period in the order of 0.5 s. If, during this limited time window, the net is presented with a transformed version of the original stimulus then not only will the initially active afferent synapses modify onto the neuron, but so also will the synapses activated by the transformed version of this stimulus. In this way the cell will learn to respond to either appearance of the original stimulus. Making such associations works in practice because it is very likely that within short time periods different aspects of the same object will be being inspected. The cell will not, however, tend to make spurious links across stimuli that are part of different objects because of the unlikelihood in the real world of one object consistently following another.

Various biological bases for this temporal trace have been advanced:

- The persistent firing of neurons for as long as 100–400 ms observed after presentations of stimuli for 16 ms (Rolls and Tovee 1994) could provide a time window within which to associate subsequent images. Maintained activity may potentially be implemented by recurrent connections between as well as within cortical areas (Rolls and Treves 1998, Rolls and Deco 2002) (see Chapter 5).
- The binding period of glutamate in the NMDA channels, which may last for 100 ms or more, may implement a trace rule by producing a narrow time window over which the average activity at each presynaptic site affects learning (Rolls 1992a, Rhodes 1992, Földiák 1992, Spruston, Jonas and Sakmann 1995, Hestrin, Sah and Nicoll 1990).
- Chemicals such as nitric oxide may be released during high neural activity and gradually decay in concentration over a short time window during which learning could be enhanced (Földiák 1992, Montague, Gally and Edelman 1991).

The trace update rule used in the baseline simulations of VisNet (Wallis and Rolls 1997) is equivalent to both Földiák’s used in the context of translation invariance and to the earlier rule of Sutton and Barto (1981) explored in the context of modelling the temporal properties of classical conditioning, and can be summarized as follows:

\[
\delta w_{ij} = \alpha \gamma x_j
\]

where

\[
\gamma = (1 - \eta)y^\gamma + \eta y^{-1}
\]

and

\[\text{The precise mechanisms involved may alter the precise form of the trace rule which should be used. Földiák (1992) describes an alternative trace rule which models individual NMDA channels. Equally, a trace implemented by extended cell firing should be reflected in representing the trace as an external firing rate, rather than an internal signal.}\]

\[\text{The prolonged firing of inferior temporal cortex neurons during memory delay periods of several seconds, and associative links reported to develop between stimuli presented several seconds apart (Miyashita 1988) are on too long a time scale to be immediately relevant to the present theory. In fact, associations between visual events occurring several seconds apart would, under normal environmental conditions, be detrimental to the operation of a network of the type described here, because they would probably arise from different objects. In contrast, the system described benefits from associations between visual events which occur close in time (typically within 1 s), as they are likely to be from the same object.}\]
To bound the growth of each neuron’s synaptic weight vector, $w_i$ for the $i$th neuron, its length is explicitly normalized (a method similarly employed by von der Malsburg (1973) which is commonly used in competitive networks, see Section B.4). An alternative, more biologically relevant implementation, using a local weight bounding operation which utilizes a form of heterosynaptic long-term depression (see Section 1.5), has in part been explored using a version of the Oja (1982) rule (see Wallis and Rolls (1997)).

4.5.1.2 The network implemented in VisNet

The network itself is designed as a series of hierarchical, convergent, competitive networks, in accordance with the hypotheses advanced above. The actual network consists of a series of four layers, constructed such that the convergence of information from the most disparate parts of the network’s input layer can potentially influence firing in a single neuron in the final layer – see Fig. 4.2. This corresponds to the scheme described by many researchers (Van Essen, Anderson and Felleman 1992, Rolls 1992a, for example) as present in the primate visual system – see Fig. 4.2. The forward connections to a cell in one layer are derived from a topologically related and confined region of the preceding layer. The choice of whether a connection between neurons in adjacent layers exists or not is based upon a Gaussian distribution of connection probabilities which roll off radially from the focal point of connections for each neuron. (A minor extra constraint precludes the repeated connection of any pair of cells.) In particular, the forward connections to a cell in one layer come from a small region of the preceding layer defined by the radius in Table 4.1 which will contain approximately 67% of the connections from the preceding layer. Figure 4.2 shows the general convergent network architecture used. Localization and limitation of connectivity in the network is intended to mimic cortical connectivity, partially because of the clear retention of retinal topology through regions of visual cortex. This architecture also encourages the gradual combination of features from layer to layer which has relevance to the binding problem, as described in Section 4.5.5.

Table 4.1 VisNet dimensions

<table>
<thead>
<tr>
<th>Layer</th>
<th>Dimensions</th>
<th># Connections</th>
<th>Radius</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer 4</td>
<td>32x32</td>
<td>100</td>
<td>12</td>
</tr>
<tr>
<td>Layer 3</td>
<td>32x32</td>
<td>100</td>
<td>9</td>
</tr>
<tr>
<td>Layer 2</td>
<td>32x32</td>
<td>100</td>
<td>6</td>
</tr>
<tr>
<td>Layer 1</td>
<td>32x32</td>
<td>272</td>
<td>6</td>
</tr>
<tr>
<td>Input layer</td>
<td>128x128x32</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Modelling topological constraints in connectivity leads to an issue concerning neurons at the edges of the network layers. In principle these neurons may either receive no input from beyond the edge of the preceding layer, or have their connections repeatedly sample neurons at the edge of the previous layer. In practice either solution is liable to introduce artificial weighting on the few active inputs at the edge and hence cause the edge to have unwanted influence over the development of the network as a whole. In the real brain such edge-effects would be naturally smoothed by the transition of the locus of cellular input from the fovea to the lower acuity periphery of the visual field. However, it poses a problem here because we are in effect only simulating the small high acuity foveal portion of the visual field in our simulations. As an alternative to the former solutions Wallis and Rolls (1997) elected to
4.5.1.3 Competition and lateral inhibition

In order to act as a competitive network some form of mutual inhibition is required within each layer, which should help to ensure that all stimuli presented are evenly represented by the neurons in each layer. This is implemented in VisNet by a form of lateral inhibition. The idea behind the lateral inhibition, apart from this being a property of cortical architecture in the brain, was to prevent too many neurons that received inputs from a similar part of the preceding layer responding to the same activity patterns. The purpose of the lateral inhibition was to ensure that different receiving neurons coded for different inputs. This is important in reducing redundancy (see Section B.4). The lateral inhibition is conceived as operating within a radius that was similar to that of the region within which a neuron received converging inputs from the preceding layer (because activity in one zone of topologically organized processing within a layer should not inhibit processing in another zone in the same layer, concerned perhaps with another part of the image)\textsuperscript{16}.

The lateral inhibition and contrast enhancement just described is actually implemented in VisNet2 (Rolls and Milward 2000) in two stages, to produce filtering of the type illustrated in Fig. 4.21. This lateral inhibition is implemented by convolving the activation of the neurons in a layer with a spatial filter, \( I \), where \( \delta \) controls the contrast and \( \sigma \) controls the width, and \( a \) and \( b \) index the distance away from the centre of the filter.

\[
I_{a,b} = \begin{cases} 
-\delta e^{-\frac{(x^2+y^2)^2}{\sigma^2}} & \text{if } a \neq 0 \text{ or } b \neq 0, \\
1 - \sum_{a\neq0,b\neq0} I_{a,b} & \text{if } a = 0 \text{ and } b = 0.
\end{cases}
\]  

(4.5)

form the connections into a toroid, such that connections wrap back onto the network from opposite sides. This wrapping happens at all four layers of the network, and in the way an image on the ‘retina’ is mapped to the input filters. This solution has the advantage of making all of the boundaries effectively invisible to the network. Further, this procedure does not itself introduce problems into evaluation of the network for the problems set, as many of the critical comparisons in VisNet involve comparisons between a network with the same architecture trained with the trace rule, or with the Hebb rule, or not trained at all. In practice, it is shown below that only the network trained with the trace rule solves the problem of forming invariant representations.

\textsuperscript{16}Although the extent of the lateral inhibition actually investigated by Wallis and Rolls (1997) in VisNet operated over adjacent pixels, the lateral inhibition introduced by Rolls and Milward (2000) in what they named VisNet2 and which has been used in subsequent simulations operates over a larger region, set within a layer to approximately half of the radius of convergence from the preceding layer. Indeed, Rolls and Milward (2000) showed in a problem in which invariant representations over 49 locations were being used with a 17 face test set, that the best performance was with intermediate range lateral inhibition, using the parameters for \( \sigma \) shown in Table 4.3. These values of \( \sigma \) set the lateral inhibition radius within a layer to be approximately half that of the spread of the excitatory connections from the preceding layer.
Table 4.2 Sigmoid parameters for the runs with 25 locations by Rolls and Milward (2000)

<table>
<thead>
<tr>
<th>Layer</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentile</td>
<td>99.2</td>
<td>98</td>
<td>88</td>
<td>91</td>
</tr>
<tr>
<td>Slope $\beta$</td>
<td>190</td>
<td>40</td>
<td>75</td>
<td>26</td>
</tr>
</tbody>
</table>

Table 4.3 Lateral inhibition parameters for the 25-location runs

<table>
<thead>
<tr>
<th>Layer</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radius, $\sigma$</td>
<td>1.38</td>
<td>2.7</td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td>Contrast, $\delta$</td>
<td>1.5</td>
<td>1.5</td>
<td>1.6</td>
<td>1.4</td>
</tr>
</tbody>
</table>

This is a filter that leaves the average activity unchanged. A modified version of this filter designed as a difference of Gaussians with the same inhibition but shorter range local excitation is being tested to investigate whether the self-organizing maps that this promotes (Section B.4.6) helps the system to provide some continuity in the representations formed. The concept is that this may help the system to code efficiently for large numbers of untrained stimuli that fall between trained stimuli in similarity space.

The second stage involves contrast enhancement. In VisNet (Wallis and Rolls 1997), this was implemented by raising the neuronal activations to a fixed power and normalizing the resulting firing within a layer to have an average firing rate equal to 1.0. In VisNet2 (Rolls and Milward 2000) and in subsequent simulations a more biologically plausible form of the activation function, a sigmoid, was used:

$$y = f^{\text{sigmoid}}(r) = \frac{1}{1 + e^{-\beta(r-\alpha)}}$$  \hspace{1cm} (4.6)

where $r$ is the activation (or firing rate) of the neuron after the lateral inhibition, $y$ is the firing rate after the contrast enhancement produced by the activation function, and $\beta$ is the slope or gain and $\alpha$ is the threshold or bias of the activation function. The sigmoid bounds the firing rate between 0 and 1 so global normalization is not required. The slope and threshold are held constant within each layer. The slope is constant throughout training, whereas the threshold is used to control the sparseness of firing rates within each layer. The (population) sparseness of the firing within a layer is defined (Rolls and Treves 1998) as:

$$\alpha = \frac{(\sum y_i / n)^2}{\sum y_i^2 / n}$$  \hspace{1cm} (4.7)

where $n$ is the number of neurons in the layer. To set the sparseness to a given value, e.g. 5%, the threshold is set to the value of the 95th percentile point of the activations within the layer. (Unless otherwise stated here, the neurons used the sigmoid activation function as just described.)

In most simulations with VisNet2 and later, the sigmoid activation function was used with parameters (selected after a number of optimization runs) as shown in Table 4.2.

In addition, the lateral inhibition parameters normally used in VisNet2 simulations are as shown in Table 4.3\textsuperscript{17}.

\textsuperscript{17}Where a power activation function was used in the simulations of Wallis and Rolls (1997), the power for layer 1 was 6, and for the other layers was 2.
4.5.1.4 The input to VisNet

VisNet is provided with a set of input filters which can be applied to an image to produce inputs to the network which correspond to those provided by simple cells in visual cortical area 1 (V1). The purpose of this is to enable within VisNet the more complicated response properties of cells between V1 and the inferior temporal cortex (IT) to be investigated, using as inputs natural stimuli such as those that could be applied to the retina of the real visual system. This is to facilitate comparisons between the activity of neurons in VisNet and those in the real visual system, to the same stimuli. In VisNet no attempt is made to train the response properties of simple cells, but instead we start with a defined series of filters to perform fixed feature extraction to a level equivalent to that of simple cells in V1, as have other researchers in the field (Hummel and Biederman 1992, Buhmann, Lange, von der Malsburg, Vorbrüggen and Würtz 1991, Fukushima 1980), because we wish to simulate the more complicated response properties of cells between V1 and the inferior temporal cortex (IT). The elongated orientation-tuned input filters used accord with the general tuning profiles of simple cells in V1 (Hawken and Parker 1987) and are computed by weighting the difference of two Gaussians by a third orthogonal Gaussian as described by Wallis, Rolls and Foldiak (1993) and Wallis and Rolls (1997). Each individual filter is tuned to spatial frequency (0.0625 to 0.5 cycles/pixel over four octaves); orientation (0° to 135° in steps of 45°); and sign (±1). Of the 272 layer 1 connections, the number to each group is as shown in Table 4.4. In the VisNet2 (Rolls and Milward 2000) (used for most VisNet simulations) only even symmetric – ‘bar detecting’ – filter shapes are used, which take the form of a Gaussian shape along the axis of orientation tuning for the filter, and a difference of Gaussians along the perpendicular axis.

This filter is referred to as an oriented difference of Gaussians, or DOG filter. It was chosen for VisNet in preference to the often used Gabor filter on the grounds of its better fit to available neurophysiological data including its zero D.C. response (Hawken and Parker 1987, Wallis, Rolls and Foldiak 1993). Any zero D.C. filter can of course produce a negative as well as positive output, which would mean that this simulation of a simple cell would permit negative as well as positive firing. In contrast to some other models the response of each filter is zero thresholded, and the negative results used to form a separate anti-phase input to the network. The filter outputs are also normalized across scales to compensate for the low frequency bias in the images of natural objects. However, Gabor filters have also been tested, and also produce good results with VisNet (Deco and Rolls 2004) (see Section 6.13).

Cells of layer 1 receive a topologically consistent, localized, random selection of the filter responses in the input layer, under the constraint that each cell samples every filter spatial frequency and receives a constant number of inputs. Figure 4.22 shows pictorially the general filter sampling paradigm, and Fig. 4.23 the typical connectivity to a layer 1 cell from the filters of the input layer. The blank squares indicate that no connection exists between the layer 1 cell chosen and the filters of that particular orientation, sign, and spatial frequency.

4.5.1.5 Measures for network performance

A neuron can be said to have learnt an invariant representation if it discriminates one set of stimuli from another set, across all transformations. For example, a neuron’s response is translation invariant if its response to one set of stimuli irrespective of presentation is

---

Table 4.4 VisNet layer 1 connectivity. The frequency is in cycles per pixel

<table>
<thead>
<tr>
<th>Frequency</th>
<th>0.5</th>
<th>0.25</th>
<th>0.125</th>
<th>0.0625</th>
</tr>
</thead>
<tbody>
<tr>
<td># Connections</td>
<td>201</td>
<td>50</td>
<td>13</td>
<td>8</td>
</tr>
</tbody>
</table>

---

18 Professor R. Watt, of Stirling University, is thanked for assistance with the implementation of this filter scheme.
Computational issues in feature hierarchies

Fig. 4.22 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.

consistently higher than for all other stimuli irrespective of presentation location. Note that we state ‘set of stimuli’ since neurons in the inferior temporal cortex are not generally selective for a single stimulus but rather a subpopulation of stimuli (Baylis, Rolls and Leonard 1985, Abbott, Rolls and Tovee 1996, Rolls, Treves and Tovee 1997b, Rolls and Treves 1998, Rolls and Deco 2002, Rolls 2007i, Franco, Rolls, Aggelopoulos and Jerez 2007) (see Appendix C). The measure of network performance used in VisNet1 (Wallis and Rolls 1997), the ‘Fisher metric’ (referred to in some figure labels as the Discrimination Factor), reflects how well a neuron discriminates between stimuli, compared to how well it discriminates between different locations (or more generally the images used rather than the objects, each of which is represented by a set of images, over which invariant stimulus or object representations must be learned). The Fisher measure is very similar to taking the ratio of the two F values in a two-way ANOVA, where one factor is the stimulus shown, and the other factor is the position in which a stimulus is shown. The measure takes a value greater than 1.0 if a neuron has more different responses to the stimuli than to the locations. That is, values greater than 1 indicate invariant representations when this measure is used in the following figures. Further details of how the measure is calculated are given by Wallis and Rolls (1997).

Measures of network performance based on information theory and similar to those used in the analysis of the firing of real neurons in the brain (see Appendix C) were introduced by Rolls and Milward (2000) for VisNet2, and are used in later papers. A single cell information measure was introduced which is the maximum amount of information the cell has about any one stimulus / object independently of which transform (e.g. position on the retina) is shown. Because the competitive algorithm used in VisNet tends to produce local representations (in which single cells become tuned to one stimulus or object), this information measure can approach \( \log_2 N_S \) bits, where \( N_S \) is the number of different stimuli. Indeed, it is an advantage of this measure that it has a defined maximal value, which enables how well the network is performing to be quantified. Rolls and Milward (2000) showed that the Fisher and single cell information measures were highly correlated, and given the advantage just noted of the information measure, it was adopted in Rolls and Milward (2000) and subsequent papers. Rolls and Milward (2000) also introduced a multiple cell information measure, which has the advantage that it provides a measure of whether all stimuli are encoded by different neurons in the network. Again, a high value of this measure indicates good performance.
Invariant visual object recognition learning

Fig. 4.23 Typical connectivity before training between a single cell in the first layer of the network and the input layer, represented by plotting the receptive fields of every input layer cell connected to the particular layer 1 cell. Separate input layer cells have activity that represents a positive (P) or negative (N) output from the bank of filters which have different orientations in degrees (the columns) and different spatial frequencies (the rows). Here the overall receptive field of the layer 1 cell is centred just below the centre-point of the retina. The connection scheme allows for relatively fewer connections to lower frequency cells than to high frequency cells in order to cover a similar region of the input at each frequency. A blank square indicates that there is no connection to the layer 1 neuron from an input neuron with that particular filter type.

For completeness, we provide further specification of the two information theoretic measures, which are described in detail by Rolls and Milward (2000) (see Appendix C for introduction of the concepts). The measures assess the extent to which either a single cell, or a population of cells, responds to the same stimulus invariantly with respect to its location, yet responds differently to different stimuli. The measures effectively show what one learns about which stimulus was presented from a single presentation of the stimulus at any randomly chosen location. Results for top (4th) layer cells are shown. High information measures thus show that cells fire similarly to the different transforms of a given stimulus (object), and differently to the other stimuli. The single cell stimulus-specific information, \( I(s, R) \), is the amount of information the set of responses, \( R \), about a specific stimulus, \( s \) (see Rolls, Treves, Tovee and Panzeri (1997d) and Rolls and Milward (2000)). \( I(s, R) \) is given by

\[
I(s, R) = \sum_{r \in R} P(r|s) \log_2 \frac{P(r|s)}{P(r)}
\]  

(4.8)

where \( r \) is an individual response from the set of responses \( R \) of the neuron. For each cell the performance measure used was the maximum amount of information a cell conveyed about any one stimulus. This (rather than the mutual information, \( I(S, R) \) where \( S \) is the whole set of stimuli \( s \)), is appropriate for a competitive network in which the cells tend to become tuned to one stimulus\(^{19}\).

If all the output cells of VisNet learned to respond to the same stimulus, then the information about the set of stimuli \( S \) would be very poor, and would not reach its maximal value of \( \log_2 \) of the number of stimuli (in bits). The second measure that is used here is the information

\(^{19}\)\( I(s, R) \) has more recently been called the stimulus-specific surprise, see DeWeese and Meister (1999). Its average across stimuli is the mutual information \( I(S, R) \).
provided by a set of cells about the stimulus set, using the procedures described by Rolls, Treves and Tovee (1997b) and Rolls and Milward (2000). The multiple cell information is the mutual information between the whole set of stimuli \( S \) and of responses \( R \) calculated using a decoding procedure in which the stimulus \( s' \) that gave rise to the particular firing rate response vector on each trial is estimated. (The decoding step is needed because the high dimensionality of the response space would lead to an inaccurate estimate of the information if the responses were used directly, as described by Rolls, Treves and Tovee (1997b) and Rolls and Treves (1998).) A probability table is then constructed of the real stimuli \( s \) and the decoded stimuli \( s' \). From this probability table, the mutual information between the set of actual stimuli \( S \) and the decoded estimates \( S' \) is calculated as

\[
I(S, S') = \sum_{s, s'} P(s, s') \log_2 \frac{P(s, s')}{P(s)P(s')}
\]  

(4.9)

This was calculated for the subset of cells which had as single cells the most information about which stimulus was shown. In particular, in Rolls and Milward (2000) and subsequent papers, the multiple cell information was calculated from the first five cells for each stimulus that had maximal single cell information about that stimulus, that is from a population of 35 cells if there were seven stimuli (each of which might have been shown in for example 9 or 25 positions on the retina).

### 4.5.2 Initial experiments with VisNet

Having established a network model Wallis and Rolls (1997) (following a first report by Wallis, Rolls and Foldiak (1993)) described four experiments in which the theory of how invariant representations could be formed was tested using a variety of stimuli undergoing a number of natural transformations. In each case the network produced neurons in the final layer whose responses were largely invariant across a transformation and highly discriminating between stimuli or sets of stimuli.

#### 4.5.2.1 ‘T’, ‘L’ and ‘+’ as stimuli: learning translation invariance

One of the classical properties of inferior temporal cortex face cells is their invariant response to face stimuli translated across the visual field (Tovee, Rolls and Azzopardi 1994). In this first experiment, the learning of translation invariant representations by VisNet was investigated.

In order to test the network a set of three stimuli, based upon probable 3D edge cues – consisting of a ‘T’, ‘L’ and ‘+’ shape – was constructed\(^{20}\). The actual stimuli used are shown in Fig. 4.24. These stimuli were chosen partly because of their significance as form cues, but on a more practical note because they each contain the same fundamental features – namely a horizontal bar conjoined with a vertical bar. In practice this means that the oriented

\(^{20}\)Chakravarty (1979) describes the application of these shapes as cues for the 3D interpretation of edge junctions, and Tanaka et al. (1991) have demonstrated the existence of cells responsive to such stimuli in IT.
simple cell filters of the input layer cannot distinguish these stimuli on the basis of which features are present. As a consequence of this, the representation of the stimuli received by the network is non-orthogonal and hence considerably more difficult to classify than was the case in earlier experiments involving the trace rule described by Földiák (1991). The expectation is that layer 1 neurons would learn to respond to spatially selective combinations of the basic features thereby helping to distinguish these non-orthogonal stimuli. The trajectory followed by each stimulus consisted of sweeping left to right horizontally across three locations in the top row, and then sweeping back, right to left across the middle row, before returning to the right hand side across the bottom row – tracing out a ‘Z’ shape path across the retina. Unless stated otherwise this pattern of nine presentation locations was adopted in all image translation experiments described by Wallis and Rolls (1997).

Training was carried out by permutatively presenting all stimuli in each location a total of 800 times. The sequence described above was followed for each stimulus, with the sequence start point and direction of sweep being chosen at random for each of the 800 training trials.

Figures 4.25 and 4.26 show the response after training of a first layer neuron selective for the ‘T’ stimulus. The weighted sum of all filter inputs reveals the combination of horizontally and vertically tuned filters in identifying the stimulus. In this case many connections to the lower frequency filters have been reduced to zero by the learning process, except at the relevant orientations. This contrasts strongly with the random wiring present before training, as seen previously in Fig. 4.23. It is important that neurons at early stages of feature hierarchy networks respond to combinations of features in defined relative spatial positions, before invariance is built into the system, as this is part of the way that the binding problem is solved, as described in more detail in Section 4.5.5 and by Elliffe, Rolls and Stringer (2002). The feature combination tuning is illustrated by the VisNet layer 1 neuron shown in Figs. 4.25 and 4.26.

Likewise, Fig. 4.27 depicts two neural responses, but now from the two intermediate layers of the network, taken from the top 30 most highly invariant cells, not merely the top two or three. The gradual increase in the discrimination indicates that the tolerance to shifts of the preferred stimulus gradually builds up through the layers.

The results for layer 4 neurons are illustrated in Fig. 4.28. By this stage translation-invariant, stimulus-identifying, cells have emerged. The response profiles confirm the high level of neural selectivity for a particular stimulus irrespective of location.
Computational issues in feature hierarchies

The same cell as in the previous figure and the same input reconstruction results but separated into four rows of differing spatial frequency, and eight columns representing the four filter tuning orientations in positive and negative complementary pairs.

Fig. 4.26

The same cell as in the previous figure and the same input reconstruction results but separated into four rows of differing spatial frequency, and eight columns representing the four filter tuning orientations in positive and negative complementary pairs.

Fig. 4.27

Response profiles for two intermediate layer neurons – discrimination factors 1.34 and 1.64 – in the L, T, and + experiment.

Figure 4.29 contrasts the measure of invariance, or discrimination factor, achieved by cells in the four layers, averaged over five separate runs of the network. Translation invariance clearly increases through the layers, with a considerable increase in translation invariance between layers 3 and 4. This sudden increase may well be a result of the geometry of the network, which enables cells in layer 4 to receive inputs from any part of the input layer.

Having established that invariant cells have emerged in the final layer, we now consider the role of the trace rule, by assessing the network tested under two new conditions. Firstly, the performance of the network was measured before learning occurs, that is with its initially random connection weights. Secondly, the network was trained with $\eta$ in the trace rule set to 0, which causes learning to proceed in a traceless, standard Hebbian, fashion. (Hebbian learning is purely associative, as shown for example in equation B.19.)
Figure 4.28 Response profiles for two fourth layer neurons – discrimination factors 4.07 and 3.62 – in the L, T, and + experiment.

Figure 4.29 Variation in neural discrimination factors as a measure of performance for the top 30 most highly discriminating cells through the four layers of the network, averaged over five runs of the network in the L, T, and + experiment.

Figure 4.30 Variation in neural discrimination factors as a measure of performance for the top 30 most highly discriminating cells in the fourth layer for the three training regimes, averaged over five runs of the network.

Figure 4.30 shows the results under the three training conditions. The results show that the trace rule is the decisive factor in establishing the invariant responses in the layer 4 neurons. It is interesting to note that the Hebbian learning results are actually worse than those achieved by chance in the untrained net. In general, with Hebbian learning, the most highly discriminating cells barely rate higher than 1. This value of discrimination corresponds to the case in which a cell responds to only one stimulus and in only one location. The poor performance with the Hebb rule comes as a direct consequence of the presentation paradigm being employed.

If we consider an image as representing a vector in multidimensional space, a particular image in the top left-hand corner of the input retina will tend to look more like any other image in that same location than the same image presented elsewhere. A simple competitive
network using just Hebbian learning will thus tend to categorize images by where they are rather than what they are – the exact opposite of what the net was intended to learn. This comparison thus indicates that a small memory trace acting in the standard Hebbian learning paradigm can radically alter the normal vector averaging, image classification, performed by a Hebbian-based competitive network.

One question that emerges about the representation in the final layer of the network relates to how evenly the network divides up its resources to represent the learnt stimuli. It is conceivable that one stimulus stands out among the set of stimuli by containing very distinctive features which would make it easier to categorize. This may produce an unrepresentative number of neurons with high discrimination factors which are in fact all responding to the same stimulus. It is important that at least some cells code for (or provide information about) each of the stimuli. As a simple check on this, the preferred stimulus of each cell was found and the associated measure of discrimination added to a total for each stimulus. This measure in practice never varied by more than a factor of 1.3:1 for all stimuli. The multiple cell information measure used in some later figures addresses the same issue, with similar results.

4.5.2.2 ‘T’, ‘L’, and ‘+’ as stimuli: Optimal network parameters

The second series of investigations described by Wallis and Rolls (1997) using the ‘T’, ‘L’ and ‘+’ stimuli, centred upon finding optimal parameters for elements of the network, such as the optimal trace time constant $\tau$, which controls the relative effect of previous activities on current learning as described above. The network performance was gauged using a single 800 epoch training run of the network with the median discrimination factor (with the upper and lower quartile values) for the top sixteen cells of the fourth layer being displayed at each parameter value.

Figure 4.31 displays the effect of varying the value of $\tau$ for the nine standard presentation locations. The optimal value of $\tau$ might conceivably change with the alteration of the number of training locations, and indeed one might predict that it would be smaller if the number of presentation locations was reduced. To confirm this, network performance was also measured for presentation sweeps over only five locations. Figure 4.32 shows the results of this experiment, which confirm the expected shift in the general profile of the curve towards shorter time constant values. Of course, the optimal value of $\tau$ derived is in effect a compromise between optimal values for the three layers in which the trace operates. Since neurons in each layer have different effective receptive field sizes, one would expect each layer’s neurons to be exposed to different portions of the full sweep of a particular stimulus. This would in turn suggest that the optimal value of $\tau$ will grow through the layers.

4.5.2.3 Faces as stimuli: translation invariance

The aim of the next set of experiments described by Wallis and Rolls (1997) was to start to address the issues of how the network operates when invariant representations must be learned for a larger number of stimuli, and whether the network can learn when much more complicated, real biological stimuli (faces) are used. The set of face images used appears in Fig. 4.33. In practice, to equalize luminance the D.C. component of the images was removed. In addition, so as to minimize the effect of cast shadows, an oval Hamming window was applied to the face image which also served to remove any hard edges of the image relative to the plain background upon which they were set.

The results of training in the translation invariance paradigm with 7 faces each in 9 locations are shown in Figs. 4.34, 4.35 and 4.36. The network produces neurons with high discrimination factors, and this only occurs if it is trained with the trace rule. Some layer 4 neurons showed a somewhat distributed representation, as illustrated in the examples of layer 4 neurons shown in Fig. 4.34.
In order to check that there was an invariant representation in layer 4 of VisNet that could be read by a receiving population of neurons, a fifth layer was added to the net which fully sampled the fourth layer cells. This layer was in turn trained in a supervised manner using gradient descent or with a Hebbian associative learning rule. Wallis and Rolls (1997) showed that the object classification performed by the layer 5 network was better if the network had been trained with the trace rule than when it was untrained or was trained with a Hebb rule.

4.5.2.4 Faces as stimuli: view invariance

Given that the network had been shown to be able to operate usefully with a more difficult translation invariance problem, we next addressed the question of whether the network can solve other types of transform invariance, as we had intended. The next experiment addressed this question, by training the network on the problem of 3D stimulus rotation, which produces non-isomorphic transforms, to determine whether the network can build a view-invariant categorization of the stimuli (Wallis and Rolls 1997). The trace rule learning paradigm should, in conjunction with the architecture described here, prove capable of learning any of the transforms tolerated by IT neurons, so long as each stimulus is presented in short sequences during which the transformation occurs and can be learned. This experiment continued with the use of faces but now presented them centrally in the retina in a sequence of different views of a face. The images used are shown in Fig. 4.37. The faces were again smoothed at the
edges to erase the harsh image boundaries, and the D.C. term was removed. During the 800
epochs of learning, each stimulus was chosen at random, and a sequence of preset views of it
was shown, rotating the face either to the left or to the right.

Although the actual number of images being presented is smaller, some 21 views in all,
there is good reason to think that this problem may be harder to solve than the previous
translation experiments. This is simply due to the fact that all 21 views exactly overlap with
one another. The net was indeed able to solve the invariance problem, with examples of
invariant layer 4 neuron response profiles appearing in Fig. 4.38.

Figure 4.39 confirms the improvement in invariant stimulus representation found through
the layers, and that layer 4 provides a considerable improvement in performance over the
previous layers. Figure 4.40 shows the Hebb trained and untrained nets performing equally
poorly, whilst the trace trained net shows good invariance across the entire 30 cells selected.

Fig. 4.34  Response profiles for two neurons in the fourth layer – discrimination factors 2.64 and 2.10. The
net was trained on 7 faces each in 9 locations.

Fig. 4.35  Variation in network performance for the
top 30 most highly discriminating cells through the
four layers of the network, averaged over five runs
of the network. The net was trained on 7 faces
each in 9 locations.

Fig. 4.36  Variation in network performance for the
top 30 most highly discriminating cells in the fourth
layer for the three training regimes, averaged over
five runs of the network. The net was trained on 7
faces each in 9 locations.
Fig. 4.37 Three faces in seven different views used as stimuli in an experiment by Wallis and Rolls (1997).

Fig. 4.38 Response profiles for cells in the last two layers of the network – discrimination factors 11.12 and 12.40 – in the experiment with seven different views of each of three faces.

4.5.3 The optimal parameters for the temporal trace used in the learning rule

The trace used in VisNet enables successive features that, based on the natural statistics of the visual input, are likely to be from the same object or feature complex to be associated together. For good performance, the temporal trace needs to be sufficiently long that it covers the period in which features seen by a particular neuron in the hierarchy are likely to come from the same object. On the other hand, the trace should not be so long that it produces associations between features that are parts of different objects, seen when for example the eyes move to another object. One possibility is to reset the trace during saccades between different objects. If explicit trace resetting is not implemented, then the trace should, to optimize the compromise implied by the above, lead to strong associations between temporally close stimuli, and increasingly weaker associations between temporally more distant stimuli. In fact, the trace implemented in VisNet has an exponential decay, and it has been shown that this form is optimal in the situation where the exact duration over which the same object is being viewed varies, and where the natural statistics of the visual input happen also to show a decreasing probability that the same object is being viewed as the time period in question increases (Wallis and...
Baddeley 1997). Moreover, as is made evident in Figs. 4.32 and 4.31, performance can be enhanced if the duration of the trace does at the same time approximately match the period over which the input stimuli are likely to come from the same object or feature complex. Nevertheless, good performance can be obtained in conditions under which the trace rule allows associations to be formed only between successive items in the visual stream (Rolls and Milward 2000, Rolls and Stringer 2001a).

It is also the case that the optimal value of \( \eta \) in the trace rule is likely to be different for different layers of VisNet, and for cortical processing in the ‘what’ visual stream. For early layers of the system, small movements of the eyes might lead to different feature combinations providing the input to cells (which at early stages have small receptive fields), and a short duration of the trace would be optimal. However, these small eye movements might be around the same object, and later layers of the architecture would benefit from being able to associate together their inputs over longer times, in order to learn about the larger scale properties that characterize individual objects, including for example different views of objects observed as an object turns or is turned. Thus the suggestion is made that the temporal trace could be effectively longer at later stages (e.g. inferior temporal visual cortex) compared to early stages (e.g. V2 and V4) of processing in the visual system. In addition, as will be shown in Section 4.5.5, it is important to form feature combinations with high spatial precision before invariance learning supported by a temporal trace starts, in order that the feature combinations and not the individual features have invariant representations. This leads to the suggestion that the trace rule should either not operate, or be short, at early stages of cortical visual processing such as V1. This is reflected in the operation of VisNet2, which does not use a temporal trace in layer 1 (Rolls and Milward 2000).

### 4.5.4 Different forms of the trace learning rule, and their relation to error correction and temporal difference learning

The original trace learning rule used in the simulations of Wallis and Rolls (1997) took the form

\[
\delta w_j = \alpha \eta r x^T_j
\]  (4.10)
where the trace $\mathbb{T}$ is updated according to

$$
\mathbb{T} = (1 - \eta)y^\tau + \eta \mathbb{T}^{-1}.
$$

(4.11)

The parameter $\eta \in [0, 1]$ controls the relative contributions to the trace $\mathbb{T}$ from the instantaneous firing rate $y^\tau$ and the trace at the previous time step $\mathbb{T}^{-1}$, where for $\eta = 0$ we have $\mathbb{T} = y^\tau$ and equation 4.10 becomes the standard Hebb rule

$$
\delta w_j = \alpha y^\tau x_j^\tau.
$$

(4.12)

At the start of a series of investigations of different forms of the trace learning rule, Rolls and Milward (2000) demonstrated that VisNet’s performance could be greatly enhanced with a modified Hebbian learning rule that incorporated a trace of activity from the preceding time steps, with no contribution from the activity being produced by the stimulus at the current time step. This rule took the form

$$
\delta w_j = \alpha \mathbb{T}^{-1} x_j^\tau.
$$

(4.13)

The trace shown in equation 4.13 is in the postsynaptic term, and similar effects were found if the trace was in the presynaptic term, or in both the pre- and the postsynaptic terms. The crucial difference from the earlier rule (see equation 4.10) was that the trace should be calculated up to only the preceding timestep, with no contribution to the trace from the firing on the current trial to the current stimulus. How might this be understood?

One way to understand this is to note that the trace rule is trying to set up the synaptic weight on trial $\tau$ based on whether the neuron, based on its previous history, is responding to that stimulus (in other transforms, e.g. position). Use of the trace rule at $\tau - 1$ does this, that is it takes into account the firing of the neuron on previous trials, with no contribution from the firing being produced by the stimulus on the current trial. On the other hand, use of the trace at time $\tau$ in the update takes into account the current firing of the neuron to the stimulus in that particular position, which is not a good estimate of whether that neuron should be allocated to invariantly represent that stimulus. Effectively, using the trace at time $\tau$ introduces a Hebbian element into the update, which tends to build position-encoded analyzers, rather than stimulus-encoded analyzers. (The argument has been phrased for a system learning translation invariance, but applies to the learning of all types of invariance.) A particular advantage of using the trace at $\tau - 1$ is that the trace will then on different occasions (due to the randomness in the location sequences used) reflect previous histories with different sets of positions, enabling the learning of the neuron to be based on evidence from the stimulus present in many different positions. Using a term from the current firing in the trace (i.e. the trace calculated at time $\tau$) results in this desirable effect always having an undesirable element from the current firing of the neuron to the stimulus in its current position.

4.5.4.1 The modified Hebbian trace rule and its relation to error correction

The rule of equation 4.13 corrects the weights using a postsynaptic trace obtained from the previous firing (produced by other transforms of the same stimulus), with no contribution to the trace from the current postsynaptic firing (produced by the current transform of the stimulus). Indeed, insofar as the current firing $y^\tau$ is not the same as $\mathbb{T}^{-1}$, this difference can be thought of as an error. This leads to a conceptualization of using the difference between the current firing and the preceding trace as an error correction term, as noted in the context of modelling the temporal properties of classical conditioning by Sutton and Barto (1981), and developed next in the context of invariance learning (see Rolls and Stringer (2001a)).

First, we re-express the rule of equation 4.13 in an alternative form as follows. Suppose we are at timestep $\tau$ and have just calculated a neuronal firing rate $y^\tau$ and the corresponding trace
\( \mathbf{f} \) from the trace update equation 4.11. If we assume \( \eta \in (0,1) \), then rearranging equation 4.11 gives

\[
\mathbf{f}^{-1} = \frac{1}{\eta}(\mathbf{f} - (1-\eta)y^\tau), \tag{4.14}
\]

and substituting equation 4.14 into equation 4.13 gives

\[
\delta w_j = \frac{1}{\eta}(\mathbf{f} - (1-\eta)y^\tau)x_j^\tau = \frac{1-\eta}{\eta}(1-\eta\mathbf{f} - y^\tau)x_j^\tau = \hat{\alpha}(\beta\mathbf{f} - y^\tau)x_j^\tau \tag{4.15}
\]

where \( \hat{\alpha} = \frac{1-\eta}{\eta} \) and \( \beta = \frac{1}{1-\eta} \). The modified Hebbian trace learning rule (4.13) is thus equivalent to equation 4.15 which is in the general form of an error correction rule (Hertz, Krogh and Palmer 1991). That is, rule (4.15) involves the subtraction of the current firing rate \( y^\tau \) from a target value, in this case \( \beta\mathbf{f} \).

Although above we have referred to rule (4.13) as a modified Hebbian rule, we note that it is only associative in the sense of associating previous cell firing with the current cell inputs. In the next section we continue to explore the error correction paradigm, examining five alternative examples of this sort of learning rule.

### 4.5.4.2 Five forms of error correction learning rule

Error correction learning rules are derived from gradient descent minimization (Hertz, Krogh and Palmer 1991), and continually compare the current neuronal output to a target value \( t \) and adjust the synaptic weights according to the following equation at a particular timestep \( \tau \)

\[
\delta w_j = \alpha(t - y^\tau)x_j^\tau. \tag{4.16}
\]

In this usual form of gradient descent by error correction, the target \( t \) is fixed. However, in keeping with our aim of encouraging neurons to respond similarly to images that occur close together in time it seems reasonable to set the target at a particular timestep, \( t^\tau \), to be some function of cell activity occurring close in time, because encouraging neurons to respond to temporal classes will tend to make them respond to the different variants of a given stimulus (Földiák 1991, Rolls 1992a, Wallis and Rolls 1997). For this reason, Rolls and Stringer (2001a) explored a range of error correction rules where the targets \( t^\tau \) are based on the trace of neuronal activity calculated according to equation 4.11. We note that although the target is not a fixed value as in standard error correction learning, nevertheless the new learning rules perform gradient descent on each timestep, as elaborated below. Although the target may be varying early on in learning, as learning proceeds the target is expected to become more and more constant, as neurons settle to respond invariantly to particular stimuli. The first set of five error correction rules we discuss are as follows.

\[
\delta w_j = \alpha(\beta\mathbf{f}^{-1} - y^\tau)x_j^\tau, \tag{4.17}
\]

\[
\delta w_j = \alpha(\beta y^{-1} - y^\tau)x_j^\tau, \tag{4.18}
\]

\[
\delta w_j = \alpha(\beta\mathbf{f} - y^\tau)x_j^\tau, \tag{4.19}
\]

\[
\delta w_j = \alpha(\beta\mathbf{f}^{-1} - y^\tau)x_j^\tau, \tag{4.20}
\]

\[
\delta w_j = \alpha(\beta y^{-1} - y^\tau)x_j^\tau, \tag{4.21}
\]

where updates (4.17), (4.18) and (4.19) are performed at timestep \( \tau \), and updates (4.20) and (4.21) are performed at timestep \( \tau + 1 \). (The reason for adopting this convention is that the
Numerical results with the five error correction rules (4.17), (4.18), (4.19), (4.20), (4.21) (with positive clipping of synaptic weights) trained on 7 faces in 9 locations. On the left are single cell information measures, and on the right are multiple cell information measures. (After Rolls and Stringer 2001a.)

These rules are all similar except for their targets \( t^\tau \), which are all functions of a temporally nearby value of cell activity. In particular, rule (4.19) is directly related to rule (4.15), but is more general in that the parameter \( \beta = \frac{1}{1+\eta} \) is replaced by an unconstrained parameter \( \beta \). In addition, we also note that rule (4.17) is closely related to a rule developed in Peng, Sha, Gan and Wei (1998) for view invariance learning. The above five error correction rules are biologically plausible in that the targets \( t^\tau \) are all local cell variables (see Appendix B and Rolls and Treves (1998)). In particular, rule (4.19) uses the trace \( \tau \) from the current time level \( \tau \), and rules (4.18) and (4.21) do not need exponential trace values \( \tau \) instead relying only on the instantaneous firing rates at the current and immediately preceding timesteps. However, all five error correction rules involve decrementing of synaptic weights according to an error which is calculated by subtracting the current activity from a target.

Numerical results with the error correction rules trained on 7 faces in 9 locations are presented in Fig. 4.41. For all the results shown the synaptic weights were clipped to be positive during the simulation, because it is important to test that decrementing synaptic weights purely within the positive interval \( w \in [0, \infty) \) will provide significantly enhanced performance. That is, it is important to show that error correction rules do not necessarily require possibly biologically implausible modifiable negative weights. For each of the rules (4.17), (4.18), (4.19), (4.20), (4.21), the parameter \( \beta \) has been individually optimized to the following respective values: 4.9, 2.2, 2.2, 3.8, 2.2. On the left and right are results with the single and multiple cell information measures, respectively. Comparing Fig. 4.41 with Fig. 4.42 shows that all five error correction rules offer considerably improved performance over both the standard trace rule (4.10) and rule (4.13). From the left-hand side of Fig. 4.41 it can be seen that rule (4.17) performs best, and this is probably due to two reasons. Firstly, rule (4.17) incorporates an exponential trace \( \tau^{-1} \) in its target \( t^\tau \), and we would expect this to help neurons to learn more quickly to respond invariantly to a class of inputs that occur close together in time. Hence, setting \( \eta = 0 \) as in rule (4.18) results in reduced performance. Secondly, unlike rules (4.19) and (4.20), rule (4.17) does not contain any component of \( \tau^{\tau} \) in its target. If we examine rules (4.19), (4.20), we see that their respective targets \( \tau^\tau \), \( \tau^{\tau+1} \) contain significant components of \( \tau^{\tau} \).
4.5.4.3 Relationship to temporal difference learning

Rolls and Stringer (2001a) not only considered the relationship of rule (4.13) to error correction, but also considered how the error correction rules shown in equations (4.17), (4.18), (4.19), (4.20) and (4.21) are related to temporal difference learning (Sutton 1988, Sutton and Barto 1998). Sutton (1988) described temporal difference methods in the context of prediction learning. These methods are a class of incremental learning techniques that can learn to predict final outcomes through comparison of successive predictions from the preceding time steps. This is in contrast to traditional supervised learning, which involves the comparison of predictions only with the final outcome. Consider a series of multistep prediction problems in which for each problem there is a sequence of observation vectors, \( x^1, x^2, \ldots, x^m \), at successive timesteps, followed by a final scalar outcome \( z \). For each sequence of observations temporal difference methods form a sequence of predictions \( y^1, y^2, \ldots, y^m \), each of which is a prediction of \( z \). These predictions are based on the observation vectors \( x^\tau \) and a vector of modifiable weights \( w \); i.e. the prediction at time step \( \tau \) is given by \( y^\tau(x^\tau, w) \), and for a linear dependency the prediction is given by \( y^\tau = w^T x^\tau \). (Note here that \( w^T \) is the transpose of the weight vector \( w \).) The problem of prediction is to calculate the weight vector \( w \) such that the predictions \( y^\tau \) are good estimates of the outcome \( z \).

The supervised learning approach to the prediction problem is to form pairs of observation vectors \( x^\tau \) and outcome \( z \) for all time steps, and compute an update to the weights according to the gradient descent equation

\[
\delta w = \alpha (z - y^\tau) \nabla_w y^\tau \tag{4.22}
\]

where \( \alpha \) is a learning rate parameter and \( \nabla_w \) indicates the gradient with respect to the weight vector \( w \). However, this learning procedure requires all calculation to be done at the end of the sequence, once \( z \) is known. To remedy this, it is possible to replace method (4.22) with a temporal difference algorithm that is mathematically equivalent but allows the computational workload to be spread out over the entire sequence of observations. Temporal difference methods are a particular approach to updating the weights based on the values of successive predictions, \( y^\tau, y^{\tau+1} \). Sutton (1988) showed that the following temporal difference algorithm is equivalent to method (4.22)

\[
\delta w = \alpha (y^{\tau+1} - y^\tau) \sum_{k=1}^{\tau} \nabla_w y^k. \tag{4.23}
\]
where $y^{m+1} \equiv z$. However, unlike method (4.22) this can be computed incrementally at each successive time step since each update depends only on $y^{r+1}, y^r$ and the sum of $\nabla_w y^k$ over previous time steps $k$. The next step taken in Sutton (1988) is to generalize equation (4.23) to the following final form of temporal difference algorithm, known as ‘TD($\lambda$)’

$$\delta w = \alpha (y^{r+1} - y^r) \sum_{k=1}^{r} \lambda^{r-k} \nabla_w y^k$$

(4.24)

where $\lambda \in [0, 1]$ is an adjustable parameter that controls the weighting on the vectors $\nabla_w y^k$. Equation (4.24) represents a much broader class of learning rules than the more usual gradient descent-based rule (4.23), which is in fact the special case TD(1).

A further special case of equation (4.24) is for $\lambda = 0$, i.e. TD(0), as follows

$$\delta w = \alpha (y^{r+1} - y^r) \nabla_w y^r.$$  

(4.25)

But for problems where $y^r$ is a linear function of $x^r$ and $w$, we have $\nabla_w y^r = x^r$, and so equation (4.25) becomes

$$\delta w = \alpha (y^{r+1} - y^r) x^r.$$  

(4.26)

If we assume the prediction process is being performed by a neuron with a vector of inputs $x^r$, synaptic weight vector $w$, and output $y^r = w^T x^r$, then we see that the TD(0) algorithm (4.26) is identical to the error correction rule (4.21) with $\beta = 1$. In understanding this comparison with temporal difference learning, it may be useful to note that the firing at the end of a sequence of the transformed exemplars of a stimulus is effectively the temporal difference target $z$. This establishes a link to temporal difference learning (described further in Section B.15.3). Further, we note that from learning epoch to learning epoch, the target $z$ for a given neuron will gradually settle down to be more and more fixed as learning proceeds.

We now explore in more detail the relation between the error correction rules described above and temporal difference learning. For each sequence of observations with a single outcome the temporal difference method (4.26), when viewed as an error correction rule, is attempting to adapt the weights such that $y^{r+1} = y^r$ for all successive pairs of time steps – the same general idea underlying the error correction rules (4.17), (4.18), (4.19), (4.20), (4.21). Furthermore, in Sutton and Barto (1998), where temporal difference methods are applied to reinforcement learning, the TD($\lambda$) approach is again further generalized by replacing the target $y^{r+1}$ by any weighted average of predictions $y$ from arbitrary future timesteps, e.g. $t^r = \frac{1}{\beta} y^{r+3} + \frac{\beta}{\beta+1} y^{r+2}$, including an exponentially weighted average extending forward in time. So a more general form of the temporal difference algorithm has the form

$$\delta w = \alpha (t^r - y^r) x^r,$$

(4.27)

where here the target $t^r$ is an arbitrary weighted average of the predictions $y$ over future timesteps. Of course, with standard temporal difference methods the target $t^r$ is always an average over future timesteps $k = r + 1, r + 2$, etc. But in the five error correction rules this is only true for the last exemplar (4.21). This is because with the problem of prediction, for example, the ultimate target of the predictions $y^1, ..., y^m$ is a final outcome $y^{m+1} \equiv z$. However, this restriction does not apply to our particular application of neurons trained to respond to temporal classes of inputs within VisNet. Here we only wish to set the firing rates $y^1, ..., y^m$ to the same value, not some final given value $z$. However, the more general error correction rules clearly have a close relationship to standard temporal difference algorithms. For example, it can be seen that equation (4.18) with $\beta = 1$ is in some sense a temporal mirror image of equation (4.26), particularly if the updates $\delta w_j$ are added to the weights $w_j$.
only at the end of a sequence. That is, rule (4.18) will attempt to set \( y^1, ..., y^m \) to an initial value \( y^0 = z \). This relationship to temporal difference algorithms allows us to begin to exploit established temporal difference analyses to investigate the convergence properties of the error correction methods (Rolls and Stringer 2001a).

Although the main aim of Rolls and Stringer (2001a) in relating error correction rules to temporal difference learning was to begin to exploit established temporal difference analyses, they observed that the most general form of temporal difference learning, TD(\( \lambda \)), in fact suggests an interesting generalization to the existing error correction learning rules for which we currently have \( \lambda = 0 \). Assuming \( y^{\tau} = w^{\tau} x^{\tau} \) and \( \nabla_w y^{\tau} = x^{\tau} \), the general equation (4.24) for TD(\( \lambda \)) becomes

\[
\delta w = \alpha (y^{\tau+1} - y^{\tau}) \sum_{k=1}^{\tau} \lambda^{\tau-k} x^k
\]

(4.28)

where the term \( \sum_{k=1}^{\tau} \lambda^{\tau-k} x^k \) is a weighted sum of the vectors \( x^k \). This suggests generalizing the original five error correction rules (4.17), (4.18), (4.19), (4.20), (4.21) by replacing the term \( x^j \) by a weighted sum \( \tilde{x}^j = \sum_{k=1}^{\tau} \lambda^{\tau-k} x^k \) with \( \lambda \in [0, 1] \). In Sutton (1988) \( \tilde{x}^j \) is calculated according to

\[
\tilde{x}^j = x^j + \lambda \tilde{x}^{j-1}
\]

(4.29)

with \( \tilde{x}^0 = 0 \). This gives the following five temporal difference-inspired error correction rules

\[
\delta w_j = \alpha (\beta y^{\tau+1} - y^{\tau}) \tilde{x}^j,
\]

(4.30)

\[
\delta w_j = \alpha (\beta y^{\tau+1} - y^{\tau}) \tilde{x}^j,
\]

(4.31)

\[
\delta w_j = \alpha (\beta y^{\tau+1} - y^{\tau}) \tilde{x}^j,
\]

(4.32)

\[
\delta w_j = \alpha (\beta y^{\tau+1} - y^{\tau}) \tilde{x}^j,
\]

(4.33)

\[
\delta w_j = \alpha (\beta y^{\tau+1} - y^{\tau}) \tilde{x}^j,
\]

(4.34)

where it may be readily seen that equations (4.31) and (4.34) are special cases of equations (4.30) and (4.33) respectively, with \( \eta = 0 \). As with the trace \( T^\tau \), the term \( \tilde{x}^j \) is reset to zero when a new stimulus is presented. These five rules can be related to the more general TD(\( \lambda \)) algorithm, but continue to be biologically plausible using only local cell variables. Setting \( \lambda = 0 \) in rules (4.30), (4.31), (4.32), (4.33), (4.34), gives us back the original error correction rules (4.17), (4.18), (4.19), (4.20), (4.21) which may now be related to TD(0).

Numerical results with error correction rules (4.30), (4.31), (4.32), (4.33), (4.34), and \( \tilde{x}^j \) calculated according to equation (4.29) with \( \lambda = 1 \), with positive clipping of weights, trained on 7 faces in 9 locations are presented in Fig. 4.43. For each of the rules (4.30), (4.31), (4.32), (4.33), (4.34), the parameter \( \beta \) has been individually optimized to the following respective values: 1.7, 1.8, 1.5, 1.6, 1.8. On the left and right are results with the single and multiple cell information measures, respectively. Comparing these five temporal difference-inspired rules it can be seen that the best performance is obtained with rule (4.34) where many more cells reach the maximum level of performance possible with respect to the single cell information measure. In fact, this rule offered the best such results. This may well be due to the fact that this rule may be directly compared to the standard TD(1) learning rule, which itself may be related to classical supervised learning for which there are well known optimality results, as discussed further by Rolls and Stringer (2001a).

From the simulations described by Rolls and Stringer (2001a) it appears that the form of optimization described above associated with TD(1) rather than TD(0) leads to better performance within VisNet. Comparing Figs. 4.41 and 4.43 shows that the TD(1)-like rule (4.34)
with $\lambda = 1.0$ and $\beta = 1.8$ gives considerably superior results to the TD(0)-like rule (4.21) with $\beta = 2.2$. In fact, the former of these two rules provided the best single cell information results in these studies. We hypothesize that these results are related to the fact that only a finite set of image sequences is presented to VisNet, and so the type of optimization performed by TD(1) for repeated presentations of a finite data set is more appropriate for this problem than the form of optimization performed by TD(0).

**4.5.4.4 Discussion of the different training rules**

In terms of biological plausibility, we note the following. First, all the learning rules investigated by Rolls and Stringer (2001a) are local learning rules, and in this sense are biologically plausible (see Appendix B and Rolls and Treves (1998)). (The rules are local in that the terms used to modify the synaptic weights are potentially available in the pre- and post-synaptic elements.)

Second we note that all the rules do require some evidence of the activity on one or more previous stimulus presentations to be available when the synaptic weights are updated. Some of the rules, e.g. learning rule (4.19), use the trace $f^{\tau}$ from the current time level, while rules (4.18) and (4.21) do not need to use an exponential trace of the neuronal firing rate, but only the instantaneous firing rates $y$ at two successive time steps. It is known that synaptic plasticity does involve a combination of separate processes each with potentially differing time courses (Koch 1999), and these different processes could contribute to trace rule learning. Another mechanism suggested for implementing a trace of previous neuronal activity is the continuing firing for often 300 ms produced by a short (16 ms) presentation of a visual stimulus (Rolls and Tovee 1994) which is suggested to be implemented by local cortical recurrent attractor networks (Rolls and Treves 1998).

Third, we note that in utilizing the trace in the targets $t^{\tau}$, the error correction (or temporal difference inspired) rules perform a comparison of the instantaneous firing $y^{\tau}$ with a temporally nearby value of the activity, and this comparison involves a subtraction. The subtraction provides an error, which is then used to increase or decrease the synaptic weights. This is a somewhat different operation from long-term depression (LTD) as well as long term potentiation (LTP), which are associative changes which depend on the pre- and post-synaptic activity. However, it is interesting to note that an error correction rule which appears to involve a subtraction of current firing from a target might be implemented by a combination of an associative process operating with the trace, and an anti-Hebbian process operating to remove the effects of the current firing. For example, the synaptic updates $\delta w_j = \alpha (t^{\tau} - y^{\tau}) x_j^{\tau}$ can
be decomposed into two separate associative processes \( ax^T x_j^T \) and \(-\alpha q x_j^T \), that may occur independently. (The target, \( t^T \), could in this case be just the trace of previous neural activity from the preceding trials, excluding any contribution from the current firing.) Another way to implement an error correction rule using associative synaptic modification would be to force the post-synaptic neuron to respond to the error term. Although this has been postulated to be an effect which could be implemented by the climbing fibre system in the cerebellum (Ito 1989, Ito 1984, Rolls and Treves 1998), there is no similar system known for the neocortex, and it is not clear how this particular implementation of error correction might operate in the neocortex.

In Section 4.5.4.2 we describe five learning rules as error correction rules. We now discuss an interesting difference of these error correction rules from error correction rules as conventionally applied. It is usual to derive the general form of error correction learning rule from gradient descent minimization in the following way (Hertz, Krogh and Palmer 1991). Consider the idealized situation of a single neuron with a number of inputs \( x_j \) and output \( y = \sum_j w_j x_j \), where \( w_j \) are the synaptic weights. We assume that there are a number of input patterns and that for the \( k \)th input pattern, \( x^k = [x^k_1, x^k_2, \ldots]^T \), the output \( y^k \) has a target value \( t^k \). Hence an error measure or cost function can be defined as

\[
e(w) = \frac{1}{2} \sum_k (t^k - y^k)^2 = \frac{1}{2} \sum_k (t^k - \sum_j w_j x_j^k)^2.
\]

(4.35)

This cost function is a function of the input patterns \( x^k \) and the synaptic weight vector \( w = [w_1, w_2, \ldots]^T \). With a fixed set of input patterns, we can reduce the error measure by employing a gradient descent algorithm to calculate an improved set of synaptic weights. Gradient descent achieves this by moving downhill on the error surface defined in \( w \) space using the update

\[
\delta w_j = -\alpha \frac{\partial e}{\partial w_j} = \alpha \sum_k (t^k - y^k)x_j^k.
\]

(4.36)

If we update the weights after each pattern \( k \), then the update takes the form of an error correction rule

\[
\delta w_j = \alpha (t^k - y^k)x_j^k,
\]

(4.37)

which is also commonly referred to as the delta rule or Widrow–Hoff rule (see Widrow and Hoff (1960) and Widrow and Stearns (1985)). Error correction rules continually compare the neuronal output with its pre-specified target value and adjust the synaptic weights accordingly. In contrast, the way Rolls and Stringer (2001a) introduced of utilizing error correction is to specify the target as the activity trace based on the firing rate at nearby timesteps. Now the actual firing at those nearby time steps is not a pre-determined fixed target, but instead depends on how the network has actually evolved. This effectively means the cost function \( e(w) \) that is being minimized changes from timestep to timestep. Nevertheless, the concept of calculating an error, and using the magnitude and direction of the error to update the synaptic weights, is the similarity Rolls and Stringer (2001a) made to gradient descent learning.

To conclude this discussion, the error correction and temporal difference rules explored by Rolls and Stringer (2001a) provide interesting approaches to help understand invariant pattern recognition learning. Although we do not know whether the full power of these rules is expressed in the brain, we provided suggestions about how they might be implemented. At the same time, we note that the original trace rule used by Földiák (1991), Rolls (1992a), and Wallis and Rolls (1997) is a simple associative rule, is therefore biologically very plausible, and, while not as powerful as many of the other rules introduced by Rolls and Stringer (2001a), can nevertheless solve the same class of problem. Rolls and Stringer (2001a) also
emphasized that although they demonstrated how a number of new error correction and temporal difference rules might play a role in the context of view invariant object recognition, they may also operate elsewhere where it is important for neurons to learn to respond similarly to temporal classes of inputs that tend to occur close together in time.

4.5.5 The issue of feature binding, and a solution

In this section we investigate two key issues that arise in hierarchical layered network architectures, such as VisNet, other examples of which have been described and analyzed by Fukushima (1980), Ackley, Hinton and Sejnowski (1985), Rosenblatt (1961), and Riesenhuber and Poggio (1999b). One issue is whether the network can discriminate between stimuli that are composed of the same basic alphabet of features. The second issue is whether such network architectures can find solutions to the spatial binding problem. These issues are addressed next and by Elliffe, Rolls and Stringer (2002).

The first issue investigated is whether a hierarchical layered network architecture of the type exemplified by VisNet can discriminate stimuli that are composed of a limited set of features and where the different stimuli include cases where the feature sets are subsets and supersets of those in the other stimuli. An issue is that if the network has learned representations of both the parts and the wholes, will the network identify that the whole is present when it is shown, and not just that one or more parts is present. (In many investigations with VisNet, complex stimuli (such as faces) were used where each stimulus might contain unique features not present in the other stimuli.) To address this issue, Elliffe, Rolls and Stringer (2002) used stimuli that are composed from a set of four features which are designed so that each feature is spatially separate from the other features, and no unique combination of firing caused for example by overlap of horizontal and vertical filter outputs in the input representation distinguishes any one stimulus from the others. The results described in Section 4.5.5.4 show that VisNet can indeed learn correct invariant representations of stimuli which do consist of feature sets where individual features do not overlap spatially with each other and where the stimuli can be composed of sets of features which are supersets or subsets of those in other stimuli. Fukushima and Miyake (1982) did not address this crucial issue where different stimuli might be composed of subsets or supersets of the same set of features, although they did show that stimuli with partly overlapping features could be discriminated by the Neocognitron.

In Section 4.5.5.5 we address the spatial binding problem in architectures such as VisNet. This computational problem that needs to be addressed in hierarchical networks such as the primate visual system and VisNet is how representations of features can be (e.g. translation) invariant, yet can specify stimuli or objects in which the features must be specified in the correct spatial arrangement. This is the feature binding problem, discussed for example by von der Malsburg (1990), and arising in the context of hierarchical layered systems (Ackley, Hinton and Sejnowski 1985, Fukushima 1980, Rosenblatt 1961). The issue is whether or not features are bound into the correct combinations in the correct relative spatial positions, or if alternative combinations of known features or the same features in different relative spatial positions would elicit the same responses. All this has to be achieved while at the same time producing position invariant recognition of the whole combination of features, that is, the object. This is a major computational issue that needs to be solved for memory systems in the brain to operate correctly. This can be achieved by what is effectively a learning process that builds into the system a set of neurons in the hierarchical network that enables the recognition process to operate correctly with the appropriate position, size, view etc. invariances.
4.5.5.1 Syntactic binding of separate neuronal ensembles by synchronization

The problem of syntactic binding of neuronal representations, in which some features must be bound together to form one object, and other simultaneously active features must be bound together to represent another object, has been addressed by von der Malsburg (see von der Malsburg (1990)). He has proposed that this could be performed by temporal synchronization of those neurons that were temporarily part of one representation in a different time slot from other neurons that were temporarily part of another representation. The idea is attractive in allowing arbitrary relinking of features in different combinations. Singer, Engel, Konig, and colleagues (Singer, Gray, Engel, Konig, Artola and Brocher 1990, Engel, Konig, Kreiter, Schillen and Singer 1992, Singer and Gray 1995, Singer 1999), and others (Abeles 1991) have obtained some evidence that when features must be bound, synchronization of neuronal populations can occur (but see Shadlen and Movshon (1999)), and this has been modelled (Hummel and Biederman 1992).

Synchronization to implement syntactic binding has a number of disadvantages and limitations (see also Rolls and Treves (1998), Riesenhuber and Poggio (1999a) and Rolls and Deco (2002)). The greatest computational problem is that synchronization does not by itself define the spatial relations between the features being bound, so is not just as a binding mechanism adequate for shape recognition. For example, temporal binding might enable features 1, 2 and 3, which might define one stimulus to be bound together and kept separate from for example another stimulus consisting of features 2, 3 and 4, but would require a further temporal binding (leading in the end potentially to a combinatorial explosion) to indicate the relative spatial positions of the 1, 2 and 3 in the 123 stimulus, so that it can be discriminated from e.g. 312.

A second problem with the synchronization approach to the spatial binding of features is that, when stimulus-dependent temporal synchronization has been rigourously tested with information theoretic approaches, it has so far been found that most of the information available is in the number of spikes, with rather little, less than 5% of the total information, in stimulus-dependent synchronization (Aggelopoulos, Franco and Rolls 2005, Franco, Rolls, Aggelopoulos and Treves 2004, Rolls, Aggelopoulos, Franco and Treves 2004) (see Section C.3.7). For example, Aggelopoulos, Franco and Rolls (2005) showed that when macaques used object-based attention to search for one of two objects to touch in a complex natural scene, between 99% and 94% of the information was present in the firing rates of inferior temporal cortex neurons, and less that 5% in any stimulus-dependent synchrony that was present between the simultaneously recorded inferior temporal cortex neurons. The implication of these results is that any stimulus-dependent synchrony that is present is not quantitatively important as measured by information theoretic analyses under natural scene conditions when feature binding, segmentation of objects from the background, and attention are required. This has been found for the inferior temporal cortex, a brain region where features are put together to form representations of objects (Rolls and Deco 2002), and where attention has strong effects, at least in scenes with blank backgrounds (Rolls, Aggelopoulos and Zheng 2003a). It would of course also be of interest to test the same hypothesis in earlier visual areas, such as V4, with quantitative, information theoretic, techniques. In connection with rate codes, it should be noted that a rate code implies using the number of spikes that arrive in a given time, and that this time can be very short, as little as 20–50 ms, for very useful amounts of information to be made available from a population of neurons (Tovee, Rolls, Treves and Bellis 1993, Rolls and Tovee 1994, Rolls, Tovee, Purcell, Stewart and Azzopardi 1994b, Tovee and Rolls 1995, Rolls, Tovee and Panzeri 1999b, Rolls 2003, Rolls, Franco, Aggelopoulos and Jerez 2006b) (see Section C.3.4).
In the context of VisNet, and how the real visual system may operate to implement object recognition, the use of synchronization does not appear to match the way in which the visual system is organized. For example, von der Malsburg’s argument would indicate that, using only a two-layer network, synchronization could provide the necessary feature linking to perform object recognition with relatively few neurons, because they can be reused again and again, linked differently for different objects. In contrast, the primate uses a considerable part of its cortex, perhaps 50% in monkeys, for visual processing, with therefore what could be in the order of $6 \times 10^9$ neurons and $6 \times 10^{12}$ synapses involved (estimating from the values given in Table 1.1), so that the solution adopted by the real visual system may be one which relies on many neurons with simpler processing than arbitrary syntax implemented by synchronous firing of separate assemblies suggests. On the other hand, a solution such as that investigated by VisNet, which forms low-order combinations of what is represented in previous layers, is very demanding in terms of the number of neurons required, and this matches what is found in the primate visual system. It will be fascinating to see how research on these different approaches to processing in the primate visual system develops. For the development of both approaches, the use of well-defined neuronal network models is proving to be very helpful.

4.5.5.2 Sigma-Pi neurons

Another approach to a binding mechanism is to group spatial features based on local mechanisms that might operate for closely adjacent synapses on a dendrite (in what is a Sigma-Pi type of neuron, see Sections 4.6 and A.2.3) (Finkel and Edelman 1987, Mel, Ruderman and Archie 1998). A problem for such architectures is how to force one particular neuron to respond to the same feature combination invariantly with respect to all the ways in which that feature combination might occur in a scene.

4.5.5.3 Binding of features and their relative spatial position by feature combination neurons

The approach to the spatial binding problem that is proposed for VisNet is that individual neurons at an early stage of processing are set up (by learning) to respond to low order combinations of input features occurring in a given relative spatial arrangement and position on the retina (Rolls 1992a, Rolls 1994a, Rolls 1995b, Wallis and Rolls 1997, Rolls and Treves 1998, Elliffe, Rolls and Stringer 2002, Rolls and Deco 2002) (cf. Feldman (1985)). (By low order combinations of input features we mean combinations of a few input features. By forming neurons that respond to combinations of a few features in the correct spatial arrangement the advantages of the scheme for syntactic binding are obtained, yet without the combinatorial explosion that would result if the feature combination neurons responded to combinations of many input features so producing potentially very specifically tuned neurons which very rarely responded.) Then invariant representations are developed in the next layer from these feature combination neurons which already contain evidence on the local spatial arrangement of features. Finally, in later layers, only one stimulus would be specified by the particular set of low order feature combination neurons present, even though each feature combination neuron would itself be somewhat invariant. The overall design of the scheme is shown in Fig. 4.20. Evidence that many neurons in V1 respond to combinations of spatial features with the correct spatial configuration is now starting to appear (see Section 4.4), and neurons that respond to feature combinations (such as two lines with a defined angle between them, and overall orientation) are found in V2 (Hegde and Van Essen 2000, Ito and Komatsu 2004). The tuning of a VisNet layer 1 neuron to a combination of features in the correct relative spatial position is illustrated in Figs. 4.25 and 4.26.
Discrimination between stimuli with super- and sub-set feature combinations

Some investigations with VisNet (Wallis and Rolls 1997) have involved groups of stimuli that might be identified by some unique feature common to all transformations of a particular stimulus. This might allow VisNet to solve the problem of transform invariance by simply learning to respond to a unique feature present in each stimulus. For example, even in the case where VisNet was trained on invariant discrimination of T, L, and +, the representation of the T stimulus at the spatial filter level inputs to VisNet might contain unique patterns of filter outputs where the horizontal and vertical parts of the T join. The unique filter outputs thus formed might distinguish the T from for example the L.

Elliffe, Rolls and Stringer (2002) tested whether VisNet is able to form transform invariant cells with stimuli that are specially composed from a common alphabet of features, with no stimulus containing any firing in the spatial filter inputs to VisNet not present in at least one of the other stimuli. The limited alphabet enables the set of stimuli to consist of feature sets which are subsets or supersets of those in the other stimuli.

For these experiments the common pool of stimulus features chosen was a set of two horizontal and two vertical bars, each aligned with the sides of a 32 × 32 square. The stimuli can be constructed by arbitrary combination of these base level features. We note that effectively the stimulus set consists of four features, a top bar (T), a bottom bar (B), a left bar (L), and a right bar (R). Figure 4.44 shows the complete set used, containing every possible image feature combination. (Note that the two double-feature combinations where the features are parallel to each other are not included, in the interests of retaining symmetry and equal inter-object overlap within each feature-combination level.) Subsequent discussion will group these objects by the number of features each contains: single-; double-; triple-; and quadruple-feature objects correspond to the respective rows of Fig. 4.44. Stimuli are referred to by the list of features they contain; e.g. ‘LBR’ contains the left, bottom, and right features, while ‘TL’ contains top and left only. Further details of how the stimuli were prepared are provided by Elliffe, Rolls and Stringer (2002).

To train the network a stimulus was presented in a randomized sequence of nine locations in a square grid across the 128 × 128 input retina. The central location of the square grid was in the centre of the ‘retina’, and the eight other locations were offset 8 pixels horizontally and/or vertically from this. Two different learning rules were used, ‘Hebbian’ (4.12), and ‘trace’ (4.13), and also an untrained condition with random weights. As in earlier work (Wallis and Rolls 1997, Rolls and Milward 2000) only the trace rule led to any cells with invariant responses, and the results shown here are for networks trained with the trace rule.

The results with VisNet trained on the set of stimuli shown in Fig. 4.44 with the trace rule are as follows. Firstly, it was found that single neurons in the top layer learned to differentiate between the stimuli in that the responses of individual neurons were maximal for one of the stimuli and had no response to any of the other stimuli invariantly with respect to location. Secondly, to assess how well every stimulus was encoded for in this way, Fig. 4.45 shows the information available about each of the stimuli consisting of feature singles, feature pairs, feature triples, and the quadruple-feature stimulus ‘TLBR’. The single cell information available from the 26–85 cells with best tuning to each of the stimuli is shown. The cells in general conveyed translation invariant information about the stimulus to which they responded, with indeed cells which perfectly discriminated one of the stimuli from all others over every testing position (for all stimuli except ‘RTL’ and ‘TLBR’).

The results presented show clearly that the VisNet paradigm can accommodate networks that can perform invariant discrimination of objects which have a subset-superset relationship. The result has important consequences for feature binding and for discriminating stimuli for
Fig. 4.44 Merged feature objects. All members of the full object set are shown, using a dotted line to represent the central $32 \times 32$ square on which the individual features are positioned, with the features themselves shown as dark line segments. Nomenclature is by acronym of the features present. (After Elliffe, Rolls and Stringer 2002.)

other stimuli which may be supersets of the first stimulus. For example, a VisNet cell which responds invariantly to feature combination TL can genuinely signal the presence of exactly that combination, and will not necessarily be activated by T alone, or by TLB. The basis for this separation by competitive networks of stimuli which are subsets and supersets of each other is described in Section B.4, and by Rolls and Treves (1998, Section 4.3.6).

4.5.5.5 Feature binding in a hierarchical network with invariant representations of local feature combinations

In this section we consider the ability of output layer neurons to learn new stimuli if the lower layers are trained solely through exposure to simpler feature combinations from which the new stimuli are composed. A key question we address is how invariant representations of low order feature combinations in the early layers of the visual system are able to uniquely specify the correct spatial arrangement of features in the overall stimulus and contribute to preventing false recognition errors in the output layer.

The problem, and its proposed solution, can be treated as follows. Consider an object 1234 made from the features 1, 2, 3 and 4. The invariant low order feature combinations might represent 12, 23, and 34. Then if neurons at the next layer respond to combinations of the activity of these neurons, the only neurons in the next layer that would respond would be those tuned to 1234, not to for example 3412, which is distinguished from 1234 by the input of a pair neuron responding to 41 rather than to 23. The argument (Rolls 1992a) is that low-order spatial feature combination neurons in the early stage contain sufficient spatial information...
so that a particular combination of those low-order feature combination neurons specifies a unique object, even if the relative positions of the low-order feature combination neurons are not known, because they are somewhat invariant.

The architecture of VisNet is intended to solve this problem partly by allowing high spatial precision combinations of input features to be formed in layer 1. The actual input features in VisNet are, as described above, the output of oriented spatial-frequency tuned filters, and the combinations of these formed in layer 1 might thus be thought of in a simple way as for example a T or an L or for that matter a Y. Then in layer 2, application of the trace rule might enable neurons to respond to a T with limited spatial invariance (limited to the size of the region of layer 1 from which layer 2 cells receive their input). Then an ‘object’ such as H might be formed at a higher layer because of a conjunction of two Ts in the same small region.

To show that VisNet can actually solve this problem, Elliffe, Rolls and Stringer (2002) performed the experiments described next. They trained the first two layers of VisNet with feature pair combinations, forming representations of feature pairs with some translation invariance in layer 2. Then they used feature triples as input stimuli, allowed no more learning in layers 1 and 2, and then investigated whether layers 3 and 4 could be trained to produce invariant representations of the triples where the triples could only be distinguished if the local spatial arrangement of the features within the triple had effectively to be encoded in order to distinguish the different triples. For this experiment, they needed stimuli that could be specified in terms of a set of different features (they chose vertical (1), diagonal (2), and horizontal (3) bars) each capable of being shown at a set of different relative spatial positions...
Fig. 4.46 Feature combinations for experiments of Section 4.5.5.5: there are 3 features denoted by 1, 2 and 3 (including a blank space 0) that can be placed in any of 3 positions A, B, and C. Individual stimuli are denoted by three consecutive numbers which refer to the individual features present in positions A, B and C respectively. In the experiments in Section 4.5.5.5, layers 1 and 2 were trained on stimuli consisting of pairs of the features, and layers 3 and 4 were trained on stimuli consisting of triples. Then the network was tested to show whether layer 4 neurons would distinguish between triples, even though the first two layers had only been trained on pairs. In addition, the network was tested to show whether individual cells in layer 4 could distinguish between triples even in locations where the triples were not presented during training. (After Elliffe, Rolls and Stringer 2002.)
Table 4.5 The different training regimes used in VisNet experiments 1–4 of Section 4.5.5.5. In the no training condition the synaptic weights were left in their initial untrained random values.

<table>
<thead>
<tr>
<th></th>
<th>Layers 1, 2</th>
<th>Layers 3, 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>trained on pairs</td>
<td>trained on triples</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>no training</td>
<td>no training</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>no training</td>
<td>trained on triples</td>
</tr>
<tr>
<td>Experiment 4</td>
<td>trained on triples</td>
<td>trained on triples</td>
</tr>
</tbody>
</table>

Fig. 4.47 Numerical results for experiments 1–4 as described in Table 4.5, with the trace learning rule (4.13). On the left are single cell information measures, and on the right are multiple cell information measures. (After Elliffe, Rolls and Stringer 2002.)

Positions of the low-order feature combination neurons are not known because these neurons are somewhat translation invariant (cf. also Fukushima (1988)).

The stimuli used in the experiments of Elliffe, Rolls and Stringer (2002) were constructed from pre-processed component features as discussed in Section 4.5.5.4. That is, base stimuli containing a single feature were constructed and filtered, and then the pairs and triples were constructed by merging these pre-processed single feature images. In the first experiment layers 1 and 2 of VisNet were trained with the 18 feature pairs, each stimulus being presented in sequences of 9 locations across the input. This led to the formation of neurons that responded to the feature pairs with some translation invariance in layer 2. Then they trained layers 3 and 4 on the 6 feature triples in the same 9 locations, while allowing no more learning in layers 1 and 2, and examined whether the output layer of VisNet had developed transform invariant neurons to the 6 triples. The idea was to test whether layers 3 and 4 could be trained to produce invariant representations of the triples where the triples could only be distinguished if the local spatial arrangement of the features within the triple had effectively to be encoded in order to distinguish the different triples. The results from this experiment were compared and contrasted with results from three other experiments which involved different training regimes for layers 1, 2 and layers 3, 4. All four experiments are summarized in Table 4.5. Experiment 2 involved no training in layers 1, 2 and 3, 4, with the synaptic weights left unchanged from their initial random values. These results are included as a baseline performance with which to compare results from the other experiments 1, 3 and 4. The model parameters used in these experiments were as described by Rolls and Milward (2000) and Rolls and Stringer (2001a).

In Fig. 4.47 we present numerical results for the four experiments listed in Table 4.5. On the left are the single cell information measures for all top (4th) layer neurons ranked in order of their invariance to the triples, while on the right are multiple cell information measures. To help to interpret these results we can compute the maximum single cell information measure
according to

\[ \text{Maximum single cell information} = \log_2(\text{Number of triples}), \]  

(4.38)

where the number of triples is 6. This gives a maximum single cell information measure of 2.6 bits for these test cases. First, comparing the results for experiment 1 with the baseline performance of experiment 2 (no training) demonstrates that even with the first two layers trained to form invariant responses to the pairs, and then only layers 3 and 4 trained on feature triples, layer 4 is indeed capable of developing translation invariant neurons that can discriminate effectively between the 6 different feature triples. Indeed, from the single cell information measures it can be seen that a number of cells have reached the maximum level of performance in experiment 1. In addition, the multiple cell information analysis presented in Fig. 4.47 shows that all the stimuli could be discriminated from each other by the firing of a number of cells. Analysis of the response profiles of individual cells showed that a fourth layer cell could respond to one of the triple feature stimuli and have no response to any other of the triple feature stimuli invariantly with respect to location.

A comparison of the results from experiment 1 with those from experiment 3 (see Table 4.5 and Fig. 4.47) reveals that training the first two layers to develop neurons that respond invariantly to the pairs (performed in experiment 1) actually leads to improved invariance of 4th layer neurons to the triples, as compared with when the first two layers are left untrained (experiment 3).

Two conclusions follow from these results (Elliffe, Rolls and Stringer 2002). First, a hierarchical network that seeks to produce invariant representations in the way used by VisNet can solve the feature binding problem. In particular, when feature pairs in layer 2 with some translation invariance are used as the input to later layers, these later layers can nevertheless build invariant representations of objects where all the individual features in the stimulus must occur in the correct spatial position relative to each other. This is possible because the feature combination neurons formed in the first layer (which could be trained just with a Hebb rule) do respond to combinations of input features in the correct spatial configuration, partly because of the limited size of their receptive fields (see e.g. Fig. 4.23). The second conclusion is that even though early layers can in this case only respond to small feature subsets, these provide, with no further training of layers 1 and 2, an adequate basis for learning to discriminate in layers 3 and 4 stimuli consisting of combinations of larger numbers of features. Indeed, comparing results from experiment 1 with experiment 4 (in which all layers were trained on triples, see Table 4.5) demonstrates that training the lower layer neurons to develop invariant responses to the pairs offers almost as good performance as training all layers on the triples (see Fig. 4.47).

### 4.5.5.6 Stimulus generalization to new locations

Another important aspect of the architecture of VisNet is that it need not be trained with every stimulus in every possible location. Indeed, part of the hypothesis (Rolls 1992a) is that training early layers (e.g. 1–3) with a wide range of visual stimuli will set up feature analyzers in these early layers which are appropriate later on with no further training of early layers for new objects. For example, presentation of a new object might result in large numbers of low order feature combination neurons in early layers of VisNet being active, but the particular set of feature combination neurons active would be different for the new object. The later layers of the network (in VisNet, layer 4) would then learn this new set of active layer 3 neurons as encoding the new object. However, if the new object was then shown in a new location, the same set of layer 3 neurons would be active because they respond with spatial invariance to feature combinations, and given that the layer 3 to 4 connections had already been set up by
Fig. 4.48 Generalization to new locations: numerical results for a repeat of experiment 1 of Section 4.5.5.5 with the triples presented at only 7 of the original 9 locations during training, and with the trace learning rule (4.13). On the left are single cell information measures, and on the right are multiple cell information measures.

the new object, the correct layer 4 neurons would be activated by the new object in its new untrained location, and without any further training.

To test this hypothesis, Elliffe, Rolls and Stringer (2002) repeated the general procedure of experiment 1 of Section 4.5.5.5, training layers 1 and 2 with feature pairs, but then instead trained layers 3 and 4 on the triples in only 7 of the original 9 locations. The crucial test was to determine whether VisNet could form top layer neurons that responded invariantly to the 6 triples when presented over all nine locations, not just the seven locations at which the triples had been presented during training. The results are presented in Fig. 4.48, with single cell information measures on the left and multiple cell information measures on the right. VisNet is still able to develop some fourth layer neurons with perfect invariance, that is which have invariant responses over all nine location, as shown by the single cell information analysis. The response profiles of individual fourth layer cells showed that they can continue to discriminate between the triples even in the two locations where the triples were not presented during training. In addition, the multiple cell analysis shown in Fig. 4.48 demonstrates that a small population of cells was able to discriminate between all of the stimuli irrespective of location, even though for two of the test locations the triples had not been trained at those particular locations during the training of layers 3 and 4.

4.5.5.7 Discussion of feature binding in hierarchical layered networks

Elliffe, Rolls and Stringer (2002) thus first showed (see Section 4.5.5.4) that hierarchical feature detecting neural networks can learn to respond differently to stimuli that consist of unique combinations of non-unique input features, and that this extends to stimuli that are direct subsets or supersets of the features present in other stimuli.

Second, Elliffe, Rolls and Stringer (2002) investigated (see Section 4.5.5.5) the hypothesis that hierarchical layered networks can produce identification of unique stimuli even when the feature combination neurons used to define the stimuli are themselves partly translation invariant. The stimulus identification should work correctly because feature combination neurons in which the spatial features are bound together with high spatial precision are formed in the first layer. Then at later layers when neurons with some translation invariance are formed, the neurons nevertheless contain information about the relative spatial position of the original features. There is only then one object which will be consistent with the set of active neurons at earlier layers, which though somewhat translation invariant as combination neurons, reflect in the activity of each neuron information about the original spatial position of the features. We note that the trace rule training used in early layers (1 and 2) in Experiments
1 and 4 would set up partly invariant feature combination neurons, and yet the late layers (3 and 4) were able to produce during training neurons in layer 4 that responded to stimuli that consisted of unique spatial arrangements of lower order feature combinations. Moreover, and very interestingly, Elliffe, Rolls and Stringer (2002) were able to demonstrate that VisNet layer 4 neurons would respond correctly to visual stimuli at untrained locations, provided that the feature subsets had been trained in early layers of the network at all locations, and that the whole stimulus had been trained at some locations in the later layers of the network.

The results described by Elliffe, Rolls and Stringer (2002) thus provide one solution to the feature binding problem. The solution which has been shown to work in the model is that in a multilayer competitive network, feature combination neurons which encode the spatial arrangement of the bound features are formed at intermediate layers of the network. Then neurons at later layers of the network which respond to combinations of active intermediate layer neurons do contain sufficient evidence about the local spatial arrangement of the features to identify stimuli because the local spatial arrangement is encoded by the intermediate layer neurons. The information required to solve the visual feature binding problem thus becomes encoded by self-organization into what become hard-wired properties of the network. In this sense, feature binding is not solved at run time by the necessity to instantaneously set up arbitrary syntactic links between sets of co-active neurons. The computational solution proposed to the superset/subset aspect of the binding problem will apply in principle to other multilayer competitive networks, although the issues considered here have not been explicitly addressed in architectures such as the Neocognitron (Fukushima and Miyake 1982).

Consistent with these hypotheses about how VisNet operates to achieve, by layer 4, position-invariant responses to stimuli defined by combinations of features in the correct spatial arrangement, investigations of the effective stimuli for neurons in intermediate layers of VisNet showed as follows. In layer 1, cells responded to the presence of individual features, or to low order combinations of features (e.g. a pair of features) in the correct spatial arrangement at a small number of nearby locations. In layers 2 and 3, neurons responded to single features or to higher order combinations of features (e.g. stimuli composed of feature triples) in more locations. These findings provide direct evidence that VisNet does operate as described above to solve the feature binding problem.

A further issue with hierarchical multilayer architectures such as VisNet is that false binding errors might occur in the following way (Mozer 1991, Mel and Fiser 2000). Consider the output of one layer in such a network in which there is information only about which pairs are present. How then could a neuron in the next layer discriminate between the whole stimulus (such as the triple 123 in the above experiment) and what could be considered a more distributed stimulus or multiple different stimuli composed of the separated subparts of that stimulus (e.g. the pairs 120, 023, 103 occurring in 3 of the 9 training locations in the above experiment)? The problem here is to distinguish a single object from multiple other objects containing the same component combinations (e.g. pairs). We propose that part of the solution to this general problem in real visual systems is implemented through lateral inhibition between neurons in individual layers, and that this mechanism, implemented in VisNet, acts to reduce the possibility of false recognition errors in the following two ways.

First, consider the situation in which neurons in layer \( N \) have learned to represent low order feature combinations with location invariance, and where a neuron \( n \) in layer \( N + 1 \) has learned to respond to a particular set \( \Omega \) of these feature combinations. The problem is that neuron \( n \) receives the same input from layer \( N \) as long as the same set \( \Omega \) of feature combinations is present, and cannot distinguish between different spatial arrangements of these feature combinations. The question is how can neuron \( n \) respond only to a particular favoured spatial arrangement \( \Psi \) of the feature combinations contained within the set \( \Omega \). We suggest that as the favoured spatial arrangement \( \Psi \) is altered by rearranging the spatial
relationships of the component feature combinations, the new feature combinations that are formed in new locations will stimulate additional neurons nearby in layer $N+1$, and these will tend to inhibit the firing of neuron $n$. Thus, lateral inhibition within a layer will have the effect of making neurons more selective, ensuring neuron $n$ responds only to a single spatial arrangement $\Psi$ from the set of feature combinations $\Omega$, and hence reducing the possibility of false recognition.

The second way in which lateral inhibition may help to reduce binding errors is through limiting the sparseness of neuronal firing rates within layers. In our discussion above the spurious stimuli we suggested that might lead to false recognition of triples were obtained from splitting up the component feature combinations (pairs) so that they occurred in separate training locations. However, this would lead to an increase in the number of features present in the complete stimulus; triples contain 3 features while their spurious counterparts would contain 6 features (resulting from 3 separate pairs). For this trivial example, the increase in the number of features is not dramatic, but if we consider, say, stimuli composed of 4 features where the component feature combinations represented by lower layers might be triples, then to form spurious stimuli we need to use 12 features (resulting from 4 triples occurring in separate locations). But if the lower layers also represented all possible pairs then the number of features required in the spurious stimuli would increase further. In fact, as the size of the stimulus increases in terms of the number of features, and as the size of the component feature combinations represented by the lower layers increases, there is a combinatorial explosion in terms of the number of features required as we attempt to construct spurious stimuli to trigger false recognition. And the construction of such spurious stimuli will then be prevented through setting a limit on the sparseness of firing rates within layers, which will in turn set a limit on the number of features that can be represented. Lateral inhibition is likely to contribute in both these ways to the performance of VisNet when the stimuli consist of subsets and supersets of each other, as described in Section 4.5.5.4.

Another way is which the problem of multiple objects is addressed is by limiting the size of the receptive fields of inferior temporal cortex neurons so that neurons in IT respond primarily to the object being fixated, but with nevertheless some asymmetry in the receptive fields (see Section 4.5.10). Multiple objects are then ‘seen’ by virtue of being added to a visuo-spatial scratchpad, as addressed in Section 4.7.

A related issue that arises in this class of network is whether forming neurons that respond to feature combinations in the way described here leads to a combinatorial explosion in the number of neurons required. The solution to this issue that is proposed is to form only low-order combinations of features at any one stage of the network (Rolls (1992a); cf. Feldman (1985)). Using low-order combinations limits the number of neurons required, yet enables the type of computation that relies on feature combination neurons that is analyzed here to still be performed. The actual number of neurons required depends also on the redundancies present in the statistics of real-world images. Even given these factors, it is likely that a large number of neurons would be required if the ventral visual system performs the computation of invariant representations in the manner captured by the hypotheses implemented in VisNet. Consistent with this, a considerable part of the non-human primate brain is devoted to visual information processing. The fact that large numbers of neurons and a multilayer organization are present in the primate ventral visual system is actually thus consistent with the type of model of visual information processing described here.

### 4.5.6 Operation in a cluttered environment

In this section we consider how hierarchical layered networks of the type exemplified by VisNet operate in cluttered environments. Although there has been much work involving
object recognition in cluttered environments with artificial vision systems, many such systems typically rely on some form of explicit segmentation followed by search and template matching procedure (see Ullman (1996) for a general review). In natural environments, objects may not only appear against cluttered (natural) backgrounds, but also the object may be partially occluded. Biological nervous systems operate in quite a different manner to those artificial vision systems that rely on search and template matching, and the way in which biological systems cope with cluttered environments and partial occlusion is likely to be quite different also.

One of the factors that will influence the performance of the type of architecture considered here, hierarchically organized series of competitive networks, which form one class of approaches to biologically relevant networks for invariant object recognition (Fukushima 1980, Rolls 1992a, Wallis and Rolls 1997, Poggio and Edelman 1990, Rolls and Treves 1998), is how lateral inhibition and competition are managed within a layer. Even if an object is not obscured, the effect of a cluttered background will be to fire additional neurons, which will in turn to some extent compete with and inhibit those neurons that are specifically tuned to respond to the desired object. Moreover, where the clutter is adjacent to part of the object, the feature analysing neurons activated against a blank background might be different from those activated against a cluttered background, if there is no explicit segmentation process. We consider these issues next, following investigations of Stringer and Rolls (2000).

4.5.6.1 VisNet simulations with stimuli in cluttered backgrounds

In this section we show that recognition of objects learned previously against a blank background is hardly affected by the presence of a natural cluttered background. We go on to consider what happens when VisNet is set the task of learning new stimuli presented against cluttered backgrounds.

The images used for training and testing VisNet in the simulations described next performed by Stringer and Rolls (2000) were specially constructed. There were 7 face stimuli approximately 64 pixels in height constructed without backgrounds from those shown in Fig. 4.33. In addition there were 3 possible backgrounds: a blank background (greyscale 127, where the range is 0–255), and two cluttered backgrounds as shown in Fig. 4.49 which are 128 × 128 pixels in size. Each image presented to VisNet’s 128 × 128 input retina was composed of a single face stimulus positioned at one of 9 locations on either a blank or cluttered
Computational issues in feature hierarchies

Fig. 4.50 Numerical results for experiment 2, with the 7 faces presented on a blank background during training and a cluttered background during testing. On the left are single cell information measures, and on the right are multiple cell information measures.

The cluttered background was intended to be like the background against which an object might be viewed in a natural scene. If a background is used in an experiment described here, the same background is always used, and it is always in the same position, with stimuli moved to different positions on it. The 9 stimulus locations are arranged in a square grid across the background, where the grid spacings are 32 pixels horizontally or vertically. Before images were presented to VisNet’s input layer they were pre-processed by the standard set of input filters which accord with the general tuning profiles of simple cells in V1 (Hawken and Parker 1987); full details are given in Rolls and Milward (2000). To train the network a sequence of images is presented to VisNet’s retina that corresponds to a single stimulus occurring in a randomized sequence of the 9 locations across a background. At each presentation the activation of individual neurons is calculated, then their firing rates are calculated, and then the synaptic weights are updated. After a stimulus has been presented in all the training locations, a new stimulus is chosen at random and the process repeated. The presentation of all the stimuli across all locations constitutes 1 epoch of training. In this manner the network is trained one layer at a time starting with layer 1 and finishing with layer 4. In the investigations described in this subsection, the numbers of training epochs for layers 1–4 were 50, 100, 100 and 75 respectively.

In this experiment (see Stringer and Rolls (2000), experiment 2), VisNet was trained with the 7 face stimuli presented on a blank background, but tested with the faces presented on each of the 2 cluttered backgrounds. Figure 4.50 shows results for experiment 2, with single and multiple cell information measures on the left and right respectively. It can be seen that a number of cells have reached the maximum possible single cell information measure of 2.8 bits (log₂ of the number of stimuli) for this test case, and that the multiple cell information measures also reach the 2.8 bits indicating perfect performance. Compared to performance when shown against a blank background, there was very little deterioration in performance when testing with the faces presented on either of the two cluttered backgrounds. This is an interesting result to compare with many artificial vision systems that would need to carry out computationally intensive serial searching and template matching procedures in order to achieve such results. In contrast, the VisNet neural network architecture is able to perform such recognition relatively quickly through a simple feedforward computation. Further results from this experiment are presented in Fig. 4.51 where we show the response profiles of a 4th layer neuron to the 7 faces presented on cluttered background 1 during testing. It can be seen that this neuron achieves excellent invariant responses to the 7 faces even with the faces presented on a cluttered background. The response profiles are independent of location but differentiate between the faces in that the responses are maximal for only one of the faces and
minimal for all other faces.

This is an interesting and important result, for it shows that after learning, special mechanisms for segmentation and for attention are not needed in order for neurons already tuned by previous learning to the stimuli to be activated correctly in the output layer. Although the experiments described here tested for position invariance, we predict and would expect that the same results would be demonstrable for size and view invariant representations of objects.

In experiments 3 and 4 of Stringer and Rolls (2000), VisNet was trained with the 7 face stimuli presented on either one of the 2 cluttered backgrounds, but tested with the faces presented on a blank background. Results for this experiment showed poor performance. The results of experiments 3 and 4 suggest that in order for a cell to learn invariant responses to different transforms of a stimulus when it is presented during training in a cluttered background, some form of segmentation is required in order to separate the figure (i.e. the stimulus or object) from the background. This segmentation might be performed using evidence in the visual scene about different depths, motions, colours, etc. of the object from its background. In the visual system, this might mean combining evidence represented in different cortical areas, and might be performed by cross-connections between cortical areas to enable such evidence to help separate the representations of objects from their backgrounds in the form-representing cortical areas.

Another mechanism that helps the operation of architectures such as VisNet and the primate visual system to learn about new objects in cluttered scenes is that the receptive fields of inferior temporal cortex neurons become much smaller when objects are seen against natural backgrounds (Sections 4.5.9.1 and 4.5.9). This will help greatly to learn about new objects that are being fixated, by reducing responsiveness to other features elsewhere in the scene.

Another mechanism that might help the learning of new objects in a natural scene is attention. An attentional mechanism might highlight the current stimulus being attended to and suppress the effects of background noise, providing a training representation of the object more like that which would be produced when it is presented against a blank background. The mechanisms that could implement such attentional processes are described in Chapter 6. If such attentional mechanisms do contribute to the development of view invariance, then it follows that cells in the temporal cortex may only develop transform invariant responses to objects to which attention is directed.

Part of the reason for the poor performance in experiments 3 and 4 was probably that the stimuli were always presented against the same fixed background (for technical reasons), and thus the neurons learned about the background rather than the stimuli. Part of the difficulty that hierarchical multilayer competitive networks have with learning in cluttered environments may
Computational issues in feature hierarchies

more generally be that without explicit segmentation of the stimulus from its background, at least some of the features that should be formed to encode the stimuli are not formed properly, because the neurons learn to respond to combinations of inputs which come partly from the stimulus, and partly from the background. To investigate this, Stringer and Rolls (2000) performed experiment 5 in which layers 1–3 were pretrained with stimuli to ensure that good feature combination neurons for stimuli were available, and then allowed learning in only layer 4 when stimuli were presented in the cluttered backgrounds. Layer 4 was then trained in the usual way with the 7 faces presented against a cluttered background. The results for this experiment are shown in Fig. 4.52, with single and multiple cell information measures on the left and right respectively. It was found that prior random exposure to the face stimuli led to much improved performance. Indeed, it can be seen that a number of cells have reached the maximum possible single cell information measure of 2.8 bits for this test case, although the multiple cell information measures do not quite reach the 2.8 bits that would indicate perfect performance for the complete face set.

These results demonstrate that the problem of developing position invariant neurons to stimuli occurring against cluttered backgrounds may be ameliorated by the prior existence of stimulus-tuned feature-detecting neurons in the early layers of the visual system, and that these feature-detecting neurons may be set up through previous exposure to the relevant class of objects. When tested in cluttered environments, the background clutter may of course activate some other neurons in the output layer, but at least the neurons that have learned to respond to the trained stimuli are activated. The result of this activity is sufficient for the activity in the output layer to be useful, in the sense that it can be read off correctly by a pattern associator connected to the output layer. Indeed, Stringer and Rolls (2000) tested this by connecting a pattern associator to layer 4 of VisNet. The pattern associator had seven neurons, one for each face, and 1,024 inputs, one from each neuron in layer 4 of VisNet. The pattern associator learned when trained with a simple associative Hebb rule (equation 4.12 on page 316) to activate the correct output neuron whenever one of the faces was shown in any position in the uncluttered environment. This ability was shown to be dependent on invariant neurons for each stimulus in the output layer of VisNet, for the pattern associator could not be taught the task if VisNet had not been previously trained with a trace learning rule to produce...
Invariant representations. Then it was shown that exactly the correct neuron was activated when any of the faces was shown in any position with the cluttered background. This read-off by a pattern associator is exactly what we hypothesize takes place in the brain, in that the inferior temporal visual cortex (where neurons with invariant responses are found) projects to structures such as the orbitofrontal cortex and amygdala, where associations between the invariant visual representations and stimuli such as taste and touch are learned (Rolls and Treves 1998, Rolls 1999a, Rolls 2005) (see Chapter 3). Thus testing whether the output of an architecture such as VisNet can be used effectively by a pattern associator is a very biologically relevant way to evaluate the performance of this class of architecture.

4.5.6.2 Learning invariant representations of an object with multiple objects in the scene and with cluttered backgrounds

The results of the experiments just described suggest that in order for a neuron to learn invariant responses to different transforms of a stimulus when it is presented during training in a cluttered background, some form of segmentation is required in order to separate the figure (i.e. the stimulus or object) from the background. This segmentation might be performed using evidence in the visual scene about different depths, motions, colours, etc. of the object from its background. In the visual system, this might mean combining evidence represented in different cortical areas, and might be performed by cross-connections between cortical areas to enable such evidence to help separate the representations of objects from their backgrounds in the form-representing cortical areas.

A second way in which training a feature hierarchy network in a cluttered natural scene may be facilitated follows from the finding that the receptive fields of inferior temporal cortex neurons shrink from in the order of 70 degrees in diameter when only one object is present in a blank scene to much smaller values of as little as 5–10 degrees close to the fovea in complex natural scenes (Rolls, Aggelopoulos and Zheng 2003a). The proposed mechanism for this is that if there is an object at the fovea, this object, because of the high cortical magnification factor at the fovea, dominates the activity of neurons in the inferior temporal cortex by competitive interactions (Trappenberg, Rolls and Stringer 2002, Deco and Rolls 2004) (see Section 4.5.9). This allows primarily the object at the fovea to be represented in the inferior temporal cortex, and, it is proposed, for learning to be about this object, and not about the other objects in a whole scene.

Third, top-down spatial attention (Deco and Rolls 2004, Deco and Rolls 2005a) (see Chapter 6) could bias the competition towards a region of visual space where the object to be learned is located.

Fourth, if object 1 is presented during training with different objects present on different trials, then the competitive networks that are part of VisNet will learn to represent each object separately, because the features that are part of each object will be much more strongly associated together, than are those features with the other features present in the different objects seen on some trials during training (Stringer and Rolls 2007b, Stringer, Rolls and Tromans 2007b). It is a natural property of competitive networks that input features that co-occur very frequently together are allocated output neurons to represent the pattern as a result of the learning. Input features that do not co-occur frequently, may not have output neurons allocated to them. This principle may help feature hierarchy systems to learn representations of individual objects, even when other objects with some of the same features are present in the visual scene, but with different other objects on different trials. With this fundamental and interesting property of competitive networks, it has now become possible for VisNet to self-organize invariant representations of individual objects, even though each object is always presented during training with at least one other object present in the scene (Stringer and Rolls 2007b, Stringer, Rolls and Tromans 2007b).
4.5.6.3 VisNet simulations with partially occluded stimuli

In this section we examine the recognition of partially occluded stimuli. Many artificial vision systems that perform object recognition typically search for specific markers in stimuli, and hence their performance may become fragile if key parts of a stimulus are occluded. However, in contrast we demonstrate that the model of invariance learning in the brain discussed here can continue to offer robust performance with this kind of problem, and that the model is able to correctly identify stimuli with considerable flexibility about what part of a stimulus is visible.

In these simulations (Stringer and Rolls 2000), training and testing was performed with a blank background to avoid confounding the two separate problems of occlusion and background clutter. In object recognition tasks, artificial vision systems may typically rely on being able to locate a small number of key markers on a stimulus in order to be able to identify it. This approach can become fragile when a number of these markers become obscured. In contrast, biological vision systems may generalize or complete from a partial input as a result of the use of distributed representations in neural networks, and this could lead to greater robustness in situations of partial occlusion.

In this experiment (6 of Stringer and Rolls (2000)), the network was first trained with the 7 face stimuli without occlusion, but during testing there were two options: either (i) the top halves of all the faces were occluded, or (ii) the bottom halves of all the faces were occluded. Since VisNet was tested with either the top or bottom half of the stimuli no stimulus features were common to the two test options. This ensures that if performance is good with both options, the performance cannot be based on the use of a single feature to identify a stimulus. Results for this experiment are shown in Fig. 4.53, with single and multiple cell information measures on the left and right respectively. When compared with the performance without occlusion (Stringer and Rolls 2000), Fig. 4.53 shows that there is only a modest drop in performance in the single cell information measures when the stimuli are partially occluded.

For both options (i) and (ii), even with partially occluded stimuli, a number of cells continue to respond maximally to one preferred stimulus in all locations, while responding minimally to all other stimuli. However, comparing results from options (i) and (ii) shows that the network performance is better when the bottom half of the faces is occluded. This is consistent with psychological results showing that face recognition is performed more easily when the top halves of faces are visible rather than the bottom halves (see Bruce (1988)). The top half of a face will generally contain salient features, e.g. eyes and hair, that are particularly
helpful for recognition of the individual, and it is interesting that these simulations appear to further demonstrate this point. Furthermore, the multiple cell information measures confirm that performance is better with the upper half of the face visible (option (ii)) than the lower half (option (i)). When the top halves of the faces are occluded the multiple cell information measure asymptotes to a suboptimal value reflecting the difficulty of discriminating between these more difficult images. Further results from experiment 6 are presented in Fig. 4.54 where we show the response profiles of a 4th layer neuron to the 7 faces, with the bottom half of all the faces occluded during testing. It can be seen that this neuron continues to respond invariantly to the 7 faces, responding maximally to one of the faces but minimally for all other faces.

Thus this model of the ventral visual system offers robust performance with this kind of problem, and the model is able to correctly identify stimuli with considerable flexibility about what part of a stimulus is visible, because it is effectively using distributed representations and associative processing.

### 4.5.7 Learning 3D transforms

In this section we describe investigations of Stringer and Rolls (2002) which show that trace learning can in the VisNet architecture solve the problem of in-depth rotation invariant object recognition by developing representations of the transforms which features undergo when they are on the surfaces of 3D objects. Moreover, it is shown that having learned how features on 3D objects transform as the object is rotated in depth, the network can correctly recognize novel 3D variations within a generic view of an object which is composed of previously learned feature combinations.

Rolls’ hypothesis of how object recognition could be implemented in the brain postulates that trace rule learning helps invariant representations to form in two ways (Rolls 1992a, Rolls 1994a, Rolls 1995b, Rolls 2000a). The first process enables associations to be learned between different generic 3D views of an object where there are different qualitative shape descriptors. One example of this would be the front and back views of an object, which might have very different shape descriptors. Another example is provided by considering how the shape descriptors typical of 3D shapes, such as Y vertices, arrow vertices, cusps, and ellipse shapes, alter when most 3D objects are rotated in 3 dimensions. At some point in the 3D rotation, there is a catastrophic rearrangement of the shape descriptors as a new generic view can be seen (Koenderink 1990). An example of a catastrophic change to a new generic view is when a cup being viewed from slightly below is rotated so that one can see....
inside the cup from slightly above. The bottom surface disappears, the top surface of the cup changes from a cusp to an ellipse, and the inside of the cup with a whole set of new features comes into view. The second process is that within a generic view, as the object is rotated in depth, there will be no catastrophic changes in the qualitative 3D shape descriptors, but instead the quantitative values of the shape descriptors alter. For example, while the cup is being rotated within a generic view seen from somewhat below, the curvature of the cusp forming the top boundary will alter, but the qualitative shape descriptor will remain a cusp. Trace learning could help with both processes. That is, trace learning could help to associate together qualitatively different sets of shape descriptors that occur close together in time, and describe for example the generically different views of a cup. Trace learning could also help with the second process, and learn to associate together the different quantitative values of shape descriptors that typically occur when objects are rotated within a generic view.

We note that there is evidence that some neurons in the inferior temporal cortex may show the two types of 3D invariance. First, Booth and Rolls (1998) showed that some inferior temporal cortex neurons can respond to different generic views of familiar 3D objects. Second, some neurons do generalize across quantitative changes in the values of 3D shape descriptors while faces (Hasselmo, Rolls, Baylis and Nalwa 1989b) and objects (Tanaka 1996, Logothetis, Pauls and Poggio 1995) are rotated within generic views. Indeed, Logothetis, Pauls and Poggio (1995) showed that a few inferior temporal cortex neurons can generalize to novel (untrained) values of the quantitative shape descriptors typical of within-generic view object rotation.

In addition to the qualitative shape descriptor changes that occur catastrophically between different generic views of an object, and the quantitative changes of 3D shape descriptors that occur within a generic view, there is a third type of transform that must be learned for correct invariant recognition of 3D objects as they rotate in depth. This third type of transform is that which occurs to the surface features on a 3D object as it transforms in depth. The main aim here is to consider mechanisms that could enable neurons to learn this third type of transform, that is how to generalize correctly over the changes in the surface markings on 3D objects that are typically encountered as 3D objects rotate within a generic view. Examples of the types of perspectival transforms investigated are shown in Fig. 4.55. Surface markings on the sphere that consist of combinations of three features in different spatial arrangements undergo characteristic transforms as the sphere is rotated from 0 degrees towards -60 degrees and +60 degrees. We investigated whether the class of architecture exemplified by VisNet, and the trace learning rule, can learn about the transforms that surface features of 3D objects typically undergo during 3D rotation in such a way that the network generalizes across the change of the quantitative values of the surface features produced by the rotation, and yet still discriminates between the different objects (in this case spheres). In the cases being considered, each object is identified by surface markings that consist of a different spatial arrangement of the same three features (a horizontal, vertical, and diagonal line, which become arcs on the surface of the object).

We note that it has been suggested that the finding that neurons may offer some degree of 3D rotation invariance after training with a single view (or limited set of views) represents a challenge for existing trace learning models, because these models assume that an initial exposure is required during learning to every transformation of the object to be recognized (Riesenhuber and Poggio 1998). Stringer and Rolls (2002) showed as described here that this is not the case, and that such models can generalize to novel within-generic views of an object provided that the characteristic changes that the features show as objects are rotated have been learned previously for the sets of features when they are present in different objects.

Elliffe, Rolls and Stringer (2002) demonstrated for a 2D system how the existence of translation invariant representations of low order feature combinations in the early layers of the visual system could allow correct stimulus identification in the output layer even when the
Fig. 4.55 Learning 3D perspectival transforms of features. Representations of the 6 visual stimuli with 3 surface features (triples) presented to VisNet during the simulations described in Section 4.5.7. Each stimulus is a sphere that is uniquely identified by a unique combination of three surface features (a vertical, diagonal and horizontal arc), which occur in 3 relative positions A, B, and C. Each row shows one of the stimuli rotated through the 5 different rotational views in which the stimulus is presented to VisNet. From left to right the rotational views shown are: (i) –60 degrees, (ii) –30 degrees, (iii) 0 degrees (central position), (iv) +30 degrees, and (v) +60 degrees. (After Stringer and Rolls 2002.)

The images used for training and testing VisNet were specially constructed for the purpose of demonstrating how the trace learning paradigm might be further developed to give rise to neurons that are able to respond invariantly to novel within-generic view perspectives of an object, obtained by rotations in-depth up to 30 degrees from any perspectives encountered during learning. The proposal was that the low-order spatial feature combination neurons in the early stages contain sufficient spatial information so that a particular combination of those low-order feature combination neurons specifies a unique object, even if the relative positions of the low-order feature combination neurons are not known because these neurons are somewhat translation invariant (see Section 4.5.5.5). Stringer and Rolls (2002) extended this analysis to feature combinations on 3D objects, and indeed in their simulations described in this section therefore used surface markings for the 3D objects that consisted of triples of features.
during learning. The stimuli take the form of the surface feature combinations of 3-dimensional rotating spheres, with each image presented to VisNet’s retina being a 2-dimensional projection of the surface features of one of the spheres. Each stimulus is uniquely identified by two or three surface features, where the surface features are (1) vertical, (2) diagonal, and (3) horizontal arcs, and where each feature may be centred at three different spatial positions, designated A, B, and C, as shown in Fig. 4.55. The stimuli are thus defined in terms of what features are present and their precise spatial arrangement with respect to each other. We refer to the two and three feature stimuli as ‘pairs’ and ‘triples’, respectively. Individual stimuli are denoted by three numbers which refer to the individual features present in positions A, B and C, respectively. For example, a stimulus with positions A and C containing a vertical and diagonal bar, respectively, would be referred to as stimulus 102, where the 0 denotes no feature present in position B. In total there are 18 pairs (120, 130, 210, 230, 310, 320, 013, 021, 023, 031, 032, 102, 103, 201, 203, 301, 302) and 6 triples (123, 132, 213, 231, 312, 321).

To train the network each stimulus was presented to VisNet in a randomized sequence of five orientations with respect to VisNet’s input retina, where the different orientations are obtained from successive in-depth rotations of the stimulus through 30 degrees. That is, each stimulus was presented to VisNet’s retina from the following rotational views: (i) $-60^\circ$, (ii) $-30^\circ$, (iii) $0^\circ$ (central position with surface features facing directly towards VisNet’s retina), (iv) $30^\circ$, (v) $60^\circ$. Figure 4.55 shows representations of the 6 visual stimuli with 3 surface features (triples) presented to VisNet during the simulations. (For the actual simulations described here, the surface features and their deformations were what VisNet was trained and tested with, and the remaining blank surface of each sphere was set to the same greyscale as the background.) Each row shows one of the stimuli rotated through the 5 different rotational views in which the stimulus is presented to VisNet. At each presentation the activation of individual neurons is calculated, then the neuronal firing rates are calculated, and then the synaptic weights are updated. Each time a stimulus has been presented in all the training orientations, a new stimulus is chosen at random and the process repeated. The presentation of all the stimuli through all 5 orientations constitutes 1 epoch of training. In this manner the network was trained one layer at a time starting with layer 1 and finishing with layer 4. In the investigations described here, the numbers of training epochs for layers 1–4 were 50, 100, 100 and 75, respectively.

In experiment 1, VisNet was trained in two stages. In the first stage, the 18 feature pairs were used as input stimuli, with each stimulus being presented to VisNet’s retina in sequences of five orientations as described above. However, during this stage, learning was only allowed to take place in layers 1 and 2. This led to the formation of neurons which responded to the feature pairs with some rotation invariance in layer 2. In the second stage, we used the 6 feature triples as stimuli, with learning only allowed in layers 3 and 4. However, during this second training stage, the triples were only presented to VisNet’s input retina in the first 4 orientations (i)–(iv). After the two stages of training were completed, Stringer and Rolls (2002) examined whether the output layer of VisNet had formed top layer neurons that responded invariantly to the 6 triples when presented in all 5 orientations, not just the 4 in which the triples had been presented during training. To provide baseline results for comparison, the results from experiment 1 were compared with results from experiment 2 which involved no training in layers 1, 2 and 3, with the synaptic weights left unchanged from their initial random values.

In Fig. 4.56 numerical results are given for the experiments described. On the left are the single cell information measures for all top (4th) layer neurons ranked in order of their invariance to the triples, while on the right are multiple cell information measures. To help to interpret these results we can compute the maximum single cell information measure according to
where the number of triples is 6. This gives a maximum single cell information measure of 2.6 bits for these test cases. The information results from the experiment demonstrate that even with the triples presented to the network in only four of the five orientations during training, layer 4 is indeed capable of developing rotation invariant neurons that can discriminate effectively between the 6 different feature triples in all 5 orientations, that is with correct recognition from all five perspectives. In addition, the multiple cell information for the experiment reaches the maximal level of 2.6 bits, indicating that the network as a whole is capable of perfect discrimination between the 6 triples in any of the 5 orientations. These results may be compared with the very poor baseline performance from the control experiment, where no learning was allowed before testing. Further results from experiment 1 are presented in Fig. 4.57 where we show the response profiles of a top layer neuron to the 6 triples. It can be seen that this neuron has achieved excellent invariant responses to the 6 triples: the response profiles are independent of orientation, but differentiate between triples in that the responses are maximal for triple 132 and minimal for all other triples. In particular, the cell responses are maximal for triple 132 presented in all 5 of the orientations.

Stringer and Rolls (2002) also performed a control experiment to show that the network really had learned invariant representations specific to the kinds of 3D deformations undergone by the surface features as the objects rotated in-depth. In the control experiment the network was trained on ‘spheres’ with non-deformed surface features; and then as predicted the network failed to operate correctly when it was tested with objects with the features present in the
transformed way that they appear on the surface of a real 3D object.

Stringer and Rolls (2002) were thus able to show how trace learning can form neurons that can respond invariantly to novel rotational within-generic view perspectives of an object, obtained by within-generic view 3D rotations up to 30 degrees from any view encountered during learning. They were able to show in addition that this could occur for a novel view of an object which was not an interpolation from previously shown views. This was possible given that the low order feature combination sets from which an object was composed had been learned about in early layers of VisNet previously. The within-generic view transform invariant object recognition described was achieved through the development of true 3-dimensional representations of objects based on 3-dimensional features and feature combinations, which, unlike 2-dimensional feature combinations, are invariant under moderate in-depth rotations of the object. Thus, in a sense, these rotation invariant representations encode a form of 3-dimensional knowledge with which to interpret the visual input from the real world, that is able provide a basis for robust rotation invariant object recognition with novel perspectives. The particular finding in the work described here was that VisNet can learn how the surface features on 3D objects transform as the object is rotated in depth, and can use knowledge of the characteristics of the transforms to perform 3D object recognition. The knowledge embodied in the network is knowledge of the 3D properties of objects, and in this sense assists the recognition of 3D objects seen from different views.

The process investigated by Stringer and Rolls (2002) will only allow invariant object recognition over moderate 3D object rotations, since rotating an object through a large angle may lead to a catastrophic change in the appearance of the object that requires the new qualitative 3D shape descriptors to be associated with those of the former view. In that case, invariant object recognition must rely on the first process referred to at the start of this Section (4.5.7) in order to associate together the different generic views of an object to produce view invariant object identification. For that process, association of a few cardinal or generic views is likely to be sufficient (Koenderink 1990). The process described in this section of learning how surface features transform is likely to make a major contribution to the within-generic view transform invariance of object identification and recognition.

4.5.8 Capacity of the architecture, and incorporation of a trace rule into a recurrent architecture with object attractors

One issue that has not been considered extensively so far is the capacity of hierarchical feedforward networks of the type exemplified by VisNet that are used for invariant object recognition. One approach to this issue is to note that VisNet operates in the general mode of a competitive network, and that the number of different stimuli that can be categorized by a competitive network is in the order of the number of neurons in the output layer, as described in Section B.4. Given that the successive layers of the real visual system (V1, V2, V4, posterior inferior temporal cortex, anterior inferior temporal cortex) are of the same order of magnitude21, VisNet is designed to work with the same number of neurons in each successive layer. The hypothesis is that because of redundancies in the visual world, each layer of the system by its convergence and competitive categorization can capture sufficient of the statistics of the visual input at each stage to enable correct specification of the properties of the world that specify objects. For example, V1 does not compute all possible combinations of a few lateral geniculate inputs, but instead represents linear series of geniculate inputs to form edge-like and bar-like feature analyzers, which are the dominant arrangement of pixels

21 Of course the details are worth understanding further. V1 is for example somewhat larger than earlier layers, but on the other hand serves the dorsal as well as the ventral stream of visual cortical processing.
Invariant visual object recognition learning

found at the small scale in natural visual scenes. Thus the properties of the visual world at this stage can be captured by a small proportion of the total number of combinations that would be needed if the visual world were random. Similarly, at a later stage of processing, just a subset of all possible combinations of line or edge analyzers would be needed, partly because some combinations are much more frequent in the visual world, and partly because the coding because of convergence means that what is represented is for a larger area of visual space (that is, the receptive fields of the neurons are larger), which also leads to economy and limits what otherwise would be a combinatorial need for feature analyzers at later layers. The hypothesis thus is that the effects of redundancies in the input space of stimuli that result from the statistical properties of natural images (Field 1987), together with the convergent architecture with competitive learning at each stage, produces a system that can perform invariant object recognition for large numbers of objects. Large in this case could be within one or two orders of magnitude of the number of neurons in any one layer of the network (or cortical area in the brain). The extent to which this can be realized can be explored with simulations of the type implemented in VisNet, in which the network can be trained with natural images which therefore reflect fully the natural statistics of the stimuli presented to the real brain.

We should note that a rich variety of information in perceptual space may be represented by subtle differences in the distributed representation provided by the output of the visual system. At the same time, the actual number of different patterns that may be stored in for example a pattern associator connected to the output of the visual system is limited by the number of input connections per neuron from the output neurons of the visual system (see Section B.2). One essential function performed by the ventral visual system is to provide an invariant representation which can be read by a pattern associator in such a way that if the pattern associator learns about one view of the object, then the visual system allows generalization to another view of the same object, because the same output neurons are activated by the different view. In the sense that any view can and must activate the same output neurons of the visual system (the input to the associative network), then we can say the invariance is made explicit in the representation. Making some properties of an input representation explicit in an output representation has a major function of enabling associative networks that use visual inputs in for example recognition, episodic memory, emotion and motivation to generalize correctly, that is invariantly with respect to image transforms that are all consistent with the same object in the world (Rolls and Treves 1998).

Another approach to the issue of the capacity of networks that use trace-learning to associate together different instances (e.g. views) of the same object is to reformulate the issue in the context of autoassociation (attractor) networks, where analytic approaches to the storage capacity of the network are well developed (see Section B.3, Amit (1989), and Rolls and Treves (1998)). This approach to the storage capacity of networks that associate together different instantiations of an object to form invariant representations has been developed by Parga and Rolls (1998) and Elliffe, Rolls, Parga and Renart (2000), and is described next.

In this approach, the storage capacity of a recurrent network which performs for example view invariant recognition of objects by associating together different views of the same object which tend to occur close together in time, was studied (Parga and Rolls 1998, Elliffe, Rolls, Parga and Renart 2000). The architecture with which the invariance is computed is a little different to that described earlier. In the model of Rolls ((1992a), (1994a), (1995b), Wallis and Rolls (1997), Rolls and Milward (2000), and Rolls and Stringer (2006)), the postsynaptic memory trace enabled different afferents from the preceding stage to modify onto the same postsynaptic neuron (see Fig. 4.58). In that model there were no recurrent connections between the neurons, although such connections were one way in which it was postulated the memory trace might be implemented, by simply keeping the representation of one view or aspect active until the next view appeared. Then an association would occur between representations that
Computational issues in feature hierarchies

In this Section, the different exemplars of an object which need to be associated together are called views, for simplicity, but could at earlier stages of the hierarchy represent for example similar feature combinations (derived from the same object) in different positions in space.

22In this Section, the different exemplars of an object which need to be associated together are called views, for simplicity, but could at earlier stages of the hierarchy represent for example similar feature combinations (derived from the same object) in different positions in space.
Fig. 4.60 A schematic illustration of the first type of associations contributing to the synaptic matrix considered by Parga and Rolls (1998). Object 1 \((O_1)\) has five views labelled \(v_1\) to \(v_5\), etc. The matrix is formed by associating the pattern presented in the columns with itself, that is with the same pattern presented as rows.

In more detail, Parga and Rolls (1998) considered two main approaches. First, one could store in a synaptic weight matrix the \(s\) views of an object. This consists of equally associating all the views to each other, including the association of each view with itself. Choosing in Fig. 4.60 an example such that objects are defined in terms of five different views, this might produce (if each view produced firing of one neuron at a rate of 1) a block of \(s \times s\) pairs of views contributing to the synaptic efficacies each with value 1. Object 2 might produce another block of synapses of value 1 further along the diagonal, and symmetric about it. Each object or memory could then be thought of as a single attractor with a distributed representation involving five elements (each element representing a different view). Then the capacity of the system in terms of the number \(P_o\) of objects that can be stored is just the number of separate attractors which can be stored in the network. For random fully distributed patterns this is as shown numerically by Hopfield (1982)

\[
P_o = 0.14 \ C \quad (4.40)
\]

where there are \(C\) inputs per neuron (and \(N = C\) neurons if the network is fully connected). Now the synaptic matrix envisaged here does not consist of random fully distributed binary elements, but instead we will assume has a sparseness \(a = s/N\), where \(s\) is the number of views stored for each object, from any of which the whole representation of the object must be recognized. In this case, one can show (Gardner 1988, Tsodyks and Feigelson 1988, Treves and Rolls 1991) that the number of objects that can be stored and correctly retrieved is

\[
P_o = \frac{k \ C}{a \ \ln(1/a)} \quad (4.41)
\]

where \(C\) is the number of synapses on each neuron devoted to the recurrent collaterals from other neurons in the network, and \(k\) is a factor that depends weakly on the detailed structure of the rate distribution, on the connectivity pattern, etc., but is approximately in the order of 0.2–0.3. A problem with this proposal is that as the number of views of each object increases to a large number \((\text{e.g.} > 20)\), the network will fail to retrieve correctly the internal representation
Computational issues in feature hierarchies

Fig. 4.61 A schematic illustration of the second and main type of associations contributing to the synaptic matrix considered by Parga and Rolls (1998) and Elliffe, Rolls, Parga and Renart (2000). Object 1 (O₁) has five views labelled v₁ to v₅, etc. The association of any one view with itself has strength 1, and of any one with another view of the same object has strength b.

of the object starting from any one view (which is only a fraction 1/s of the length of the stored pattern that represents an object).

The second approach, taken by Parga and Rolls (1998) and Elliffe, Rolls, Parga and Renart (2000), is to consider the operation of the network when the associations between pairs of views can be described by a matrix that has the general form shown in Fig. 4.61. Such an association matrix might be produced by different views of an object appearing after a given view with equal probability, and synaptic modification occurring of the view with itself (giving rise to the diagonal term), and of any one view with that which immediately follows it. The same weight matrix might be produced not only by pairwise association of successive views because the association rule allows for associations over the short time scale of e.g. 100–200 ms, but might also be produced if the synaptic trace had an exponentially decaying form over several hundred milliseconds, allowing associations with decaying strength between views separated by one or more intervening views. The existence of a regime, for values of the coupling parameter between pairs of views in a finite interval, such that the presentation of any of the views of one object leads to the same attractor regardless of the particular view chosen as a cue, is one of the issues treated by Parga and Rolls (1998) and Elliffe, Rolls, Parga and Renart (2000). A related problem also dealt with was the capacity of this type of synaptic matrix: how many objects can be stored and retrieved correctly in a view invariant way? Parga and Rolls (1998) and Elliffe, Rolls, Parga and Renart (2000) showed that the number grows linearly with the number of recurrent collateral connections received by each neuron. Some of the groundwork for this approach was laid by the work of Amit and collaborators (Griniasty, Tsodyks and Amit 1993, Amit 1989).

A variant of the second approach is to consider that the remaining entries in the matrix shown in Fig. 4.61 all have a small value. This would be produced by the fact that sometimes a view of one object would be followed by a view of a different object, when for example a large saccade was made, with no explicit resetting of the trace. On average, any one object would follow another rarely, and so the case is considered when all the remaining associations between pairs of views have a low value.

Parga and Rolls (1998) and Elliffe, Rolls, Parga and Renart (2000) were able to show that invariant object recognition is feasible in attractor neural networks in the way described. The system is able to store and retrieve in a view invariant way an extensive number of objects,
each defined by a finite set of views. What is implied by extensive is that the number of objects is proportional to the size of the network. The crucial factor that defines this size is the number of connections per neuron. In the case of the fully connected networks considered in this section, the size is thus proportional to the number of neurons. To be particular, the number of objects that can be stored is $0.081 N/5$, when there are five views of each object. The number of objects is $0.073 N/11$, when there are eleven views of each object. This is an interesting result in network terms, in that $s$ views each represented by an independent random set of active neurons can, in the network described, be present in the same ‘object’ attraction basin. It is also an interesting result in neurophysiological terms, in that the number of objects that can be represented in this network scales linearly with the number of recurrent connections per neuron. That is, the number of objects $P_o$ that can be stored is approximately

$$P_o = \frac{kC}{s}$$

where $C$ is the number of synapses on each neuron devoted to the recurrent collaterals from other neurons in the network, $s$ is the number of views of each object, and $k$ is a factor that is in the region of 0.07–0.09 (Parga and Rolls 1998).

Although the explicit numerical calculation was done for a rather small number of views for each object (up to 11), the basic result, that the network can support this kind of ‘object’ phase, is expected to hold for any number of views (the only requirement being that it does not increase with the number of neurons). This is of course enough: once an object is defined by a set of views, when the network is presented with a somewhat different stimulus or a noisy version of one of them it will still be in the attraction basin of the object attractor.

Parga and Rolls (1998) thus showed that multiple (e.g. ‘view’) patterns could be within the basin of attraction of a shared (e.g. ‘object’) representation, and that the capacity of the system was proportional to the number of synapses per neuron divided by the number of views of each object.

Elliffe, Rolls, Parga and Renart (2000) extended the analysis of Parga and Rolls (1998) by showing that correct retrieval could occur where retrieval ‘view’ cues were distorted; where there was some association between the views of different objects; and where there was only partial and indeed asymmetric connectivity provided by the associatively modified recurrent collateral connections in the network. The simulations also extended the analysis by showing that the system can work well with sparse patterns, and indeed that the use of sparse patterns increases (as expected) the number of objects that can be stored in the network.

Taken together, the work described by Parga and Rolls (1998) and Elliffe, Rolls, Parga and Renart (2000) introduced the idea that the trace rule used to build invariant representations could be implemented in the recurrent collaterals of a neural network (as well as or as an alternative to its incorporation in the forward connections from one layer to another incorporated in VisNet), and provided a precise analysis of the capacity of the network if it operated in this way. In the brain, it is likely that the recurrent collateral connections between cortical pyramidal cells in visual cortical areas do contribute to building invariant representations, in that if they are associatively modifiable, as seems likely, and because there is continuing firing for typically 100–300 ms after a stimulus has been shown, associations between different exemplars of the same object that occur together close in time would almost necessarily become built into the recurrent synaptic connections between pyramidal cells.

Invariant representation of faces in the context of attractor neural networks has also been discussed by Bartlett and Sejnowski (1997) in terms of a model where different views of faces are presented in a fixed sequence (Griniasty, Tsodyks and Amit 1993). This is not however the general situation; normally any pair of views can be seen consecutively and they will become associated. The model described by Parga and Rolls (1998) treats this more general situation.
We wish to note the different nature of the invariant object recognition problem studied here, and the paired associate learning task studied by Miyashita and Chang (1988), Miyashita (1988), and Sakai and Miyashita (1991). In the invariant object recognition case no particular learning protocol is required to produce an activity of the inferior temporal cortex cells responsible for invariant object recognition that is maintained for 300 ms. The learning can occur rapidly, and the learning occurs between stimuli (e.g. different views) which occur with no intervening delay. In the paired associate task, which had the aim of providing a model of semantic memory, the monkeys must learn to associate together two stimuli that are separated in time (by a number of seconds), and this type of learning can take weeks to train. During the delay period the sustained activity is rather low in the experiments, and thus the representation of the first stimulus that remains is weak, and can only poorly be associated with the second stimulus. However, formally the learning mechanism could be treated in the same way as that used by Parga and Rolls (1998) for invariant object recognition. The experimental difference is just that in the paired associate task used by Miyashita et al., it is the weak memory of the first stimulus that is associated with the second stimulus. In contrast, in the invariance learning, it would be the firing activity being produced by the first stimulus (not the weak memory of the first stimulus) that can be associated together. It is possible that the perirhinal cortex makes a useful contribution to invariant object recognition by providing a short-term memory that helps successive views of the same objects to become associated together (Buckley, Booth, Rolls and Gaffan 2001).

The mechanisms described here using an attractor network with a trace associative learning rule would apply most naturally when a small number of representations need to be associated together to represent an object. One example is associating together what is seen when an object is viewed from different perspectives. Another example is scale, with respect to which neurons early in the visual system tolerate scale changes of approximately 1.5 octaves, so that the whole scale range could be covered by associating together a limited number of such representations (see Chapter 5 of Rolls and Deco (2002) and Fig. 4.2). The mechanism would not be so suitable when a large number of different instances would need to be associated together to form an invariant representation of objects, as might be needed for translation invariance. For the latter, the standard model of VisNet with the associative trace learning rule implemented in the feedforward connections (or trained by continuous spatial transformation learning as described in Section 4.5.11) would be more appropriate. However, both types of mechanism, with the trace rule in the feedforward or in recurrent collateral synapses, could contribute (separately or together) to achieve invariant representations. Part of the interest of the attractor approach described in this section is that it allows analytic investigation.

Another approach to training invariance is the purely associative mechanism continuous spatial transformation learning, described in Section 4.5.11. With this training procedure, the capacity is increased with respect to the number of training locations, with for example 169 training locations producing translation invariant representations for two face stimuli (Perry, Rolls and Stringer 2007). It will be of interest in future research to investigate how the VisNet architecture, whether trained with a trace or purely associative rule, scales up with respect to capacity as the number of neurons in the system increases. More distributed representations in the output layer, which may be facilitated by encouraging the system to form a self-organising map, by including short range excitation as well as longer range inhibition between the neurons (see Section 4.5.1), may also help to increase the capacity.
4.5.9 Vision in natural scenes – effects of background versus attention

Object-based attention refers to attention to an object. For example, in a visual search task the object might be specified as what should be searched for, and its location must be found. In spatial attention, a particular location in a scene is pre-cued, and the object at that location may need to be identified. Here we consider some of the neurophysiology of object selection and attention in the context of a feature hierarchy approach to invariant object recognition. The computational mechanisms of attention, including top-down biased competition, are described in Chapter 6.

4.5.9.1 Neurophysiology of object selection in the inferior temporal visual cortex

Much of the neurophysiology, psychophysics, and modelling of attention has been with a small number, typically two, of objects in an otherwise blank scene. In this Section, I consider how attention operates in complex natural scenes, and in particular describe how the inferior temporal visual cortex operates to enable the selection of an object in a complex natural scene (see also Rolls and Deco (2006)). The inferior temporal visual cortex contains distributed and invariant representations of objects and faces (Rolls 2000a, Rolls and Deco 2002, Booth and Rolls 1998, Rolls, Treves and Tovee 1997b, Rolls and Tovee 1995b, Tovee, Rolls and Azzopardi 1994, Hasselmo, Rolls and Baylis 1989a, Rolls and Baylis 1986, Rolls 2007i, Rolls 2007j, Rolls 2007f).

To investigate how attention operates in complex natural scenes, and how information is passed from the inferior temporal cortex (IT) to other brain regions to enable stimuli to be selected from natural scenes for action, Rolls, Aggelopoulos and Zheng (2003a) analyzed the responses of inferior temporal cortex neurons to stimuli presented in complex natural backgrounds. The monkey had to search for two objects on a screen, and a touch of one object was rewarded with juice, and of another object was punished with saline (see Fig. 4.4 for a schematic illustration and Fig. 4.62 for a version of the display with examples of the stimuli shown to scale). Neuronal responses to the effective stimuli for the neurons were compared when the objects were presented in the natural scene or on a plain background. It
Fig. 4.63 Summary of the receptive field sizes of inferior temporal cortex neurons to a 5 degree effective stimulus presented in either a blank background (blank screen) or in a natural scene (complex background). The stimulus that was a target for action in the different experimental conditions is marked by T. When the target stimulus was touched, a reward was obtained. The mean receptive field diameter of the population of neurons analyzed, and the mean firing rate in spikes/s, is shown. The stimuli subtended 5 deg at the retina, and occurred on each trial in a random position in the 70 deg × 55 deg screen. The dashed circle is proportional to the receptive field size. Top row: responses with one visual stimulus in a blank (left) or complex (right) background. Middle row: responses with two stimuli, when the effective stimulus was not the target of the visual search. Bottom row: responses with two stimuli, when the effective stimulus was the target of the visual search. (After Rolls, Aggelopoulos, and Zheng 2003.)

was found that the overall response of the neuron to objects was hardly reduced when they were presented in natural scenes, and the selectivity of the neurons remained. However, the main finding was that the magnitudes of the responses of the neurons typically became much less in the real scene the further the monkey fixated in the scene away from the object (see
A small receptive field size has also been found in inferior temporal cortex neurons when monkeys have been trained to discriminate closely spaced small visual stimuli (DiCarlo and Maunsell 2003).

It is proposed that this reduced translation invariance in natural scenes helps an unambiguous representation of an object which may be the target for action to be passed to the brain regions that receive from the primate inferior temporal visual cortex. It helps with the binding problem, by reducing in natural scenes the effective receptive field of at least some inferior temporal cortex neurons to approximately the size of an object in the scene.

It is also found that in natural scenes, the effect of object-based attention on the response properties of inferior temporal cortex neurons is relatively small, as illustrated in Fig. 4.63 (Rolls, Aggelopoulos and Zheng 2003a).

4.5.9.2 Attention in natural scenes – a computational account

The results summarized in Fig. 4.63 for 5 degree stimuli show that the receptive fields were large (77.6 degrees) with a single stimulus in a blank background (top left), and were greatly reduced in size (to 22.0 degrees) when presented in a complex natural scene (top right). The results also show that there was little difference in receptive field size or firing rate in the complex background when the effective stimulus was selected for action (bottom right, 19.2 degrees), and when it was not (middle right, 15.6 degrees) (Rolls, Aggelopoulos and Zheng 2003a). (For comparison, the effects of attention against a blank background were much larger, with the receptive field increasing from 17.2 degrees to 47.0 degrees as a result of object-based attention, as shown in Fig. 4.63, left middle and bottom.)

Trappenberg, Rolls and Stringer (2002) have suggested what underlying mechanisms could account for these findings, and simulated a model to test the ideas. The model utilizes an attractor network representing the inferior temporal visual cortex (implemented by the recurrent connections between inferior temporal cortex neurons), and a neural input layer with several retinotopically organized modules representing the visual scene in an earlier visual cortical area such as V4 (see Fig. 4.64). The attractor network aspect of the model produces the property that the receptive fields of IT neurons can be large in blank scenes by enabling a weak input in the periphery of the visual field to act as a retrieval cue for the object attractor. On the other hand, when the object is shown in a complex background, the object closest to the fovea tends to act as the retrieval cue for the attractor, because the fovea is given increased weight in activating the IT module because the magnitude of the input activity from objects at the fovea is greatest due to the higher magnification factor of the fovea incorporated into the model. This results in smaller receptive fields of IT neurons in complex scenes, because the object tends to need to be close to the fovea to trigger the attractor into the state representing that object. (In other words, if the object is far from the fovea, then it will not trigger neurons in IT which represent it, because neurons in IT are preferentially being activated by another object at the fovea.) This may be described as an attractor model in which the competition for which attractor state is retrieved is weighted towards objects at the fovea.

Attentional top-down object-based inputs can bias the competition implemented in this attractor model, but have relatively minor effects (in for example increasing receptive field size) when they are applied in a complex natural scene, as then as usual the stronger forward inputs dominate the states reached. In this network, the recurrent collateral connections may be thought of as implementing constraints between the different inputs present, to help arrive at firing in the network which best meets the constraints. In this scenario, the preferential weighting of objects close to the fovea because of the increased magnification factor at the fovea is a useful principle in enabling the system to provide useful output. The attentional object biasing effect is much more marked in a blank scene, or a scene with only two objects.
Computational issues in feature hierarchies

Visual Input

Object bias

V4

IT

Fig. 4.64 The architecture of the inferior temporal cortex (IT) model of Trappenberg, Rolls and Stringer (2002) operating as an attractor network with inputs from the fovea given preferential weighting by the greater magnification factor of the fovea. The model also has a top-down object-selective bias input. The model was used to analyze how object vision and recognition operate in complex natural scenes.

present at similar distances from the fovea, which are conditions in which attentional effects have frequently been examined. The results of the investigation (Trappenberg, Rolls and Stringer 2002) thus suggest that top-down attention may be a much more limited phenomenon in complex, natural, scenes than in reduced displays with one or two objects present. The results also suggest that the alternative principle, of providing strong weight to whatever is close to the fovea, is an important principle governing the operation of the inferior temporal visual cortex, and in general of the output of the visual system in natural environments. This principle of operation is very important in interfacing the visual system to action systems, because the effective stimulus in making inferior temporal cortex neurons fire is in natural scenes usually on or close to the fovea. This means that the spatial coordinates of where the object is in the scene do not have to be represented in the inferior temporal visual cortex, nor passed from it to the action selection system, as the latter can assume that the object making IT neurons fire is close to the fovea in natural scenes.

There may of course be in addition a mechanism for object selection that takes into account the locus of covert attention when actions are made to locations not being looked at. However, the simulations described in this section suggest that in any case covert attention is likely to be a much less significant influence on visual processing in natural scenes than in reduced scenes with one or two objects present.

Given these points, one might question why inferior temporal cortex neurons can have
such large receptive fields, which show translation invariance. At least part of the answer to this may be that inferior temporal cortex neurons must have the capability to be large if they are to deal with large objects. A V1 neuron, with its small receptive field, simply could not receive input from all the features necessary to define an object. On the other hand, inferior temporal cortex neurons may be able to adjust their size to approximately the size of objects, using in part the interactive effects described in Chapter 6, and need the capability for translation invariance because the actual relative positions of the features of an object could be at different relative positions in the scene. For example, a car can be recognized whichever way it is viewed, so that the parts (such as the bonnet or hood) must be identifiable as parts wherever they happen to be in the image, though of course the parts themselves also have to be in the correct relative positions, as allowed for by the hierarchical feature analysis architecture described in this chapter.

Some details of the simulations follow. Each independent module within ‘V4’ in Fig. 4.64 represents a small part of the visual field and receives input from earlier visual areas represented by an input vector for each possible location which is unique for each object. Each module was 6 degrees in width, matching the size of the objects presented to the network. For the simulations Trappenberg, Rolls and Stringer (2002) chose binary random input vectors representing objects with $N^{V4} a^{V4}$ components set to ones and the remaining $N^{V4} (1 - a^{V4})$ components set to zeros. $N^{V4}$ is the number of nodes in each module and $a^{V4}$ is the sparseness of the representation which was set to be $a^{V4} = 0.2$ in the simulations.

The structure labelled ‘IT’ represents areas of visual association cortex such as the inferior temporal visual cortex and cortex in the anterior part of the superior temporal sulcus in which neurons provide distributed representations of faces and objects (Booth and Rolls 1998, Rolls 2000a). Nodes in this structure are governed by leaky integrator dynamics (similar to those used in the mean field approach described in Section B.8.1) with time constant $\tau$

$$\frac{d h_{i}^{IT}(t)}{dt} = -h_{i}^{IT}(t) + \sum_j (w_{ij}^{IT} - e^{IT}) y_j^{IT}(t) + \sum_k w_{ik}^{V4} a^{V4}(t) + k^{IT BIAS} I_{i}^{OBJ}. \quad (4.43)$$

The firing rate $y_i^{IT}$ of the $i$th node is determined by a sigmoidal function from the activation $h_i^{IT}$ as follows

$$y_i^{IT}(t) = \frac{1}{1 + \exp[-2\beta (h_i^{IT}(t) - \alpha)]}, \quad (4.44)$$

where the parameters $\beta = 1$ and $\alpha = 1$ represent the gain and the bias, respectively.

The recognition functionality of this structure is modelled as an attractor neural network (ANN) with trained memories indexed by $\mu$ representing particular objects. The memories are formed through Hebbian learning on sparse patterns,

$$w_{ij}^{IT} = k^{IT} \sum_{\mu} (\xi_{ij}^{\mu} - a^{IT})(\xi_{ij}^{\mu} - a^{IT}), \quad (4.45)$$

where $k^{IT}$ (set to 1 in the simulations) is a normalization constant that depends on the learning rate, $a^{IT} = 0.2$ is the sparseness of the training pattern in IT, and $\xi_{ij}^{\mu}$ are the components of the pattern used to train the network. The constant $e^{IT}$ in equation (4.43) represents the strength of the activity-dependent global inhibition simulating the effects of inhibitory interneurons. The external ‘top-down’ input vector $I^{OBJ}$ produces object-selective inputs, which are used as the attentional drive when a visual search task is simulated. The strength of this object bias is modulated by the value of $k^{IT BIAS}$ in equation (4.43).
The weights $w_{ij}^{IT-V4}$ between the V4 nodes and IT nodes were trained by Hebbian learning of the form

$$w_{ij}^{IT-V4} = k^{IT-V4}(k) \sum_{\mu} (c_{ij}^{\mu} - a^{V4}) (c_{ij}^{\mu} - a^{IT}).$$

(4.46)

to produce object representations in IT based on inputs in V4. The normalizing modulation factor $k^{IT-V4}(k)$ allows the gain of inputs to be modulated as a function of their distance from the fovea, and depends on the module $k$ to which the presynaptic node belongs. The model supports translation invariant object recognition of a single object in the visual field if the normalization factor is the same for each module and the model is trained with the objects placed at every possible location in the visual field. The translation invariance of the weight vectors between each ‘V4’ module and the IT nodes is however explicitly modulated in the model by the module-dependent modulation factor $k^{IT-V4}(k)$ as indicated in Fig. 4.64 by the width of the lines connecting V4 with IT. The strength of the foveal V4 module is strongest, and the strength decreases for modules representing increasing eccentricity. The form of this modulation factor was derived from the parameterization of the cortical magnification factors given by Dow, Snyder, Vautin and Bauer (1981).

To study the ability of the model to recognize trained objects at various locations relative to the fovea the system was trained on a set of objects. The network was then tested with distorted versions of the objects, and the ‘correlation’ between the target object and the final state of the attractor network was taken as a measure of the performance. The correlation was estimated from the normalized dot product between the target object vector that was used during training the IT network, and the state of the IT network after a fixed amount of time sufficient for the network to settle into a stable state. The objects were always presented on backgrounds with some noise (introduced by flipping 2% of the bits in the scene which were not the test stimulus) in order to utilize the properties of the attractor network, and because the input to IT will inevitably be noisy under normal conditions of operation.

In the first simulation only one object was present in the visual scene in a plain (blank) background at different eccentricities from the fovea. As shown in Fig. 4.65a by the line labelled ‘blank background’, the receptive fields of the neurons were very large. The value of the object bias $k^{IT-V4}_{HAS}$ was set to 0 in these simulations. Good object retrieval (indicated by large correlations) was found even when the object was far from the fovea, indicating large IT receptive fields with a blank background. The reason that any drop is seen in performance as a function of eccentricity is because flipping 2% of the bits outside the object introduces some noise into the recall process. This demonstrates that the attractor dynamics can support translation invariant object recognition even though the translation invariant weight vectors between V4 and IT are explicitly modulated by the modulation factor $k^{IT-V4}$ derived from the cortical magnification factor.

In a second simulation individual objects were placed at all possible locations in a natural and cluttered visual scene. The resulting correlations between the target pattern and the asymptotic IT state are shown in Fig. 4.65a with the line labelled ‘natural background’. Many objects in the visual scene are now competing for recognition by the attractor network, and the objects around the foveal position are enhanced through the modulation factor derived from the cortical magnification factor. This results in a much smaller size of the receptive field of IT neurons when measured with objects in natural backgrounds.

In addition to this major effect of the background on the size of the receptive field, which parallels and may account for the physiological findings outlined above and in Section 4.5.9.1,

23 This parameterization is based on V1 data. However, it was shown that similar forms of the magnification factor hold also in V4 (Gattass, Sousa and Covey 1985). Similar results to the ones presented here can also be achieved with different forms of the modulation factor such as a shifted Gaussian.
there is also a dependence of the size of the receptive fields on the level of object bias provided to the IT network. Examples are shown in Fig. 4.65b where an object bias was used. The object bias biases the IT network towards the expected object with a strength determined by the value of \( k^{\text{IT,BIAS}} \), and has the effect of increasing the size of the receptive fields in both blank and natural backgrounds (see Fig. 4.65b compared to a). This models the effect found neurophysiologically (Rolls, Aggelopoulos and Zheng 2003a).24

Some of the conclusions are as follows (Trappenberg, Rolls and Stringer 2002). When single objects are shown in a scene with a blank background, the attractor network helps neurons to respond to an object with large eccentricities of this object relative to the fovea of the agent. When the object is presented in a natural scene, other neurons in the inferior temporal cortex become activated by the other effective stimuli present in the visual field, and these forward inputs decrease the response of the network to the target stimulus by a competitive process. The results found fit well with the neurophysiological data, in that IT operates with almost complete translation invariance when there is only one object in the scene, and reduces the receptive field size of its neurons when the object is presented in a cluttered environment. The model described here provides an explanation of the responses of real IT neurons in natural scenes.

In natural scenes, the model is able to account for the neurophysiological data that the IT neuronal responses are larger when the object is close to the fovea, by virtue of fact that objects close to the fovea are weighted by the cortical magnification factor related modulation \( k_{\text{IT} \rightarrow \text{V4}} \).

The model accounts for the larger receptive field sizes from the fovea of IT neurons in natural backgrounds if the target is the object being selected compared to when it is not selected (Rolls, Aggelopoulos and Zheng 2003a). The model accounts for this by an effect of top-down bias which simply biases the neurons towards particular objects compensating for their decreasing inputs produced by the decreasing magnification factor modulation with increasing distance from the fovea. Such object-based attention signals could originate in the

\[ k^{\text{IT,BIAS}} = \begin{cases} 0.7 & \text{for natural backgrounds} \\ 0.1 & \text{for blank backgrounds} \end{cases} \]

24 This was set to 0.7 in the experiments with a natural background and to 0.1 with a blank background, reflecting the fact that more attention may be needed to find objects in natural cluttered environments because of the noise present than in blank backgrounds. Equivalently, a given level of attention may have a smaller effect in a natural scene than in a blank background, as found neurophysiologically (Rolls, Aggelopoulos and Zheng 2003a).
Computational issues in feature hierarchies

Fig. 4.66 Cortical architecture for hierarchical and attention-based visual perception after Deco and Rolls (2004). The system is essentially composed of five modules structured such that they resemble the two known main visual paths of the mammalian visual cortex. Information from the retino-geniculo-striate pathway enters the visual cortex through area V1 in the occipital lobe and proceeds into two processing streams. The occipital-temporal stream leads ventrally through V2–V4 and IT (inferior temporal visual cortex), and is mainly concerned with object recognition. The occipito-parietal stream leads dorsally into PP (posterior parietal complex), and is responsible for maintaining a spatial map of an object’s location. The solid lines with arrows between levels show the forward connections, and the dashed lines the top-down backprojections. Short-term memory systems in the prefrontal cortex (PF46) apply top-down attentional bias to the object or spatial processing streams. (After Deco and Rolls 2004.)

Prefrontal cortex and could provide the object bias for the inferior temporal visual cortex (Renart, Parga and Rolls 2000) (see Chapter 6).

Important properties of the architecture for obtaining the results just described are the high magnification factor at the fovea and the competition between the effects of different inputs, implemented in the above simulation by the competition inherent in an attractor network.

We have also been able to obtain similar results in a hierarchical feedforward network where each layer operates as a competitive network (Deco and Rolls 2004). This network thus captures many of the properties of our hierarchical model of invariant object recognition (Rolls 1992a, Wallis and Rolls 1997, Rolls and Milward 2000, Stringer and Rolls 2000, Rolls
Deco and Rolls (2004) trained the network shown in Fig. 4.66 with two objects, and used the trace learning rule (Wallis and Rolls 1997, Rolls and Milward 2000) in order to achieve translation invariance. In a first experiment we placed only one object on the retina at different distances from the fovea (i.e. different eccentricities relative to the fovea). This corresponds to the blank background condition. In a second experiment, we also placed the object at different eccentricities relative to the fovea, but on a cluttered natural background. Larger receptive fields were found with the blank as compared to the cluttered natural background.

Deco and Rolls (2004) also studied the influence of object-based attentional top-down bias on the effective size of the receptive field of an inferior temporal cortex neuron for the case of an object in a blank or a cluttered background. To do this, they repeated the two simulations but now considered a non-zero top-down bias coming from prefrontal area 46v and impinging on the inferior temporal cortex neuron specific for the object tested. When no attentional object bias was introduced, a shrinkage of the receptive field size was observed in the complex vs the blank background. When attentional object bias was introduced, the shrinkage of the receptive field due to the complex background was somewhat reduced. This is consistent with the neurophysiological results (Rolls, Aggelopoulos and Zheng 2003a). In the framework of the model (Deco and Rolls 2004), the reduction of the shrinkage of the receptive field is due to the biasing of the competition in the inferior temporal cortex layer in favour of the specific IT neuron tested, so that it shows more translation invariance (i.e. a slightly larger receptive field). The increase of the receptive field size of an IT neuron, although small, produced by the external top-down attentional bias offers a mechanism for facilitation of the search for specific objects in complex natural scenes (see further Chapter 6).

We note that it is possible that a ‘spotlight of attention’ (Desimone and Duncan 1995) can be moved covertly away from the fovea as described in Chapter 6. However, at least during normal visual search tasks in natural scenes, the neurons are sensitive to the object at which the monkey is looking, that is primarily to the object that is on the fovea, as shown by Rolls, Aggelopoulos and Zheng (2003a) and Aggelopoulos and Rolls (2005), and described in Sections 4.5.9.1 and 4.5.10.

### 4.5.10 The representation of multiple objects in a scene

When objects have distributed representations, there is a problem of how multiple objects (whether the same or different) can be represented in a scene, because the distributed representations overlap, and it may not be possible to determine whether one has an amalgam of several objects, or a new object (Mozer 1991), or multiple instances of the same object, let alone the relative spatial positions of the objects in a scene. Yet humans can determine the relative spatial locations of objects in a scene even in short presentation times without eye movements (Biederman 1972) (and this has been held to involve some spotlight of attention). Aggelopoulos and Rolls (2005) analyzed this issue by recording from single inferior temporal cortex neurons with five objects simultaneously present in the receptive field. They found that although all the neurons responded to their effective stimulus when it was at the fovea, some could also respond to their effective stimulus when it was in some but not other parafoveal positions 10 degrees from the fovea. An example of such a neuron is shown in Fig. 4.67. The asymmetry is much more evident in a scene with 5 images present (Fig. 4.67A) than when
only one image is shown on an otherwise blank screen (Fig. 4.67B). Competition between different stimuli in the receptive field thus reveals the asymmetry in the receptive field of inferior temporal visual cortex neurons.

The asymmetry provides a way of encoding the position of multiple objects in a scene. Depending on which asymmetric neurons are firing, the population of neurons provides information to the next processing stage not only about which image is present at or close to the fovea, but where it is with respect to the fovea. This information is provided by neurons that have firing rates that reflect the relevant information, and stimulus-dependent synchrony is not necessary. Top-down attentional biasing input could thus, by biasing the appropriate neurons, facilitate bottom-up information about objects without any need to alter the time relations between the firing of different neurons. The exact position of the object with respect to the fovea, and effectively thus its spatial position relative to other objects in the scene, would then be made evident by the subset of asymmetric neurons firing.

This is thus the solution that these experiments indicate is used for the representation of multiple objects in a scene, an issue that has previously been difficult to account for in neural systems with distributed representations (Mozer 1991) and for which ‘attention’ has been a proposed solution. The learning of invariant representations of objects when multiple objects are present in a scene is considered in Section 4.5.6.2.
4.5.11 Learning invariant representations using spatial continuity: Continuous Spatial Transformation learning

The temporal continuity typical of objects has been used in an associative learning rule with a short-term memory trace to help build invariant object representations in the networks described previously in this chapter. Stringer, Perry, Rolls and Proske (2006) showed that spatial continuity can also provide a basis for helping a system to self-organize invariant representations. They introduced a new learning paradigm ‘continuous spatial transformation (CT) learning’ which operates by mapping spatially similar input patterns to the same postsynaptic neurons in a competitive learning system. As the inputs move through the space of possible continuous transforms (e.g. translation, rotation, etc.), the active synapses are modified onto the set of postsynaptic neurons. Because other transforms of the same stimulus overlap with previously learned exemplars, a common set of postsynaptic neurons is activated by the new transforms, and learning of the new active inputs onto the same postsynaptic neurons is facilitated.

The concept is illustrated in Fig. 4.68. During the presentation of a visual image at one position on the retina that activates neurons in layer 1, a small winning set of neurons in layer 2 will modify (through associative learning) their afferent connections from layer 1 to respond well to that image in that location. When the same image appears later at nearby locations, so that there is spatial continuity, the same neurons in layer 2 will be activated because some of the active afferents are the same as when the image was in the first position. The key point is that if these afferent connections have been strengthened sufficiently while the image is in the first location, then these connections will be able to continue to activate the same neurons in layer 2 when the image appears in overlapping nearby locations. Thus the same neurons in the output layer have learned to respond to inputs that have similar vector elements in common.

As can be seen in Fig. 4.68, the process can be continued for subsequent shifts, provided that a sufficient proportion of input cells stay active between individual shifts. This whole process is repeated throughout the network, both horizontally as the image moves on the retina, and hierarchically up through the network. Over a series of stages, transform invariant (e.g. location invariant) representations of images are successfully learned, allowing the network to perform invariant object recognition. A similar CT learning process may operate for other kinds of transformation, such as change in view or size.

Stringer, Perry, Rolls and Proske (2006) demonstrated that VisNet can be trained with continuous spatial transformation learning to form view invariant representations. They showed that CT learning requires the training transforms to be relatively close together spatially so that spatial continuity is present in the training set; and that the order of stimulus presentation is not crucial, with even interleaving with other objects possible during training, because it is spatial continuity rather the temporal continuity that drives the self-organizing learning with the purely associative synaptic modification rule.

Perry, Rolls and Stringer (2006) extended these simulations with VisNet of view invariant learning using CT to more complex 3D objects, and using the same training images in human psychophysical investigations, showed that view invariant object learning can occur when spatial but not temporal continuity applies in a training condition in which the images of different objects were interleaved. However, they also found that the human view invariance learning was better if sequential presentation of the images of an object was used, indicating that temporal continuity is an important factor in human invariance learning.

Perry, Rolls and Stringer (2007) extended the use of continuous spatial transformation learning to translation invariance. They showed that translation invariant representations can be learned by continuous spatial transformation learning; that the transforms must be close for this to occur; that the temporal order of presentation of each transformed image during
Fig. 4.68 An illustration of how continuous spatial transformation (CT) learning would function in a network with a single layer of forward synaptic connections between an input layer of neurons and an output layer. Initially the forward synaptic weights are set to random values. The top part (a) shows the initial presentation of a stimulus to the network in position 1. Activation from the (shaded) active input cells is transmitted through the initially random forward connections to stimulate the cells in the output layer. The shaded cell in the output layer wins the competition in that layer. The weights from the active input cells to the active output neuron are then strengthened using an associative learning rule. The bottom part (b) shows what happens after the stimulus is shifted by a small amount to a new partially overlapping position 2. As some of the active input cells are the same as those that were active when the stimulus was presented in position 1, the same output cell is driven by these previously strengthened afferents to win the competition again. The rightmost shaded input cell activated by the stimulus in position 2, which was inactive when the stimulus was in position 1, now has its connection to the active output cell strengthened (denoted by the dashed line). Thus the same neuron in the output layer has learned to respond to the two input patterns that have similar vector elements in common. As can be seen, the process can be continued for subsequent shifts, provided that a sufficient proportion of input cells stay active between individual shifts. (After Stringer, Rolls, Perry and Proske 2006.)

training is not crucial for learning to occur; that relatively large numbers of transforms can be learned; and that such continuous spatial transformation learning can be usefully combined with temporal trace training.

4.5.12 Lighting invariance

Object recognition should occur correctly even despite variations of lighting. In an investigation of this, Rolls and Stringer (2006) trained VisNet on a set of 3D objects generated with OpenGL in which the viewing angle and lighting source could be independently varied (see Fig. 4.69). After training with the trace rule on all the 180 views (separated by 1 deg, and rotated about the vertical axis in Fig. 4.69) of each of the four objects under the left lighting condition, we tested whether the network would recognize the objects correctly when they were shown again, but with the source of the lighting moved to the right so that the objects appeared different (see Fig. 4.69). Figure 4.70 shows that the single and multiple cell inform-
Lighting invariance. VisNet was trained on a set of 3D objects (cube, tetrahedron, octahedron and torus) generated with OpenGL in which for training the objects had left lighting, and for testing the objects had right lighting. Just one view of each object is shown in the Figure, but for training and testing 180 views of each object separated by 1 deg were used. (After Rolls and Stringer 2006.)

![Lighting invariance](image)

**Fig. 4.69** Lighting invariance. The performance of the network after training with 180 views of each object lit from the left, when tested with the lighting again from the left (Left-light), and when tested with the lighting from the right (Right-light). The single cell information measure shows that many single neurons in layer 4 had the maximum amount of information about the objects, 2 bits, which indicates that they responded to all 180 views of one of the objects, and none of the 180 views of the other objects. The multiple cell information shows that the cells were sufficiently different in the objects to which they responded invariantly that all of the objects were perfectly represented when tested with the training images, and very well represented (with nearly 2 bits of information) when tested in the untrained lighting condition. (After Rolls and Stringer 2006.)

Some insight into the good performance with a change of lighting is that some neurons in the inferior temporal visual cortex respond to the outlines of 3D objects (Vogels and Biederman 2002), and these outlines will be relatively consistent across lighting variations. Although the features about the object represented in VisNet will include more than the representations of the outlines, the network may because it uses distributed representations of each object generalize correctly provided that some of the features are similar to those present during training. Under very difficult lighting conditions, it is likely that the performance of the network could be improved by including variations in the lighting during training, so that the trace rule could help to build representations that are explicitly invariant with respect to lighting.
4.5.13 Invariant global motion in the dorsal visual system

A key issue in understanding the cortical mechanisms that underlie motion perception is how we perceive the motion of objects such as a rotating wheel invariantly with respect to position on the retina, and size. For example, we perceive the wheel shown in Fig. 4.71a rotating clockwise independently of its position on the retina. This occurs even though the local motion for the wheels in the different positions may be opposite. How could this invariance of the visual motion perception of objects arise in the visual system? Invariant motion representations are known to be developed in the cortical dorsal visual system. Motion-sensitive neurons in V1 have small receptive fields (in the range 1–2 deg at the fovea), and can therefore not detect global motion, and this is part of the aperture problem (Wurtz and Kandel 2000b). Neurons in MT, which receives inputs from V1 and V2 (see Fig. 1.10), have larger receptive fields (e.g. 5 degrees at the fovea), and are able to respond to planar global motion, such as a field of small dots in which the majority (in practice as few as 55%) move in one direction, or to the overall direction of a moving plaid, the orthogonal grating components of which have motion at 45 degrees to the overall motion (Movshon, Adelson, Gizzi and Newsome 1985, Newsome, Britten and Movshon 1989). Further on in the dorsal visual system, some neurons in macaque visual area MST (but not MT) respond to rotating flow fields or looming with considerable translation invariance (Graziano, Andersen and Snowden 1994, Geesaman and Andersen 1996). In the cortex in the anterior part of the superior temporal sulcus, which is a convergence zone for inputs from the ventral and dorsal visual systems, some neurons respond to object-based motion, for example to a head rotating clockwise but not anticlockwise, independently of whether the head is upright or inverted which reverses the optic flow across the retina (Hasselmo, Rolls, Baylis and Nalwa 1989b).

In a unifying hypothesis with the design of the ventral cortical visual system, Rolls and Stringer (2007) proposed that the dorsal visual system uses a hierarchical feedforward network architecture (V1, V2, MT, MSTd, parietal cortex) with training of the connections with a short-term memory trace associative synaptic modification rule to capture what is invariant at each stage. The principle is illustrated in Fig. 4.71a. Simulations showed that the proposal is computationally feasible, in that invariant representations of the motion flow fields produced by objects self-organize in the later layers of the architecture (see examples in Fig. 4.71b–e). The model produces invariant representations of the motion flow fields produced by global in-plane motion of an object, in-plane rotational motion, looming vs receding of the object. The model also produces invariant representations of object-based rotation about a principal axis, of the type illustrated in Fig. 4.9 on page 272. Thus it is proposed that the dorsal and ventral visual systems may share some unifying computational principles (Rolls and Stringer 2007). Indeed, the simulations of Rolls and Stringer (2007) used a standard version of VisNet, with the exception that instead of using oriented bar receptive fields as the input to the first layer, local motion flow fields provided the inputs.

4.6 Further approaches to invariant object recognition

A related approach to invariant object recognition is described by Riesenhuber and Poggio (1999b), and builds on the hypothesis that not just shift invariance (as implemented in the Neocognitron of Fukushima (1980)), but also other invariances such as scale, rotation and even view, could be built into a feature hierarchy system, as suggested by Rolls (1992a) and incorporated into VisNet (Wallis, Rolls and Foldiak 1993, Wallis and Rolls 1997, Rolls and Milward 2000, Rolls and Stringer 2007) (see also Perrett and Oram (1993)). The approach of Riesenhuber and Poggio (1999b) (see also Riesenhuber and Poggio (1999a), Riesenhuber and Poggio (2000) and Serre, Wold, Bileschi, Riesenhuber and Poggio (2007)) is a feature
Fig. 4.71 (a) Two rotating wheels at different locations rotating in opposite directions. The local flow field is ambiguous. Clockwise or counterclockwise rotation can only be diagnosed by a global flow computation, and it is shown how the network is expected to solve the problem to produce position invariant global motion-sensitive neurons. One rotating wheel is presented at any one time, but the need is to develop a representation of the fact that in the case shown the rotating flow field is always clockwise, independently of the location of the flow field. (b–d) Translation invariance, with training on 9 locations. (b) Single cell information measures showing that some layer 4 neurons have perfect performance of 1 bit (clockwise vs anticlockwise) after training with the trace rule, but not with random initial synaptic weights in the untrained control condition. (c) The multiple cell information measure shows that small groups of neurons have perfect performance. (d) Position invariance illustrated for a single cell from layer 4, which responded only to the clockwise rotation, and for every one of the 9 positions. (e) Size invariance illustrated for a single cell from layer 4, which after training with three different radii of rotating wheel, responded only to anticlockwise rotation, independently of the size of the rotating wheels. (After Rolls and Stringer 2007.)
Further approaches to invariant object recognition

Fig. 4.72 Sketch of Riesenhuber and Poggio's (1999a,b) model of invariant object recognition. The model includes layers of ‘S’ cells which perform template matching (solid lines), and ‘C’ cells (solid lines) which pool information by a non-linear MAX function to achieve invariance (see text). (After Riesenhuber and Poggio 1999a, b.)

hierarchy approach that uses alternate ‘simple cell’ and ‘complex cell’ layers in a way analogous to Fukushima (1980) (see Fig. 4.72). The function of each S cell layer is to build more complicated features from the inputs, and works by template matching. The function of each ‘C’ cell layer is to provide some translation invariance over the features discovered in the preceding simple cell layer (as in Fukushima (1980)), and operates by performing a MAX function on the inputs. The non-linear MAX function makes a complex cell respond only to whatever is the highest activity input being received, and is part of the process by which invariance is achieved according to this proposal. This C layer process involves ‘implicitly scanning over afferents of the same type differing in the parameter of the transformation to which responses should be invariant (for instance, feature size for scale invariance), and then selecting the best-matching afferent’ (Riesenhuber and Poggio 1999b). Brain mechanisms by which this computation could be set up are not part of the scheme, and the model is not fully self-organizing, so does not yet provide a biologically plausible model of invariant object recognition. However, the fact that the model sets out to achieve some of the processes specified by Rolls (1992a) (see Section 4.4) and implemented in VisNet (see Section 4.5) does represent useful convergent thinking towards how invariant object recognition may be implemented in the brain. Similarly, the approach of training a five-layer network with a more artificial gradient ascent approach with a goal function that does however include forming relatively time invariant representations and decorrelating the responses of neurons within each layer (Wyss, Konig and Verschure 2006) (both processes that have their counterpart in VisNet), also reflects convergent thinking.
Further evidence consistent with the approach developed in the investigations of VisNet described in this chapter comes from psychophysical studies. Wallis and Bülthoff (1999) and Perry, Rolls and Stringer (2006) describe psychophysical evidence for learning of view invariant representations by experience, in that the learning can be shown in special circumstances to be affected by the temporal sequence in which different views of objects are seen.

Another approach to the implementation of invariant representations in the brain is the use of neurons with Sigma-Pi synapses. Sigma-Pi synapses, described in Section A.2.3, effectively allow one input to a synapse to be multiplied or gated by a second input to the synapse. The multiplying input might gate the appropriate set of the other inputs to a synapse to produce the shift or scale change required. For example, the $x^c$ input in equation A.14 could be a signal that varies with the shift required to compute translation invariance, effectively mapping the appropriate set of $x_j$ inputs through to the output neurons depending on the shift required (Mel, Ruderman and Archie 1998, Mel and Fiser 2000, Olshausen, Anderson and van Essen 1993, Olshausen, Anderson and Van Essen 1995). Local operations on a dendrite could be involved in such a process (Mel, Ruderman and Archie 1998). The explicit neural implementation of the gating mechanism seems implausible, given the need to multiply and thus remap large parts of the retinal input depending on shift and scale modifying connections to a particular set of output neurons. Moreover, the explicit control signal to set the multiplication required in V1 has not been identified. Moreover, if this was the solution used by the brain, the whole problem of shift and scale invariance could in principle be solved in one layer of the system, rather than with the multiple hierarchically organized set of layers actually used in the brain, as shown schematically in Fig. 4.2. The multiple layers actually used in the brain are much more consistent with the type of scheme incorporated in VisNet. Moreover, if a multiplying system of the type hypothesized by Mel, Ruderman and Archie (1998), Olshausen, Anderson and Van Essen (1993) and Olshausen, Anderson and Van Essen (1995) was implemented in a multilayer hierarchy with the shift and scale change emerging gradually, then the multiplying control signal would need to be supplied to every stage of the hierarchy. A further problem with such approaches is how the system is trained in the first place.

4.7 Visuo-spatial scratchpad memory, and change blindness

Given the fact that the responses of inferior temporal cortex neurons are quite locked to the stimulus being viewed, it is unlikely that IT provides the representation of the visual world that we think we see, with objects at their correct places in a visual scene. In fact, we do not really see the whole visual scene, as most of it is a memory reflecting what was seen when the eyes were last looking at a particular part of the scene. The evidence for this statement comes from change blindness experiments, which show that humans rather remarkably do not notice if, while they are moving their eyes and cannot respond to changes in the visual scene, a part of the scene changes (O’Regan, Rensink and Clark 1999, Rensink 2000). A famous example is that in which a baby was removed from the mother’s arms during the subject’s eye movement, and the subject failed to notice that the scene was any different. Similarly, unless we are fixating the location that is different in two alternated versions of a visual scene, we are remarkably insensitive to differences in the scene, such as a glass being present on a dining table in one but not another picture of a dining room. Given then that much of the apparent richness of our visual world is actually based on what was seen at previously fixated positions in the scene (with this being what the inferior temporal visual cortex represents), we may ask where this ‘visuo-spatial’ scratchpad (short-term memory) is located in the brain. One
possibility is in the right parieto-occipital area, for patients with lesions in the parieto-occipital region have (dorsal) simultanagnosia, in which they can recognize objects, but cannot see more than one object at a time (Farah 2004). Alternatively, the network could be in the left inferior temporo-occipital regions, for patients with (ventral) simultanagnosia cannot recognize but can see multiple objects in a scene (Farah 2004).

The computational basis for this could be a number of separate, that is local, attractors each representing part of the space, and capable of being loaded by inputs from the inferior temporal visual cortex. According to this computational model, the particular attractor network in the visuo-spatial scratchpad memory would be addressed by information based on the position of the eyes, of covert attention, and probably the position of the head. While being so addressed, inputs to the scratch-pad from the inferior temporal cortex neurons would then provide information about the object at the fovea, together with some information about the location of objects in different parfoveal regions (see Section 4.5.10), and this would then enable object information to be associated by synaptic modification with the active neurons in the attractor network representing that location (see Fig. 4.73). Because there are separate spatially local attractors for each location, each of the attractors with its associated object information could be kept active simultaneously, to maintain for short periods information about the relative spatial position of multiple objects in a scene. The attractor for each spatial location would need to represent some information about the object present at that location (as otherwise a binding problem would arise between the multiple objects and the
multiple locations), and this may be a reason why the object information in the visuo-spatial scratchpad is not detailed. The suggestion of different attractors for different regions of the scene is consistent with the architecture of the cerebral neocortex, in which the high density of connections including those of inhibitory interneurons is mainly local, within a range of 1–2 mm (see Section 1.10). (If the inhibitory connections were more widespread, so that the inhibition became more global within the visuo-spatial scratchpad memory, the system would be more like a continuous attractor network with multiple activity packets (Stringer, Rolls and Trappenberg 2004) (see Section B.5.4).)

It may only be when the inferior temporal visual cortex is representing objects at or close to the fovea in a complex scene that a great deal of information is present about an object, given the competitive processes described in this chapter combined with the large cortical magnification factor of the fovea. The perceptual system may be built in this way for the fundamental reasons described in this chapter (which enable a hierarchical competitive system to learn invariant representations that require feature binding), which account for why it cannot process a whole scene simultaneously. The visual system can, given the way in which it is built, thereby give an output in a form that is very useful for memory systems, but mainly about one or a few objects close to the fovea.

4.8 Different processes involved in different types of object identification

To conclude this chapter, it is proposed that there are (at least) three different types of process that could be involved in object identification. The first is the simple situation where different objects can be distinguished by different non-overlapping sets of features (see Section 4.3.1). An example might be a banana and an orange, where the list of features of the banana might include yellow, elongated, and smooth surface; and of the orange its orange colour, round shape, and dimpled surface. Such objects could be distinguished just on the basis of a list of the properties, which could be processed appropriately by a competitive network, pattern associator, etc. No special mechanism is needed for view invariance, because the list of properties is very similar from most viewing angles. Object recognition of this type may be common in animals, especially those with visual systems less developed than those of primates. However, this approach does not describe the shape and form of objects, and is insufficient to account for primate vision. Nevertheless, the features present in objects are valuable cues to object identity, and are naturally incorporated into the feature hierarchy approach.

A second type of process might involve the ability to generalize across a small range of views of an object, that is within a generic view, where cues of the first type cannot be used to solve the problem. An example might be generalization across a range of views of a cup when looking into the cup, from just above the near lip until the bottom inside of the cup comes into view. This type of process includes the learning of the transforms of the surface markings on 3D objects which occur when the object is rotated, as described in Section 4.5.7. Such generalization would work because the neurons are tuned as filters to accept a range of variation of the input within parameters such as relative size and orientation of the components of the features. Generalization of this type would not be expected to work when there is a catastrophic change in the features visible, as for example occurs when the cup is rotated so that one can suddenly no longer see inside it, and the outside bottom of the cup comes into view.

The third type of process is one that can deal with the sudden catastrophic change in the features visible when an object is rotated to a completely different view, as in the cup example
just given (cf. Koenderink (1990)). Another example, quite extreme to illustrate the point, might be when a card with different images on its two sides is rotated so that one face and then the other is in view. This makes the point that this third type of process may involve arbitrary pairwise association learning, to learn which features and views are different aspects of the same object. Another example occurs when only some parts of an object are visible. For example, a red-handled screwdriver may be recognized either from its round red handle, or from its elongated silver-coloured blade.

The full view-invariant recognition of objects that occurs even when the objects share the same features, such as colour, texture, etc. is an especially computationally demanding task which the primate visual system is able to perform with its highly developed temporal lobe cortical visual areas. The neurophysiological evidence and the neuronal networks described in this chapter provide clear hypotheses about how the primate visual system may perform this task.

4.9 Conclusions

We have seen in this chapter that the feature hierarchy approach has a number of advantages in performing object recognition over other approaches (see Section 4.3), and that some of the key computational issues that arise in these architectures have solutions (see Sections 4.4 and 4.5). The neurophysiological and computational approach taken here focuses on a feature hierarchy model in which invariant representations can be built by self-organizing learning based on the statistics of the visual input.

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 Spatial Transformation learning.

The model of visual processing in the ventral cortical stream can build representations of objects that are invariant with respect to translation, view, size, and lighting.

The model uses a feature combination neuron approach with the relative spatial positions of the objects specified in the feature combination neurons, and this provides a solution to the binding problem.

The model has been extended to provide an account of invariant representations in the dorsal visual system of the global motion produced by objects such as looming, rotation, and object-based movement.

The model has been extended to incorporate top-down feedback connections to model the control of attention by biased competition in for example spatial and object search tasks (see further Chapter 6).

The model has also been extended to account for how the visual system can select single objects in complex visual scenes, how multiple objects can be represented in a scene, and how invariant representations of single objects can be learned even when multiple objects are present in the scene.

It has also been suggested in a unifying proposal that adding a fifth layer to the model and training the system in spatial environments will enable hippocampus-like spatial view neurons or place cells to develop, depending on the size of the field of view (Section 2.3.5).

We have thus seen how many of the major computational issues that arise when formulating a theory of object recognition in the ventral visual system (such as feature binding, invariance learning, the recognition of objects when they are in cluttered natural scenes, the representation of multiple objects in a scene, and learning invariant representations of single objects when there are multiple objects in the scene), could be solved in the brain, with tests
of the hypotheses performed by simulations that are consistent with complementary neurophysiological results.

The approach described in this chapter is unifying in a number of ways. First, a set of simple organizational principles involving a hierarchy of cortical areas with convergence from stage to stage, and competitive learning using a modified associative learning rule with a short-term memory trace of preceding neuronal activity, provide a basis for understanding much processing in the ventral visual stream, from V1 to the inferior temporal visual cortex.

Second, the same principles help to understand some of the processing in the dorsal visual stream by which invariant representations of the global motion of objects may be formed. Third, the same principles continued from the ventral visual stream onwards to the hippocampus help to show how spatial view and place representations may be built from the visual input. Fourth, in all these cases, the learning is possible because the system is able to extract invariant representations because it can utilize the spatio-temporal continuities and statistics in the world that help to define objects, moving objects, and spatial scenes. Fifth, a great simplification and economy in terms of brain design is that the computational principles need not be different in each of the cortical areas in these hierarchical systems, for some of the important properties of the processing in these systems to be performed.

In conclusion, we have seen in this chapter how a major form of memory, the invariant recognition of objects, involves not only the storage and retrieval of information, but also major computations to produce invariant representations. Once these invariant representations have been formed, they are used for many processes including not only recognition memory (see Section 2.2.6), but also associative learning of the rewarding and punishing properties of objects for emotion and motivation (see Chapter 3), the memory for the spatial locations of objects and rewards (see Chapter 2), the building of spatial representations based on visual input (Section 2.3.5), and as an input to short-term memory (Chapter 5), attention (Chapter 6), decision (Chapters 7 and 10), and action selection (Chapter 8) systems.