SEARCH AND ATTENTION FOR MACHINE VISION

A THESIS
SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

KEVIN BROHAN

SCHOOL OF ELECTRICAL AND ELECTRONIC ENGINEERING
2012
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>11</td>
</tr>
<tr>
<td>Declaration</td>
<td>12</td>
</tr>
<tr>
<td>Copyright</td>
<td>13</td>
</tr>
<tr>
<td>Acknowledgement</td>
<td>14</td>
</tr>
<tr>
<td>Glossary</td>
<td>15</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>18</td>
</tr>
<tr>
<td><strong>1 Introduction</strong></td>
<td>22</td>
</tr>
<tr>
<td>1.1 Organisation and Summary</td>
<td>23</td>
</tr>
<tr>
<td>1.2 Novel Work &amp; Relevance</td>
<td>24</td>
</tr>
<tr>
<td><strong>2 Background</strong></td>
<td>26</td>
</tr>
<tr>
<td>2.1 Search, Reward Harvesting and Attention</td>
<td>26</td>
</tr>
<tr>
<td>2.2 Anatomical Areas Associated with Visual Search</td>
<td>30</td>
</tr>
<tr>
<td>2.2.1 Early Processing</td>
<td>30</td>
</tr>
<tr>
<td>2.2.2 The Ventral Pathway</td>
<td>33</td>
</tr>
<tr>
<td>2.2.3 The Dorsal Pathway</td>
<td>33</td>
</tr>
<tr>
<td>2.2.4 Working Memory</td>
<td>34</td>
</tr>
<tr>
<td>2.2.5 Gaze Control</td>
<td>35</td>
</tr>
<tr>
<td>2.2.6 Inhibition of Return</td>
<td>35</td>
</tr>
<tr>
<td>2.2.7 Reward and the Basal Ganglia</td>
<td>36</td>
</tr>
<tr>
<td>2.2.8 Summary</td>
<td>36</td>
</tr>
<tr>
<td>2.3 Selection and Attention</td>
<td>38</td>
</tr>
<tr>
<td>2.3.1 Selection</td>
<td>38</td>
</tr>
<tr>
<td>2.3.2 Percepts and Selection</td>
<td>40</td>
</tr>
<tr>
<td>2.3.3 Neural Signatures of Attention and Object Recognition</td>
<td>43</td>
</tr>
<tr>
<td>2.4 Theoretical Models of Visual Attention</td>
<td>44</td>
</tr>
<tr>
<td>2.4.1 Feature Integration Theory</td>
<td>44</td>
</tr>
<tr>
<td>2.4.2 Biased Competition</td>
<td>45</td>
</tr>
<tr>
<td>2.4.3 Guided Search</td>
<td>46</td>
</tr>
<tr>
<td>2.5 Computational Models of Attention and Object Recognition</td>
<td>47</td>
</tr>
<tr>
<td>2.5.1 Selection</td>
<td>48</td>
</tr>
<tr>
<td>2.5.2 Segmentation</td>
<td>50</td>
</tr>
<tr>
<td>2.5.3 Object Recognition</td>
<td>51</td>
</tr>
</tbody>
</table>
2.6 Visual Attention for Behaving Machines ........................................... 52
2.7 Self-Organising Maps ........................................................................ 54
2.8 Summary ......................................................................................... 55

3 Framework and Implementations ......................................................... 57
  3.1 Introduction: Framework for Guidance, Attention and Learning .......... 57
  3.2 Summary of Implemented Systems .................................................. 57
  3.3 Structure ....................................................................................... 60
  3.4 Implementation Details .................................................................. 62

4 System I: Visual Search ................................................................. 63
  4.1 Introduction .................................................................................... 63
  4.2 System Description ....................................................................... 64
    4.2.1 Overview ................................................................................ 64
    4.2.2 Low-level Feature Extraction ................................................ 65
    4.2.3 Selection ................................................................................ 68
    4.2.4 Attention and Learning .......................................................... 70
  4.3 Results: System Ia ....................................................................... 73
    4.3.1 Implementation Details .......................................................... 74
      Experiment I .................................................................................. 74
    4.3.2 Experiment II ......................................................................... 76
    4.3.3 Experiment III ........................................................................ 79
  4.4 Results: System Ib ....................................................................... 84
    4.4.1 Implementation Details .......................................................... 84
    4.4.2 Experiment IV ....................................................................... 85
  4.5 Discussion ...................................................................................... 89

5 System II: Object Occlusion in Selection and Recognition ............... 91
  5.1 Introduction .................................................................................... 91
  5.2 System Description ....................................................................... 92
  5.3 Results .......................................................................................... 101
    5.3.1 Experiment Ia ....................................................................... 102
    5.3.2 Experiment Ib ....................................................................... 105
    5.3.3 Experiment II ........................................................................ 107
  5.4 Discussion ...................................................................................... 112

6 System III: Shape-Based Guidance of Attention ..................................... 114
  6.1 Introduction .................................................................................... 114
  6.2 System Description ....................................................................... 115
  6.3 Results .......................................................................................... 120
    6.3.1 Experiment I .......................................................................... 120
    6.3.2 Experiment II ........................................................................ 124
    6.3.3 Experiment III ....................................................................... 126
  6.4 Discussion ...................................................................................... 130

7 Discussion & Conclusion ................................................................. 132
  7.1 Summary ....................................................................................... 132
  7.2 Comparison of Models with Previous Work .................................... 133
  7.3 Contributions ............................................................................... 135
    7.3.1 Self-Organising Internal Representations .................................. 135
    7.3.2 Reward Harvesting with Limited Resources ............................. 136
2.1 An active visual search task. An 'eye' views a region of the entire visual scene, which contains cues on an invisible background. Different targets compete for attention, and gaze is directed towards the selected target.

2.2 Functional stages of the attention and object recognition processes. See text for details.

2.3 Schematic of locations of the anatomical regions which subserve visual attention and object recognition in the human brain. Shown connections relate to the functional stages of processing in Fig. 2.2. Pale grey labels indicate subcortical regions and cortical areas on the inside surface (i.e. along the parasagittal plane) of the brain and dark grey labels indicate areas along the outer surface. Image modified from Fig. 726, (Gray, 1918)

2.4 Example of a phenomenal T-junction providing a cue by which surface order can be inferred: square A is occluded by square B. The line ik belongs to the closer object while the terminated line hj belongs to the occluded object.

2.5 Figure taken from Rensink & Enns (1998). Search for the notched square was rapid in 1A, due to its distinct shape. In 1B, the search became much slower because of the early completion of the occluded region behind the white circle: the distinct notched feature was no longer present, which suggests that the black figure was perceived as a completed square.

2.6 Diagram from Neisser and Becklen (1975) in which the authors demonstrate that attention could select one of the overlapping objects. Subjects were instructed to count the number of times the hands clap, which severely impaired their ability to report changes in the ball game scene.

2.7 Illustration of perceptual grouping; (a) the factor of proximity causes the dots to be grouped into pairs where possible; (b) demonstration of a hierarchical grouping effect: the factor of similarity overrules the factor of proximity. Images from Wertheimer (1923).

3.0 Schematic of the proposed framework for visual attention; (a) System Ia; (b) System Ib; (c) System II; (d) System III.

4.1 Network diagram for System I, see text for details. Maps beginning with R were retinotopic, with S were world maps, and with F were in featural coordinates, with the exception of F^{SOM}, which was of dimensions 768 x 768 pixels.

4.2 Image of the robot and cues in System Ib; (a) photograph with labels; (b) illustration of experimental setup.
4.3 Illustration of the process of generating vector $C_i$ from $R^0$ and $R^{90}$. The distance $d$ describes the scale of the feature detectors, which was hand-selected for the presented stimuli. ................................. 68

4.4 Use of reward prediction error to modulate $\alpha_t$ and $\sigma_t$ over time. (a) $\delta$ was the reward prediction error; (b) $y$ was a non-linear function of the error, which filtered out small errors; (c) $\xi$ was the leaky integral of $y$ over time, the dashed line marks the reset threshold, $T_\xi = 40$; (d) $\sigma/\sigma_0$ was the width of the neighbourhood function used during SOM training. ................................. 73

4.5 Colours used as stimuli in System Ia, and their locations in RGB space. 74

4.6 Effect of the number of discrete stimuli $N$ and the size of the reward window on the fraction of rewards harvested; (a) online learning of SOM weights; (b) no learning of SOM weights. ................................. 76

4.7 Fraction of harvested rewards as a function of $\gamma_L$ for different values of $A$ and $\sigma_L$: (a) $A = 0$; (b) $A = 0.25$; (c) $A = 0.5$; (d) $A = 0.75$; (e) $A = 1$ 78

4.8 Number of saccades to each colour over 10000 time steps; (a) working memory & modulated SOM; (b) working memory & static SOM; (c) no working memory & static SOM ................................. 81

4.9 Saccades to each colour in bins of 100 time steps; colour indicates the fraction of saccades made to each colour index. In the control experiment, the rate of selection was similar for all cues. ................................. 82

4.10 Examples of modulated SOMs preferentially representing rewarding regions of input space; (a) yellow, Epoch 1; (b) green, Epoch 2; (c) blue, Epoch 3; (d) magenta, Epoch 4 ................................. 83

4.11 The total number of rewards received over time. The error bars mark one standard deviation across 10 trials. For clarity, every 250th point has been plotted ................................. 83

4.12 Number of selections as a function of the Euclidean distance between the colour of the selected cue and a reward in RGB space ................................. 84

4.13 Image of the digit representations at each of the 16 SOM nodes after training. The grayscale intensity of the bars indicates the magnitude of the weight for each dimension of the feature detector. The subscript shows the node index. ................................. 85

4.14 The cumulative number of saccades made to each digit over time ................................. 86

4.15 The $F^M$ activity for each neuron over time for each SOM position ................................. 87

4.16 The fraction of saccades which activated each region on the SOM; (a) Epoch 1: no working memory; (b) reward for saccades to the digit 2; (c) reward for saccades to digit 6 ................................. 88

4.17 Normalised Euclidean distances in RGB space to each point on the SOM for ideal stimuli. This provides an indication of the quality of the representation of a digit on the SOM. (a) Digit 2; (b) Digit 3; (c) Digit 4; (d) Digit 5; (e) Digit 6 ................................. 88

5.1 Overview of System II. Unimplemented connections were shown with dashed lines ................................. 92
5.2 Illustration of the perceptual grouping process; (a) the stimulus consisted of a trapezoid, which occluded a rectangle. The scene was initially parsed into three surfaces $S_1$, $S_2$, $S_3$; (b) ownership was assigned to the borders of each surface, based on the orientation of T-junctions (circles). (c) By extending the terminated lines of the T-junctions and searching for intersections, $S_1$ and $S_3$ were linked as a single percept to form two new surfaces, $S'_1$ and $S'_2$. (d) The borders of the final surfaces. Unowned borders are shown in grey.

5.3 Illustration of the grouping mechanism, which was used to unite surfaces into single percepts. Stopped lines were extended beyond the occluding border (light grey). (a) No intersection of the extended lines; (b) area of the intersection (dark grey) was small for perpendicular lines; (c) as the angle approaches $180^\circ$, the area of the intersection increased; (d) the largest common area was obtained for the apparent continuation of a line behind an occluder. Ellipses were shown for clarity in this figure: diffuse projections were used in the implemented system.

5.4 Grouping and border ownership assignment for the three test conditions. The leftmost column (1) shows the stimulus and the other figures in each of the rows (2 – 5) show the border ownership for each of the figures prior to selection. Positive activity indicated that border ownership was assigned to the attended surface and negative activity indicated that border ownership was not assigned to the attended surface; (A): border ownership when an occluder was present; (B) border ownership when no occluder was present; (C) border ownership for an unambiguous arrangement.

5.5 Activity of each recognition neuron ($I_l$) when each surface was attended (indicated by trace colour). The shapes to which the recognition neurons were trained are shown at the top of the figure; A: response of each neuron to each shape when border ownership information was used; B: when border ownership information was not used; C: unambiguous control case.

5.6 Activity of the triangle recognition neuron (black) and the square recognition neuron (gray) while the triangle in Table 5.4 was attended; recognition neuron activity when border ownership information was used (BO) and when it was not used (NBO) during the recognition process. (a) $\Delta O = 0$; (b) $\Delta O = 0.6$.

5.7 Ratio of the activity of the triangle recognition neuron to the activity of the square recognition neuron for each shape in Table 5.4 when border ownership information was used (circle) and when border ownership information was not used (square). (a) $\Delta O = 0$; (b) $\Delta O = 0.6$.

5.8 Fraction of correct recognitions as a function of the number of local features used during recognition. Trace colour indicates the number of objects used in the experiment. (a) $\Delta O = 0.5$: a contribution of 50% activity was made by the occlusion neurons for potentially missing features; (b) $\Delta O = 0$: the occlusion neurons had no effect on the recognition process.

5.9 Ratio of recognition neuron responses (border ownership vs. no border ownership); (a) for different numbers of local features; (b) more explicit comparison of (a) for 6 and 7 local features.

6.1 Flow diagram for System III, see text for details.

6.2 Black dots indicate the locations of the points in $d_j$ on a pair of stimuli consisting of a rectangle and a notched rectangle.
6.3 First set of objects; (a) the arrangement of the objects; (b) normalised saliency of each object in (a) .................................................. 121
6.4 Second set of objects; (a) the arrangement of the objects; (b) normalised saliency of each object in (a) .................................................. 123
6.5 Stimulus and priorities for Experiment II; (a) stimulus image with partially visible shapes at the edge. Numbers indicate the order of the objects in (b); (b) saliency of each object ............................................. 125
6.6 Visual world in Experiment III. Cues consisted of squares and notched squares, with occluding circles ........................................... 127
6.7 Plot of the total number of saccades to the notched square and the square for each vase; the system was biased to select the notched square in cases 1 & 2, and the square in cases 3 & 4 ............................................. 128
6.8 Images of the scan paths for cases 1 and 2; (a) when border ownership information was used; (b) when border ownership information was not used129

A.1 Graph showing the effect of the resolution of the colour space and the size of the reward window on harvesting for different values of N when \( B = 0 \), no online learning of SOM weights ........................................... 143
A.2 Reward harvesting rate for \( F = 0 \) .................................................. 144
A.3 Reward harvesting rate for \( F = 1 \) .................................................. 145
A.4 Reward harvesting rate for \( F = 2 \) .................................................. 146
A.5 Reward harvesting rate for \( F = 3 \) .................................................. 147
A.6 Reward harvesting rate for \( F = 4 \) .................................................. 148
A.7 The cumulative number of rewards received over time .......................... 149
A.8 Reward harvesting rate for no changes in target, \( A = B \); (a) \( \gamma = 1 \); (b) \( \gamma = 0.95 \); (c) \( \gamma = 0.9 \); (d) \( \gamma = 0.85 \); (e) \( \gamma = 0.8 \); ............................................. 150
A.9 Reward harvesting rate for 1 change in target, \( A = B \); (a) \( \gamma = 1 \); (b) \( \gamma = 0.95 \); (c) \( \gamma = 0.9 \); (d) \( \gamma = 0.85 \); (e) \( \gamma = 0.8 \); ............................................. 151
A.10 Reward harvesting rate for 2 changes in target, \( A = B \); (a) \( \gamma = 1 \); (b) \( \gamma = 0.95 \); (c) \( \gamma = 0.9 \); (d) \( \gamma = 0.85 \); (e) \( \gamma = 0.8 \); ............................................. 152
A.11 Reward harvesting rate for 3 changes in target, \( A = B \); (a) \( \gamma = 1 \); (b) \( \gamma = 0.95 \); (c) \( \gamma = 0.9 \); (d) \( \gamma = 0.85 \); (e) \( \gamma = 0.8 \); ............................................. 153
A.12 Reward harvesting rate for 4 changes in target, \( A = B \); (a) \( \gamma = 1 \); (b) \( \gamma = 0.95 \); (c) \( \gamma = 0.9 \); (d) \( \gamma = 0.85 \); (e) \( \gamma = 0.8 \); ............................................. 154
# LIST OF TABLES

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Summary of the parameters and their default values for System I</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>Summary of the parameters and their default values for System II</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>Summary of the parameters and their default values for System III</td>
<td>21</td>
</tr>
<tr>
<td>3.1</td>
<td>Summary of the framework implementations in this thesis</td>
<td>61</td>
</tr>
<tr>
<td>4.1</td>
<td>Key to symbols in Fig. 4.1. The dimensions of each neural layer are shown.</td>
<td>66</td>
</tr>
<tr>
<td>5.1</td>
<td>Features and weights ($G$ &amp; $W$) for equations 5.1 and 5.2</td>
<td>95</td>
</tr>
<tr>
<td>5.2</td>
<td>Example of the recognition process; (a) the stimulus consisted of two identical squares, one of which was partially occluded by the other; (b) feature detectors were tuned to recognise the four corners; (c) learned positions from which local feature detector activity was pooled during recognition. The middle row shows the recognition process when the upper surface (d) was attended; (e) black dots indicate peaks of activity at the locations where each of the features had been identified; (f) activity was pooled from the learned relative locations in (c) by a recognition neuron; separate feature detectors and recognition neurons existed for different objects. The bottom row shows the recognition process when the lower surface was attended (g); (h) when the lower square was attended the black dots indicate locations where features had been identified; (i) any feature which was expected within the region occupied by the upper square made a weak contribution to the recognition process</td>
<td>101</td>
</tr>
<tr>
<td>5.3</td>
<td>Stimuli and local features used in Experiment Ia. Side image: combined rectangle, trapezoid and small triangle</td>
<td>102</td>
</tr>
<tr>
<td>5.4</td>
<td>Stimuli used in Experiment Ib</td>
<td>105</td>
</tr>
<tr>
<td>5.5</td>
<td>Local feature positions in Experiment Ib</td>
<td>105</td>
</tr>
<tr>
<td>5.6</td>
<td>Objects (white) and local feature positions (black) used in Experiment II.</td>
<td>108</td>
</tr>
<tr>
<td>5.7</td>
<td>Examples of stimuli and the borders in Experiment II. White borders were owned and black borders were unowned by the attended surface. Discounted stimuli are shown in the bottom row</td>
<td>109</td>
</tr>
<tr>
<td>6.1</td>
<td>Working memory maps different objects and distractors: positive working memory activity is shown in white and negative in black; (a) diamond target and notched diamond distractor; (b) notched diamond target and diamond distractor; (c) rectangle target and notched rectangle distractor; (d) notched rectangle target and rectangle distractor</td>
<td>117</td>
</tr>
</tbody>
</table>
6.2 The choice of origin for the mapping process depended upon the occluded regions of the surface; top row: images of a set of surfaces; second row: borders for each surface when border ownership information was used; unowned borders are shown in black; third row: position of the origin for a suitable map when border ownership information was used; fourth row: borders when no ownership information was used; fifth row: position of the origin when no border ownership information was used. The choice of origin depended upon the location of unowned borders on the surface; (a) the notched rectangle was unoccluded and owned all of its borders: the default axis (top-left) was chosen for mapping the shape onto the working memory in both ownership cases; (b) the rectangle appeared at the left edge of the visual field and the left border was unowned: the top-right was used as the origin for mapping the shape onto working memory in the case with border ownership, while the origin was incorrect placed when no border ownership information was used; (c) the upper half of the rectangle lay outside the visual field: the bottom-left axis was used as the origin for mapping to working memory when border ownership information was used. When no border ownership information was used, the surface was incorrectly mapped onto the working memory; (d) the right half of the rectangle lay outside the visual field: the top-left corner was used as the origin for mapping the shape to working memory in both cases.
SEARCH AND ATTENTION FOR MACHINE VISION
A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY
KEVIN BROHAN
JULY 24, 2012

This thesis addresses the generation of behaviourally useful, robust representations
of the sensory world in the context of machine vision and behaviour. The goals of the
work presented in this thesis are to investigate strategies for representing the visual world
in a way which is behaviourally useful, to investigate the use of a neurally inspired early
perceptual organisation system upon high-level processing in an object recognition system
and to investigate the use of a perceptual organisation system on driving an object-based
selection process.

To address these problems, a biologically inspired framework for machine attention
has been developed at a high level of neural abstraction, which has been heavily inspired
by the psychological and physiological literature. The framework is described in this the-
sis, and three system implementations, which investigate the above issues, are described
and analysed in detail.

The primate brain has access to a coherent representation of the external world, which
appears as objects at different spatial locations. It is through these representations that
appropriate behavioural responses may be generated. For example, we do not become
confused by cluttered scenes or by occluded objects.

The representation of the visual scene is generated in a hierarchical computing struc-
ture in the primate brain: while shape and position information are able to drive attentional
selection rapidly, high-level processes such as object recognition must be performed seri-
ally, passing through an attentional bottleneck.

Through the process of attentional selection, the primate visual system identifies be-
vaviourally relevant regions of the visual scene, which allows it to prioritise serial atten-
tional shifts towards certain locations. In primates, the process of attentional selection is
complex, operating upon surface representations which are robust to occlusion.

Attention itself suppresses neural activity related to distractor objects, while sustaining
activity relating to the target, allowing the target object to have a clear neural representa-
tion upon which the recognition process can operate.

This thesis concludes that dynamic representations that are both early and robust
against occlusion have the potential to be highly useful in machine vision and behaviour
applications.
DECLARATION

The University of Manchester
PhD Candidate Declaration

Candidate Name: Kevin Brohan

Faculty: Engineering and Physical Sciences

Thesis Title: Search and Attention for Machine Vision

Declaration to be completed by the candidate:

I declare that no portion of this work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Signed: Date: March 21, 2012
COPYRIGHT

The author of this thesis (including any appendices and/or schedules to this thesis) owns any copyright in it (the "Copyright")\(^1\) and s/he has given The University of Manchester the right to use such Copyright for any administrative, promotional, educational and/or teaching purposes.

Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the John Rylands University Library of Manchester. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.

The ownership of any patents, designs, trade marks and any and all other intellectual property rights except for the Copyright (the "Intellectual Property Rights") and any reproductions of copyright works, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.

Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on Presentation of Theses.

\(^1\)This excludes material already printed in academic journals, for which the copyright belongs to said journal and publisher. Pages for which the author does not own the copyright are numbered differently from the rest of the thesis.
I gratefully acknowledge the contribution of my supervisor, Piotr Dudek, to this work.

This research was funded in part by the School of Electrical and Electronic Engineering, and the EPSRC under the REVERB Project (EP/C516303) and as a Doctoral Training Grant.
Attention: the process which allows a clear representation of a percept access to higher cortical processing, such as object recognition. At a neural level, there is evidence that attention has the effect of suppressing neural activity relating to clutter in the scene, so that it appears as though the object was presented in isolation.

Covert Attention: attention may be shifted to a target without an explicit eye movement.

Feature: refers to a measurable property of the image, in a single modality (e.g. colour, form, orientation). In psychology, a subset of features, known as guiding features, can guide attentional selection. Examples of guiding features include orientation, shape and 3D appearance, while examples of non-guiding features include object identity and feature conjunctions (e.g. red and vertical). A feature is generally the attribute of a percept. In this thesis, the psychological sense of the word is used.

Object: the underlying entity which possesses the features. It is the binding of different features, such as shape, colour, and local form. The extent to which binding occurs prior to attention is still debated.

Object-based attention: theories of object-based attention maintain that the object is the fundamental unit to which attention is directed. All features within the object gain access to higher cortical processing once the object is attended.

Overt Attention: an attentional shift involving a saccade.
Saccade: a ballistic eye movement which requires a target. These often accompany an attentional shift. Other categories of eye movements include smooth pursuit and fixation (microsaccades).

Saliency: in this thesis, the term 'saliency' will be used to describe the combined contribution of bottom-up saliency (i.e., the statistical uniqueness of a region of the visual field, by some measure) and top-down behavioural goals. In this sense, the saliency of a target refers to the behavioural priority of shifting attention to that target.

Selection: the process of choosing an attentional target, which takes place prior to the deployment of attention. This is generally conceived of as a competitive process between different potential targets.

Spatial attention: describes the selection of a region of retinotopic space as the attended entity. Early models of attention use a spatial spotlight analogy in which a sub-window of the visual scene enters working memory. Theories of attention have become more object-based in recent years.

Basal Ganglia (BG): these sub-cortical areas are involved in the reinforcement learning of rewarding behaviours and in action selection.

Dorsolateral prefrontal cortex (dlPFC): this region subserves the control of working memory, which encodes behavioural goals. Once a target has been attended, its identity gains access to working memory.

Fourth visual area (V4): cells in this area are selective to simple shape and colour features and the topology is coarsely retinotopic.

Frontal Eye Fields (FEF): this area is involved in biasing V4 neuron selectivity towards a behavioural target.
**Lateral Geniculate Nucleus (LGN):** area of the thalamus involved in visual processing in primates. Axons from the retina project to the LGN and from there to primary visual cortex (V1).

**Lateral Intraparietal Area (LIP):** this area is believed to implement a visual saliency map in a retinotopic topology.

**Inferiortemporal Area IT:** these cells are selective to complex shape features, such as hands, faces or other objects. Object identity is understood to be encoded at a population activity level, which is strongly modulated by attention.

**Primary visual cortex (V1):** first cortical area associated with visual processing. Classically, V1 cells are tuned to orientation, which may be phase sensitive (simple cells) or may not be phase sensitive (complex cells). This area is organised retinotopically.

**Secondary visual cortex (V2):** classically, V2 cells have similar responses to V1, but with a greater proportion of complex cells. In recent decades, a large number of non-classical receptive fields have been found in V2, with selectivity towards border ownership and illusory contours. This area is organised retinotopically.

**Superior Colliculus (SC):** this sub-cortical structure is associated with the execution of saccades. The topology of the SC is retinotopic.

**Working Memory (WM):** provides a short-term buffer for storing information about behavioural goals. This function is associated with pre-frontal cortex (PFC).
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC</td>
<td>Biased competition</td>
</tr>
<tr>
<td>BG</td>
<td>Basal Ganglia</td>
</tr>
<tr>
<td>FEF</td>
<td>Frontal Eye Fields</td>
</tr>
<tr>
<td>FIT</td>
<td>Feature integration theory</td>
</tr>
<tr>
<td>GS</td>
<td>Guided search</td>
</tr>
<tr>
<td>IT</td>
<td>Inferior Temporal</td>
</tr>
<tr>
<td>LGN</td>
<td>Lateral Geniculate Nucleus</td>
</tr>
<tr>
<td>LIP</td>
<td>Lateral Intraparietal (Area)</td>
</tr>
<tr>
<td>MT</td>
<td>Medial-Temporal (Area)</td>
</tr>
<tr>
<td>PFC</td>
<td>Prefrontal Cortex</td>
</tr>
<tr>
<td>SC</td>
<td>Superior Colliculus</td>
</tr>
<tr>
<td>SOM</td>
<td>Self-organising map</td>
</tr>
<tr>
<td>WM</td>
<td>Working memory</td>
</tr>
<tr>
<td>Parameter</td>
<td>Equation</td>
</tr>
<tr>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td>$\alpha_D$</td>
<td>4.7</td>
</tr>
<tr>
<td>$Th_l$</td>
<td>4.9</td>
</tr>
<tr>
<td>$\gamma_L$</td>
<td>4.13</td>
</tr>
<tr>
<td>$Th_p$</td>
<td>4.13</td>
</tr>
<tr>
<td>$Th_n$</td>
<td>4.13</td>
</tr>
<tr>
<td>A</td>
<td>4.13</td>
</tr>
<tr>
<td>B</td>
<td>4.13</td>
</tr>
<tr>
<td>$\sigma_L$</td>
<td>4.13</td>
</tr>
<tr>
<td>$\theta$</td>
<td>4.16</td>
</tr>
<tr>
<td>$\gamma_P$</td>
<td>4.16</td>
</tr>
<tr>
<td>$\lambda_P$</td>
<td>4.17</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>4.17</td>
</tr>
<tr>
<td>$\sigma_0$</td>
<td>4.17</td>
</tr>
<tr>
<td>$T_\xi$</td>
<td>NA</td>
</tr>
<tr>
<td>d</td>
<td>4.3</td>
</tr>
<tr>
<td>$T$</td>
<td>4.9</td>
</tr>
<tr>
<td>$Th_p$</td>
<td>4.13</td>
</tr>
<tr>
<td>$Th_n$</td>
<td>4.13</td>
</tr>
<tr>
<td>A</td>
<td>4.13</td>
</tr>
<tr>
<td>B</td>
<td>4.13</td>
</tr>
<tr>
<td>$\sigma_L$</td>
<td>4.13</td>
</tr>
<tr>
<td>$\gamma_L$</td>
<td>4.13</td>
</tr>
</tbody>
</table>

Table 1: Summary of the parameters and their default values for System I
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Equation</th>
<th>System</th>
<th>Meaning</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_o$</td>
<td>5.1</td>
<td>II</td>
<td>Broadness of tuning curve for orientation detector</td>
<td>20</td>
</tr>
<tr>
<td>$\beta_e$</td>
<td>5.2</td>
<td>II</td>
<td>Broadness of tuning curve for line end detector</td>
<td>20</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>5.6</td>
<td>II</td>
<td>Parameter of Gabor filter</td>
<td>0.01</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>5.6</td>
<td>II</td>
<td>Parameter of Gabor filter</td>
<td>6</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>5.6</td>
<td>II</td>
<td>Parameter of Gabor filter</td>
<td>10</td>
</tr>
<tr>
<td>$\tau$</td>
<td>5.9</td>
<td>II</td>
<td>Parameter of sigmoid function in the line termination detector</td>
<td>5</td>
</tr>
<tr>
<td>$\mathcal{N}$</td>
<td>5.6</td>
<td>II</td>
<td>Dimension of the receptive field for low-level feature detectors</td>
<td>7 x 7</td>
</tr>
<tr>
<td>$N^O$</td>
<td>5.6</td>
<td>II</td>
<td>Number of orientation detectors</td>
<td>12</td>
</tr>
<tr>
<td>$N^E$</td>
<td>5.6</td>
<td>II</td>
<td>Number of line end detectors</td>
<td>24</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>5.11</td>
<td>II</td>
<td>Threshold activity for an intersection</td>
<td>70</td>
</tr>
<tr>
<td>$L_{Th}$</td>
<td>NA</td>
<td>II</td>
<td>Threshold luminance for line continuation detector</td>
<td>100</td>
</tr>
<tr>
<td>$A_{Th}$</td>
<td>NA</td>
<td>II</td>
<td>Threshold area for line continuation detector</td>
<td>7 pixels</td>
</tr>
</tbody>
</table>

Table 2: Summary of the parameters and their default values for System II
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Equation</th>
<th>System</th>
<th>Meaning</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m$</td>
<td>NA</td>
<td>III</td>
<td>Minimum distance between two points on $d_j$</td>
<td>4</td>
</tr>
<tr>
<td>$\alpha_p$</td>
<td>NA</td>
<td>III</td>
<td>Amplitude of the increase in activity in working memory</td>
<td>1</td>
</tr>
<tr>
<td>$\alpha_n$</td>
<td>NA</td>
<td>III</td>
<td>Amplitude of the decrease in activity in working memory</td>
<td>-1</td>
</tr>
<tr>
<td>$\sigma_L$</td>
<td>NA</td>
<td>III</td>
<td>Spread of the injected activity in working memory</td>
<td>1.7</td>
</tr>
<tr>
<td>$\Delta_O$</td>
<td>6.3</td>
<td>III</td>
<td>Gain for occluded regions</td>
<td>$5 - 7 \cdot 10^{-4}$</td>
</tr>
<tr>
<td>$\Delta_{\text{min}}$</td>
<td>NA</td>
<td>III</td>
<td>Minimum distance from an unowned border for a point to contribute to $\Delta$</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 3: Summary of the parameters and their default values for System III
Although organising sensory information, such that it can drive potentially advantageous behaviour, appears to be a formidable task, the primate body provides a proof that such a system can exist. The human brain recognises many thousand objects in a single day, with complete success. Sounds are identified and tactile information is used to intelligently direct behaviour towards accomplishing goals.

Coherent internal representations of objects are of fundamental importance in human decision making: since a condition in which both the internal state and external stimulus are exactly repeated is extremely unlikely to occur, it is necessary to abstract the behaviourally relevant aspects of this sensory information, which can be used as part of an internal model to drive behaviour. Internal states such as hunger and tiredness provide motivation for behaviours, such as foraging or finding shelter, and rewards and punishments allowing meaningful responses to be learned: it is necessary to map sensory inputs onto a reward space so that the outcomes of different decisions can be predicted, and the most advantageous action chosen.

At any instant, only a small subset of the available information will actually be relevant to the current task. In the mammalian nervous system a coarse, early analysis of this information identifies the regions of the sensory world which are likely to be important. Attention is used to exclude other sensory information from further processing, sustaining a coherent representation of the selected target. Evidence supports the hypothesis that these targets are abstracted to behaviourally meaningful representations at even this early
The field of machine intelligence is concerned with developing similar a autonomy of learning and behaviour in artificial agents. While machine vision algorithms for object recognition have long been inspired by the primate visual pathway, there has recently been an increasing interest in primate attention, and its possible functional role in driving these behaviours.

This thesis addresses a number of specific questions regarding selection and attention:

• How can sensory inputs be mapped onto a reward space in order to drive useful behaviour, especially in a changing reward scenario? What strategies are the most useful for these tasks?

• How can the visual scene be pre-processed in a useful way to facilitate selection? Especially, how can representations which are robust against occlusion be developed during early visual processing?

• How do early mechanisms for scene understanding facilitate higher behaviour? How can attention be used to create clear object representations as a mid-point in the process of generating useful behaviours?

• What is the purpose of attention? What does it achieve and how are such processes relevant to the task of driving purposeful behaviour in machines?

1.1 Organisation and Summary

A framework of the functional processes involved in selection, recognition and the learning of rewarding behaviours has been developed for this thesis. Two aspects of the framework have been studied in detail: learning useful representations in a changing reward harvesting task and the use of clutter information in creating coherent percepts.

The emphasis of the work in this thesis is on the processes which subserve intelligent behaviour in primates and their potential applications to machine behaviour problems. This work does not have the goal of presenting a developed model of the primate visual
cortex from which temporal predcitions can be made. It follows that this system has been implemented at a high level of neural abstraction.

The next chapter contains a review of the relevant literature regarding visual search, visual attention and object recognition in primates, as well as machine vision and robotic implementations of search and recognition algorithms.

These systems are presented as part of an overall framework for visual attention and behaviour, which is introduced in Chapter 3.

System I (Chapter 4) was concerned with the reward harvesting and representation problems. A simple reward harvesting system has been developed, which learns at two different levels in the processing hierarchy: firstly it learned to organise sensory information into useful representations, which it then learns to map onto a reward space. Treating this as a generic system, the effect of different learning strategies in extracting rewards from the world was then investigated.

Chapter 5, in which System II has been presented, deals with real-world aspects of such an implementation, specifically with the problem of segmenting the visual world in a meaningful way. The system in this chapter was much more closely inspired by the psychology literature than models in the preceding sections.

Chapter 6 contains details of the final system (System III), which was a unification of aspects from the previous models. Top-down bias was introduced and it is demonstrated that the system can be biased to select a particular group of objects based upon their shape.

A general discussion of the models is contained in Chapter 7 and possible future courses of research are proposed in Section 7.4. Final conclusions and a summary of the work are found in Section 7.5.

1.2 Novel Work & Relevance

The work is relevant to the domain of machine intelligence in the respect that it aims to extend previous models of machine attention by including more faithful systems for early perceptual organisation processes. Systems II and III are not presented as a competing approach to object recognition, but rather as a generic pre-processing stage which could potentially improve the performance of any system.
As far as I am aware, the following aspects of this work are novel:

- Our model of reward harvesting at different levels of abstraction is novel in the respect that learning takes place at different hierarchical levels, which can be of use to improve performance at a behavioural task when limited processing resources are available. While self-organising maps are commonly used as classifiers, I am not aware of any system which addresses the reorganisation of internal classifiers in order to improve behavioural performance during a visual search task.

- While very many models of visual selection have been developed, few systems address the effect of attention upon higher-level processes, such as object recognition. This is the only implementation of an object recognition system which explicitly addresses robustness to occlusion during recognition using border ownership information. Many systems do not model the effect of attention except as the enhancement of resolution which is achieved through the generation of saccades.

- Few machine vision systems for visual attention exist which include perceptual grouping mechanisms, and which use complex percepts as the entities which compete for attention.

- This thesis contains the only presentation of a search system in which visual attention can be guided by shape and which is also robust to occlusion.

- This framework is also unusual in the respect that it includes a segmentation process prior to attentional selection, which has more frequently been a late process in such systems.
This chapter contains a review of the research context in which our framework has been developed. Section 2.1 contains an overview of visual search tasks and the functional stages of visual attentive processing. Section 2.2 contains a review of the cortical areas which subserve the processes of visual search. Attentional selection, spatial and object-based attention, and also work on segmentation and border ownership, as well as the neural and behavioural effects of attention, especially regarding object recognition, are discussed in section 2.3. Theoretical models of visual attention are reviewed in section 2.4, followed by a review of computational models for visual attention and object recognition in section 2.5. The final part, section 2.6 contains a discussion of the role of attention in machine vision and image understanding applications.

2.1 Search, Reward Harvesting and Attention

Fig. 2.1 shows an example of a visual search task in which an 'eye' explores a visual scene. Only a portion of the scene is visible at any time, and saccades relocate the centre of gaze, allowing the visual scene to be explored.

In a visual search task, the goal is to drive the eye to attend to a particular class of target while avoiding distractors. This can be extended to a reward harvesting task, in which the agent receives a reward for a saccade to a target location or object, and the goal of the task is to maximise the number of rewards received over time. The system can
be initialised with information about targets and distractors, or this can be learned online
through reinforcement.

Serially testing every position in the environment until a reward is found provides an
obvious, but inefficient, approach to the problem; neither is it the normal strategy which
the primate brain employ during such tasks. Instead, rapidly processed visual information
from across the entire visual field is used to identify likely target locations. These loca-
tions are subsequently investigated through attentional shifts, which increase the clarity
of the target representation in cortex by suppressing activity related to distractors. Object
recognition does not usually take place in parallel: an object must become the target of
attention before it can be recognised. As the search task progresses, attention passes from
target to target, and each attentional shift results in a saccade to the target location.

Figure 2.1: An active visual search task. An 'eye' views a region of the entire visual
scene, which contains cues on an invisible background. Different targets compete for
attention, and gaze is directed towards the selected target.
Fig. 2.2 contains a diagram of the functional components of the attention and object recognition processes in primates (Knudsen, 2007), which is described as involving four main processes: feature extraction, segmentation, selection and object recognition.

Visual saliency, the statistical uniqueness of a region of the visual scene, can exert a very strong control on attentional selection. For example, a vertical bar appears much more distinct (i.e. likely to capture attention) among horizontal bars than is does among identical vertical bars. Top-down modulation of early feature detector activity is also informed by working memory, which can direct selection towards a particular feature.

Selection is generally conceived of as a competitive process, after which the winning percept enters working memory. In this thesis, selection presented as a competition between different percepts, which may be guided by certain stimulus features. Although percept-based information is available at an early stage of visual processing, recognition, in this thesis, is conceived of as a downstream process which is constantly active, but
which will typically yield no useful results until an attentional selection had been made. A saccade to the selected target may follow (overt attention), but this is not essential (covert attention). If a reward is available for making an attentional shift to a particular class of target, then this can bias future selection towards the rewarding features, which are held in working memory. The process by which potential targets are identified prior to the competition is complex, for example a completion process is understood to act upon partially occluded shapes prior to the competition for attention, (Rensink and Enns, 1998).

Attention facilitates the object recognition process in cluttered scenes by sustaining activity related to a single percept in the higher levels of the recognition system. The nature of the attended entity has been the subject of much investigation, though evidence broadly supports the hypothesis that objects form the entities to which attention is directed (O’Craven et al., 1999). Attention is not an absolute prerequisite for object recognition, and it seems possible to recover some information about the visual scene at an early stage of processing. Attention, however, has the effect of reducing the signal-to-noise ratio of the recognition signal (Winkler et al., 2005; Zhang et al., 2011).
2.2 Anatomical Areas Associated with Visual Search

The visual pathway in primates is poorly understood in terms of the specific functions which are subserved by the different anatomical areas, especially those of the frontal and parietal areas. This section describes the processes that are currently understood to take place in the different anatomical regions whose function is related to visual attention and object recognition. The relationship between anatomy and function remains the subject of intense investigation, and the description in this review is by no means definitive.

Unlike the functions that each region subserves, the anatomical areas of the human visual pathway are well defined, Fig. 2.3. The signal passes from the retina through the lateral geniculate nucleus (LGN) of the thalamus to V1 and V2, which subserve early feature extraction and scene segmentation. From V1 and V2, visual processing takes place in two hierarchical pathways: a ventral what pathway and a dorsal where pathway (Ungerleider and Mishkin, 1982; Perrett and Oram, 1993). Processing in these pathways addresses within-object and between-object representations respectively (Humphreys, 1998). A second thalamic pathway exists from the retina to the pulvinar nuclei and thence to V1 (Grieve et al., 2000). The pulvinar pathway is associated with the processing of visual saliency. Knudsen (2007) has identified four process which drive attention: working memory, which is implemented in dorsolateral prefrontal cortex (dLPFC); top-down selection, subserved by the frontal eye fields (FEF); bottom-up saliency, which is a factor in the activity of the lateral intraparietal area (LIP); and spatial selectivity control from areas V2, LIP and V4. Mechanisms of perceptual organisation are hypothesised to be subserved by regions V2, V4 and LIP (Qiu et al., 2007).

2.2.1 Early Processing

Under normal lighting conditions, colour and luminance information is transduced by cone cells of the retina, which are massively concentrated around the fovea (Osterberg, 1935). Initial processing in the form of colour-opposition spatio-temporal filtering takes place within the retina, which has at least 12 separate axonal output channels (Callaway, 2005), and the majority of these axons project to the LGN of the thalamus, with minor projections to the superior colliculus (SC) (saccade generation) and the pretectum (reflex
control of the pupil), (Kandel et al., 2000a). From the LGN, axons project to the primary visual cortex (V1), as well as to superior colliculus, pretectum and hypothalamus.

Figure 2.3: Schematic of locations of the anatomical regions which subserve visual attention and object recognition in the human brain. Shown connections relate to the functional stages of processing in Fig. 2.2. Pale grey labels indicate subcortical regions and cortical areas on the inside surface (i.e. along the parasagittal plane) of the brain and dark grey labels indicate areas along the outer surface. Image modified from Fig. 726, (Gray, 1918)

Classical simple cell in V1 are selective to line orientations and terminations (Hubel and Wiesel, 1962, 1977) and functionally, these neurons are understood to implement low-level feature detectors. Classical complex cells are believed to introduce early phase invariant responses to orientation (Hubel and Wiesel, 1962), and hypercomplex cells show selectivity to line terminations with a particular orientations (Hubel and Wiesel, 1965). The retina is topographically mapped onto V1, i.e. neurons on the retina which lie close together, project to proximal regions of V1 (Grill-Spector and Malach, 2004). A large fraction of V1 cells are tuned to the region of the visual field around the fovea, and the distribution of receptive fields from the fovea to periphery is approximately log-polar (Schwartz, 1977). Receptive fields in V1 are small (1/4° (Hubel, 1988)) and cells typically respond to a single stimulus feature e.g. orientation, colour etc..

Approximately 66% of cells in V1 follow these classical receptive field models (70%
in V2), but there are also large numbers of non-classical receptive fields, such as those tuned to illusory contours (Peterhans and von der Heydt, 1989) or border ownership in V2 (Zhou et al., 2000). Over 50% of the V2 cells were found to show an increased response when the surface which owned a border line was in the preferred direction and these cells are predicted to play an important role in the perceptual grouping and segmentation process, and in the control of spatial selectivity as a result of attentional feedback (Qiu et al., 2007; Craft et al., 1997; Mihalas et al., 2011). While many cues which indicate surface order exist, T-junctions represent the most studied of these categories. Fig. 2.4 illustrates the way in which border ownership information can be inferred from local T-junctions: the T-junctions contain the cues which show B to be the closest surface, i.e. the surface which owns all of its borders. From this, it can be inferred that the common border between A and B is not part of A, but that a region of A is occluded. Mihalas et al. (2011) have developed a neural architecture in which the feedback from hypothetical attention modulated grouping cells in V4 can be used to select only the owned borders of an attended target. The non-classical responses to illusory contours and border ownership suggest that visual processing is more intimately linked with creating behaviourally relevant representations, at some of the earliest stages of visual processing: the visual world is organised into surface-like percepts. A number of cues have been identified that can drive this process, of which T-junctions have, historically, been the most prominent, possibly because of their compatibility with the monocular line drawings that have frequently been used as stimuli. To date, no T-junction selective cell has been found in cortex.

Figure 2.4: Example of a phenomenal T-junction providing a cue by which surface order can be inferred: square A is occluded by square B. The line ik belongs to the closer object while the terminated line hj belongs to the occluded object.
2.2.2 The Ventral Pathway

The ventral pathway includes areas V1-V2-V4-IT and is believed to subserve the process of object recognition. As one progresses through the ventral hierarchy, the receptive field sizes of the neurons increase, as do the complexity of the stimuli to which neurons respond: most cells in the V4 region have large receptive fields which are responsive to both colour and form, and almost all receive some projections from the fovea (Tanaka, 1996).

In the highest region of the ventral pathway, the inferior temporal (IT) area, receptive fields cover a very large portion of the visual field and cells typically respond to very complex stimuli such as a particular face, a hand, fingers together, fingers apart (i.e. not specific to a single hand), etc. (53 % of recorded cells in AIT (Tanaka et al., 1996)). From their systematic study of cell responses in IT, Tanaka et. al (1996) concluded that the cells were selective to shape primitives, but not selective to object categories directly. Cell responses were also found to be robust against changes in angle in the horizontal plane, illumination, position and scale (Tanaka et al., 1996; Ito et al., 1995), but to remain sensitive to vertical orientation, i.e. the same cell would not respond to an object if it was presented upside-down. It has been estimated that the IT cortex contains approximately 600 feature primitives (Tanaka, 1996).

2.2.3 The Dorsal Pathway

For the purpose of this review, the where pathway is principally composed of areas LIP and FEF (areas with motion selective neurons such as V5 are also commonly assigned to this pathway). Information in the dorsal pathway is typically monochromatic, and neuronal responses represent the location of different features as activity on a variety of topographic maps (Colby and Goldberg, 1999). Beyond this, the doral pathway is associated with the guidance of action, for example, a disruption to the function of posterior parietal cortex affects the ability to correctly execute grasping movements (Gallese et al., 1997).

The LIP area, which is organised as a retinotopic mapping of the visual field, is understood to form a saliency map for attentional shifts: the magnitude of the activities of
LIP neurons at a location indicated the behavioural priority of that location as a target for an attentional shift, but activity at a location in LIP does not mean that that location is or will become the target of attention. It is important to emphasise that the activity in LIP is not purely a representation of the distinctness of different regions of the visual scene: a location will also have an increased response when the target is behaviourally relevant (Bisley et al., 2011).

The FEF are associated with the top-down control of attention and they are understood to subserve the transformation of activity from dIPFC into appropriate modulation for V4 (Squire et al., In Review). The very large anatomical projection from the FEF to V4 and then to V2 and V1 is likely to form one of the largest pathways for attentional modulation of visual signals in temporal-occipital cortex. FMRI studies have also found that microstimulation of the FEF with an implanted electrode increased the activation of corresponding retinotopic regions in many other visual areas. Overall, evidence suggests that FEF activity has a strong modulatory influence on visual signals throughout cortex (Moore and Fallah, 2001; Moore and Armstrong, 2003).

Although the activities of LIP and FEF neurons are closely related, the time courses of responses indicate that selection of the attentional target occurs first in LIP (Bisley et al., 2011). Studies which measure the onset latency of attention modulation have attempted to uncover the site of origin of the increased synchrony and the activity modulation associated with attention control (Niebur et al., 2002). Both FEF and LIP share many common properties, and the distinction between their functional roles, and the degree of redundancy in these roles, is poorly understood (Squire et al., In Review).

2.2.4 Working Memory

Once selection has taken place, an attended object will gain access to working memory, from which future behaviour may be directed (Genovesio et al., 2006; Ranganath, 2006; Knudsen, 2007). Working memory is multimodal and each domain involves different regions of PFC (verbal → ventrolateral PFC; visuospatial → dorsolateral PFC etc. (Smith et al., 1996).

Working memory is believed to be distributed through many cortical regions (Fuster,
2.2. ANATOMICAL AREAS ASSOCIATED WITH VISUAL SEARCH

1997), but to be controlled from PFC, which is understood to encode behavioural rules (Assad et al., 2000). These rules are used to bias early selection pathways via the FEF-V4 pathway (Noudoost et al., 2010).

Holding a percept in working memory is sufficient to cause it to guide the selection of future attentional targets (Soto et al., 2008). As the load on working memory increases, it becomes more difficult to bias top-down selection in a useful manner (Lavie, 2005). The contents of working memory have also been demonstrated to affect search times (Downing, 2000) and links between different modes of representation in WM are automatic: verbal cueing of a target has the same effect as visual cueing (Potter, 1975).

Neural correlates of visual working memory have been measured as delay-period activity during stimulus matching tasks, in which stimuli are sequentially presented to a primate, when tasked with releasing a lever when a match to a target stimulus was obtained. Neural recordings during these tasks have found a large fraction of PFC neurons remain persistently active until a match target has been found, suggesting that they are encoding the behavioural goal of identifying a match (Miller and Cohen, 2001).

2.2.5 Gaze Control

Saccade generation and attentional control are two separate processes in the brain, though they share several overlapping brain structures. Since the resolution of the eye is much greater at the fovea (Grill-Spector and Malach, 2004), directing gaze to an object of interest has obvious benefits in improving the resolution of the attended entity and also for the exploration of the visual scene. However, this gain in resolution also comes with the cost of the time required to make the saccade, and so may not represent the most desirable outcome of attentional selection. The activity of the superior colliculus (SC) is closely related to saccade control and target selection (McPeek and Keller, 2004). For example, a projection from the retina to the superficial SC (the inferior SC controls saccades, which is activated via the pulvinar and lateral posterior nuclei of the thalamus) can trigger an automatic saccade to a sudden stimulus onset stimulus (Kandel et al., 2000b), while slower behaviourally motivated saccades can be generated by the projection from LIP. The SC encodes the target location in retinotopic coordinates (Klier et al., 2001).
2.2.6 Inhibition of Return

Another important phenomenon associated with attention is inhibition of return (IOR): regions of space which have recently been attended have been observed to be less likely to re-become the target of attention for a short period of time (∼ 2 s). It has been argued that this process facilitates exploration of the visual scene (Posner and Cohen, 1984).

Theories of inhibition of return specifically preclude the interpretation of the phenomenon as an explicit working memory mechanism for avoiding previously visited locations, but instead describe it as a lower-level mechanism for inhibiting return saccades. Lesion studies indicate that this phenomenon originates in the SC (Klein, 2001). During actual foraging behaviours, return saccades are rare (Gilchrist et al., 2001), but when a previously attended location becomes behaviourally relevant, then the effect of IOR can be cancelled (Farrel et al., 2010).

2.2.7 Reward and the Basal Ganglia

The basal ganglia are a group of structures in the basal forebrain which have been implicated in action selection, working memory and reward-based tasks (Gurney et al., 2001; Hazy et al., 2007; Ponzi, 2008). The BG has a large projection to the FEF (Alexander, 1986), which has been predicted to play a role in action selection, (Chambers, 2006).

Dopamine signals in the BG are understood to subserve the learning of behavioural responses (Barnes et al., 2005). O’Reilly and Frank have presented a model of the basal ganglia as the actor-critic structure which gates access to working memory based on reward prediction (2006).

2.2.8 Summary

At present, there is a basic understanding of the likely functional processes which are subserved by the various regions of the cortex. The view which emerges is one of a visual system with a pre-processing stage, followed by processing in two distinct pathways: a dorsal pathway which addresses the selection of an attention target, and a ventral pathway which addresses the recognition of the selected object. Activity encoding featural information in the ventral pathway can influence the competition for attention, which is
mediated by higher cortical areas, such as dlPFC and the FEF. These areas are believed to encode behavioural rules, and to translate these rules into appropriate modulatory activity.

Object recognition greatly aided by visual attention, which is believed to have the effect of suppressing activity which is not related to the target in the ventral pathway (Zhang et al., 2011). Attentional selection is also closely related to the guidance of eye movements, though the relationship is not simple and the selection of an attentional target is frequently, but not always, accompanied by a saccadic movement. During reward harvesting tasks, the BG drives reward-based learning.

Following from this summary of the perceived stages of attention and the anatomical areas associated with primate attentional selection and object recognition, the next section will review the psychology literature surrounding these processes in greater detail.
2.3 Selection and Attention

The relationship between anatomy and function was discussed in the previous section. This section contains a more detailed discussion of psychological results regarding selection and attention. Three aspects of this are examined in this section: the features which can guide selection, the nature of the selected target and the interaction between selection and object recognition.

2.3.1 Selection

The ability of different visual features to capture attention has been typically investigated through search tasks. These generally take the following form: a number of stimuli are presented on a neutral background and the subject reports whether a stimulus is present or not as quickly as possible (e.g. by pressing one of two buttons: one for a positive response and one for a negative). When the number of cues was compared with the reaction time for an accurate response, the slope was typically linear for many different kinds of cue (Treisman and Gelade, 1980; Duncan and Humphreys, 1989; Wolfe and Horowitz, 2004). For certain tasks, such as reporting if a colour was present among distractors, the magnitude of the slope was almost 0 and increasing the number of distractors only slightly increased the search time (∼6 ms/item). Other features (e.g. facial expression (Wolfe and Horowitz, 2004)) cannot be quickly distinguished from distractors and it appears that the only way to locate such a target is to serially inspect each potential target, meaning that the slope is much larger (∼60 ms/item), i.e. these features cannot guide selection towards a particular target. For conjunctions of features (such as a green and horizontal among vertical or red distractors) or for certain shapes (such as a T among Fs), the search slope is typically intermediate. These observations led to the description of visual search as consisting of serial and parallel processes.

Wolfe & Horowitz (2004) have produced a review of such experiments, in which different attributes have been categorised by the strength of the evidence that they can or cannot guide attention. Colour, motion orientation and size (including length and spatial frequency) were identified as the certain cases for early selection, while stereoscopic depth, closure, line termination and shape were among the other attributes which appear
very likely to guide attention. In contrast, faces, semantic category (e.g. animal) and intersection were considered as being unlikely to play a role in guiding attention. The often-quoted early selection features of colour and orientation may contribute to an impression that the features which guide attention are both simple and localised, but other features which can capture attention are not simple. For example, Enns and Rensink (1990) have demonstrated that 3D appearance is likely to be a guiding feature.

In general, very little progress has been made in understanding how form guides the selection process. It is clear that some forms can capture attention quickly, while most cannot. The difficulty in defining the form space (colour can at least be mapped onto a 3D space) makes it especially challenging to understand the target and distractor attributes which allow this to happen.

Attentional selection is also partly robust against occlusion. Perceptual completion of occluded objects has been shown to take place rapidly (Gerbino and Salmase, 1987), and it is the completed entities which have been found to be accessed by the mechanisms of attentional selection. In an experiment by Rensink and Enns (1998), the reaction time of a subject was tested when searching for a distinct object (notched square) among distractors, Fig. 2.5. The distinct feature of the notched square allowed it to be rapidly detected (Fig. 2.5, 1A). When border ownership became ambiguous, the search time increased greatly (2.5, 1B), indicating that the notched square was no longer distinct because the occluded region had been preattentively completed to form a whole square. Although this result indicates that there is an early representation of shape which is heavily pre-processed, object-background borders may be altered once attention has been directed to the object (Wolfe, 1998).
2.3. SELECTION AND ATTENTION

Figure 2.5: Figure taken from Rensink & Enns (1998). Search for the notched square was rapid in 1A, due to its distinct shape. In 1B, the search became much slower because of the early completion of the occluded region behind the white circle: the distinct notched feature was no longer present, which suggests that the black figure was perceived as a completed square.

2.3.2 Percepts and Selection

At the beginning of the 1980s, the effect of attending was compared to directing a spatial spotlight or zoom-lens upon a region of the visual field (Treisman and Gelade, 1980; Posner, 1980). This spatial spotlight analogy was also supported by experiments in which the location in which a target would appear could be primed by a preceding cue, and that this would speed the response time, since attention would already be engaged at that location. According to the theory, anything which fell within the field of view of this spotlight gained access to the higher object recognition process, which could then be used to drive behaviour. Psychophysical evidence already existed to suggest that this was not exactly the case, for example Neisser and Becklen (1975) had shown that it was possible to attend to one of two superimposed line diagrams (Fig. 2.6) such that changes in the unattended diagram went unnoticed, indicating that attention at least involves something more selective than the attended region of space. To accommodate these object-based effects in cluttered conditions, the spatial spotlight of attention became deformable around surfaces (LaBerge and Brown, 1989). Around this time, Palmer (1983) also proposed a more object-based scheme in which attention placed the object in a canonical, object-based coordinate frame, removing issues related to scale and orientation in the visual plane, but not to orientation in depth.
Duncan (1984) also experimented with divided attention between objects: subjects were found to be poorer at reporting features of two separate objects than they were at reporting features within a single entity, even when both objects occupied the same region of space.

In a study on symmetry, Baylis and Driver (1995) have shown that it was easier (i.e. it required less time) to shift attention between features in a single object than from an in-object feature to a feature exterior to the object. Rensink and Enns (1998) also demonstrated a similar effect: it took less time to report on the length of an occluded line than to report on the distance between a line end and the occluder and that attention could be driven by an occluded feature. The fast response times for object-based reports indicate that object-like percepts are formed at a very early stage of processing.

The need for a more object-based theory of attention was further supported by the observation of same-object advantages during comparison tasks: a subject could compare two features on a single object faster than comparing the features on two separate objects, even if the object appeared to be split into two by an occluder (Egly et al., 1994; Behrmann et al., 1998). This has been interpreted to indicate that the object itself is the entity which is attended and that this process is robust against occlusion. The extra time required to compare features on two separate objects was the time required to disengage and re-engage attention, which was not necessary when the features appeared on a single object: features within the single object were compared in parallel.

Not only are occluded objects completed, but completion also takes place around blind spots on the retina, caused by scars or blood cells and the occlusion of cone cells by blood
vessels and scars in the retina (Ramachandran and Gregory, 1991).

These examples suggest that the entity which is selected by attention is not obvious. One insight is found in the medical condition of simultanagnosia, whereby a patient is unable to make comparisons between two objects. Holmes and Horax (1919, in (Scholl, 2001)) observed that patients could not compare the lengths of two lines, but that they could make the comparison if the same lines were presented as the sides of a trapezoid. All of these examples appear to indicate that objects are the units of attentional selection.

Research on this topic is heavily influenced by the work of the Gestalt psychologists in the 1920s, who produced the first qualitative descriptions of the principles upon which the smallest entities which are perceived are formed into perceptual groups, (Wertheimer, 1923). Proximity, similarity, common fate, closure, equilibrium, symmetry, (continuation of) direction, good-Gestalt, figure-ground relation and habit (e.g. upon seeing the characters 314cm they are segmented as 314 _ cm and not 31 _ 4cm) were identified as the important factors in the formation of percepts. These principles are hierarchically organised and the results were considered universal and apparent to the degree that no experiments were judged necessary. (Fig. 2.7) shows an example of hierarchical groupings: in Fig. 2.7a, the black dots are grouped into pairs through proximity, while in Fig. 2.7b, different pairs of dots becomes the selected entities. The process of quantifying these descriptions has been troublesome and models are hindered in dealing with the computational complexity of the global operations which the word gestalt implies, though some interesting progress has been made (Feldman, 1999). Grossberg (Grossberg, 1994) has also extensively discussed the role of perceptual organisation processes in primate vision.
2.3. SELECTION AND ATTENTION

(a) Figure 2.7: Illustration of perceptual grouping; (a) the factor of proximity causes the dots to be grouped into pairs where possible; (b) demonstration of a hierarchical grouping effect: the factor of similarity overrules the factor of proximity. Images from Wertheimer (1923)

If attention was purely feature-based (as opposed to object-based), then only task-based attribute of the selected target should be enhanced. This has been tested in an experiment by O’Craven et al. (1999), in which the subject was instructed to attend to either a face or a house, which were presented as transparent overlayed images. Recognition related activity could be separated, with faces being recognised in the fusiform face area (FFA) and houses in the parahippocampal place area (PPA). If one of the objects moved, and if the subject was instructed to attend to the motion, then enhanced activity related to the processing of the features of the moving object was observed in the relevant area. In a purely feature-based account of attention, this would not happen. From this result the authors argue that while it was possible to attend to individual features (Maunsell and Treue, 2006), objects form the primary entities of attention.

2.3.3 Neural Signatures of Attention and Object Recognition

Early studies have shown that gist information is rapidly accessible (Thorpe et al., 1996; Rousselet et al., 2002; Li, 2002), indicating that although attention alters the neural signal in higher cortical processing areas, the pre-attentive signal can also provide some identity information. At a neural level, the object identity is represented as a population code in IT cortex (Hung et al., 2005). Some feature binding is observed prior to the competition for attention (Winkler et al., 2005), suggesting that attention does not act as a gate to the object recognition process, as has been modelled by Olshausen et al. (Olshausen
et al., 1993). Attention appears to restore the IT population activity to a similar condition as if the object was presented in isolation (Zhang et al., 2011).

This description of IT is highly compatible with the observation that border ownership selective activity is modulated by attention (Qiu et al., 2007; Bartels, 2009): that the mechanisms of early segmentation are also used to restore higher representations of objects through attention. It is also consistent with the observation that within-object features are represented differently to inter-object features in cortex (Hayworth et al., 2011). Phenomenally, attention has the effect of improving contrast sensitivity, which is commonly used in psychophysical experiments (e.g. Moore et al. (2001)).

Humphries (1998) has argued that once the target object has been selected, the recognition of features within an object surface is a parallel process. The relative position of the local features is also likely to play a strong role in this process, either as a structural description (e.g. Biederman (1987)) or as a feature hierarchy (e.g. Riesenhuber & Poggio (1999)).

2.4 Theoretical Models of Visual Attention

The most prominent models of attention are Treisman’s feature integration theory (FIT) (Treisman and Gelade, 1980; Treisman, 1988, 1998, 2006), the biased competition hypothesis of Duncan et al. (Duncan and Humphreys, 1989; Desimone and Duncan, 1995) and Wolfe’s guided search hypothesis (GS) (Wolfe, 2007).

2.4.1 Feature Integration Theory

Treisman and Gelade (1980) developed a very successful early model of visual attention that has been expanded on several occasions (e.g. Treisman (2006)). The model follows the division of search into parallel and serial stages. Textural discontinuities were used to segment the visual scene and in an early parallel stage features were extracted and stored in object files. The process of selection could be driven by a single dimension: if a single dimension (e.g. red or vertical) was sufficient to distinguish the target from its distractors then the target could be selected immediately, but if the identity of the target
was dependent upon a feature conjunction (e.g. red and vertical), then a serial search would have to be used, during which potential targets (or all colours and orientations) were sequentially excluded from the search process until the target was identified. Such a serial attentional shift was required to read the contents of an object file (colour and orientation), while without attention, a single feature dimension could be compared across all files (colour or orientation). Neural activities in areas beyond the window of attention were suppressed, allowing objects within the target region to be clearly processed and conjoined features to become bound together and identified and labelled as objects (Kahneman et al., 1992).

This theory explained why single features, such as colour and orientation, could quickly drive search, but intersections such as are found in T and L, or conjunction of shape and colour could not. It is in this way that attention provides the necessary ‘glue’ of binding between different features (Treisman and Gelade, 1980). More recently, Treisman has argued that in the initial pass of activity through the object recognition hierarchy, neuronal responses are not conjunction-selective, and therefore IT neurons should become active for all possible feature conjunctions across the entire visual scene, even though some of these hypothetical objects are not actually present (Treisman, 2006).

While this theory has been very important in the history of understanding visual attention, it has largely made way for two successors: biased competition and guided search, largely because of the rigidity of the distinction which the theory makes between serial and parallel processes and its apparent incompatibility with more recent psychophysical results.

### 2.4.2 Biased Competition

The Biased Competition Hypothesis states that different representations within the brain compete with each other for processing resources. A series of psychophysical measurements (1989) motivated the development of a new theory by Duncan and Humphries. They observed that search efficiency was not so much a function of the feature itself as it was of the similarity between targets and distractors and between different distractors in themselves.
Competition between systems was integrated such that the same winner emerged from all hierarchical levels and it was possible to bias the competition with top down signals (Duncan, 1997). Because attention was manifested as a distributed competition in this framework, it was proposed that there should be no ‘attentional module’ within the brain (Humphreys and Riddoch, 1994), which increasingly appears to be the case, although it also appears that the competition for selection does not take place in all areas, but within the dorsal pathway at a level in which the representation is behaviourally meaningful (see section 2.2.3).

### 2.4.3 Guided Search

Guided search was initially an extension of FIT (Wolfe et al., 1989) which allowed conjunction searches to be guided by information from a single target feature (at this point FIT predicted that search under conjunction conditions was totally random (Treisman and Gelade, 1980)).

When searching for the red vertical bar among mixed greens and horizontals, the attribute of red or vertical can be used to inform the serial search process. When there is no guiding attribute (T among L) the target is found by random search (the graph of reaction time against the number of targets has a large slope). When the guidance is clear (e.g. red among green, horizontal among vertical), the search appears parallel; for intermediate conditions (red horizontal among green or vertical), the slope of the reaction time against the number of targets has an intermediate value. The model has been extended a number of times, and is now in its fourth generation (Wolfe, 2007).

Structurally, the model comprises of selective and non-selective pathways (Wolfe et al., 2011b). The control systems which drive attention have direct access to certain kinds of information (such as colour) via the non-selective pathway. The selective pathway contains a bottleneck in that attention must be deployed before the information can be used in the object recognition process. Attention is guided by both top-down information and by information from the non-selective pathway. An important feature of this model, which Wolfe has recently re-emphasised (Wolfe et al., 2011a) is that the output of the early visual processes is not necessarily identical to the guidance signal.
tasks can be entirely accomplished with the non-selective pathway, such as gist extraction (Rousselet et al., 2002). A second bottleneck is related to "attentional blink", in which it is difficult to report a change within 200 - 500 ms of an attention shift (Raymond et al., 1992). Even though these items which are presented during the "blink" may not be reported, there is evidence that they are still processed in some detail and that this bottleneck is a late process, e.g. unnoticed words can still lead to anticipatory responses (Luck et al., 1996). Only one item can be selected at a time, though an item does not have to be fully processed before the next can be selected in a process that is analogous to a pipeline.

The process of guidance is complex: guidance is not based on early visual processing directly, but it is more categorically driven (Wolfe, 2007). Selection is only driven by information for a single feature dimension at any time. Object recognition is modelled as an accumulation process, in which evidence for the identity of an object slowly builds, until a threshold is exceeded. Although an object must be selected before recognition can begin, recognition does not have to be completed before another object can be selected: once the information has propigated to the next neural layer, the system is free to initiate the next selection.

One of the main appeals of this model is that it has been regularly updated to account for new psychophysical phenomena, such as the attentional blink, and the domain of the model has been broadened to include overt eye movements and target-absent trials. Another interesting way in which the model has been expanded is to explore search for rare items (Wolfe et al., 2005).

2.5 Computational Models of Attention and Object Recognition

The models in Section 2.4 provide a conceptual framework through which attention can be described. This section contains a review of the computational models of attention and recognition that are relevant to this thesis. The level of biological realism varies considerably between these models though they typically take the form of a feature extraction process, followed by a competition for selection that can be biased through top-down
control. This review divides models into the following categories:

- **Selection:** these address early visual processing, the choice of a target and biasing selection towards that target in future.

- **Segmentation:** these models are either concerned with the division of the visual scene into regions which can be usefully treated as single entities by higher processes.

- **Recognition:** these models tend to deal less with attentional effects and typically attempt to recognise objects in an image in a purely feed-forward manner.

### 2.5.1 Selection

The most widely implemented model of attentional selection to date has been the visual saliency map (Koch and Ullman, 1985; Itti et al., 1998), in which a probability distribution for an attentional shift to each retinotopic location is calculated from the uniqueness of elementary visual features such as colour and orientation. These maps are used to identify regions of the visual scene which are intrinsically attention-grabbing, though they are frequently criticised for failing to reproduce gaze patterns accurately (Tatler et al., 2011). Usually activity from a number of these maps is pooled together to form a conspicuity map which then drives selection.

A winner-take-all structure is then used to select the most salient point in the visual scene. Since targets are selected in a purely bottom-up manner, the selected regions of the visual scene may be of little behavioural consequence (Najemnik and Geisler, 2005).

Olshausen et al. (1993) presented an early large-scale model of attentional selection to address the problem of learning scale-invariant representations of object. A saliency map was used to identify relevant regions of the visual scene and a set of control neurons were used to gate access of information for processing in higher cortex. By dynamically altering the size of the window of attention (though gating), the intended target could be correctly scaled and routed to the working memory. The memory could be used to bias selection towards target features on future trials. The need for an anatomically distinct gate which routes information to IT was a major theoretical problem with the model,
though it had the advantage of allowing the object recognition process to be scale invariant. The authors made a number of predictions about changes in receptive field positions and sizes during attentional modulation and predict that the pulvinar region was the site of the gating mechanism within the brain. The system was also unable to deal with clutter and cues had to be presented in isolation on a neutral background. Amit & Mascaro (2003) extended this model further to include a more satisfying recognition process based on several aspects of the HMAX model of Poggio et al. (1999). Both models provide schemes for object-based selection.

While the above models use covert attention with a static retina, the models of Lanyon & Denham (2004; 2005) have used an active visual search scheme, in which the retina is too small to encompass the entire world and eye movements must be generated in order to explore the entire scene.

Lanyon and Denham’s (2004; 2005) models of active visual search have combined both object-based and spatial attention to reproduce temporally accurate psychophysical scan paths. Conceptually, this system was an implementation of a biased competition scheme. A gate was not required for attention (but scale invariance is also unaddressed), and the model was novel in two respects: firstly because it dealt with the different time courses of spatial and object-based attention and secondly because of the hierarchy of features which could drive visual search, i.e. colour taking priority over orientations (Motter and Belky, 1998). Point-objects were represented in both ventral and dorsal pathways and competition between different stimuli on a retinotopic saliency map resulted in the selection of a winning location. Cues were red or green, horizontal or vertical bars and inhibition of return prevented gaze from being directed to recently attended locations. Top-down bias could be both spatial or object-based, and was introduced though feedback from IT and LIP layers directly.

The interaction between object features and attentional selection has been addressed in a model by van der Velde and de Kamps (2001). This work suggested a specific microcircuit which could provide feedback between different cortical regions in order to suppress activity which was unrelated to the selected target. While the modelled IT could not encode spatial information and the LIP area could not encode featural information,
2.5. **COMPUTATIONAL MODELS OF ATTENTION AND OBJECT RECOGNITION**

A V4 area which weakly encoded both properties was used to transfer information between the areas, suppressing activity related to distractors. The network used correlations between the feedback and feed-forward signals in the network: target related feedback correlated with target-related feed-forward signals, which in turn did not correlate with feedback distractor signals. When the signals were correlated, there was sufficient activity to disinhibit the inhibitory neurons which projected to competing regions, allowing a single retinotopic population to win. The system reproduced the time courses of recordings of IT neuron activities during attentional selection, (Chelazzi et al., 1993).

Several models address selection in the context of machine vision, while remaining broadly inspired by the psychology literature (Sun and Fisher, 2003; Sun et al., 2008; Amit and Mascaro, 2003; Frintrop et al., 2005; Aziz and Mertsching, 2008).

Sun and Fisher’s model (2003) is a purely bottom-up biased competition system which was based around the idea of hierarchical groupings of features. In the early versions of this model, groupings were formed from similar features across the image. The saliency of each grouping was equal to the sum of the individual saliences at each position within the group and competition took place between groupings at a number of spatial scales. The grouping mechanism allowed regions of the image which are likely to belong to a single object to compete as a single entity. One major problem with this model was that any region which was not assigned to a grouping enters the competition as a single (pixel) entity. As a consequence, the manually selected groupings inevitably become the targets of attention (Sun and Fisher (2003), Fig. 14). The model was shown to produce more realistic scan paths than the model of Itti and Koch (1998), which was attributed to its ability to group salient features into competitive units. While groupings were manually selected in the first version, this model was later extended to include a more realistic segmentation mechanism using the EDISON algorithm (Christoudias et al., 2002) and overt eye movements (Sun et al., 2008). No quantitative comparisons were made with actual eye movements. The system was particularly interesting in examining the effect of early grouping mechanisms upon selection and attention.
2.5. COMPUTATIONAL MODELS OF ATTENTION AND OBJECT RECOGNITION

2.5.2 Segmentation

Sajda & Finkel, (1995), Craft et al. (1997) and Mihalas et al. (2011) have addressed the interaction between segmentation of the visual scene and attention. Mihalas et al. (2011) extend the model of Craft et al. (1997) to include attentional selection in the same feedback circuit which achieved segmentation. In these models, segmentation relied upon border ownership which was informed by the orientation of T-junctions in the figure. Robustness to occlusion was developed through the different interactions of owned and unowned borders across different perceptual surfaces. A related model by Domijan and Šetić (2008) uses a recurrent network to provide figure-ground assignment. Another model by Sajda and Finkel (1995) explicitly used border ownership to segment the visual scene.

Early object descriptions were also addressed by Walther and Koch (2006), though these "proto-objects" were based on contiguous regions of uniform activity on the visual saliency map, as opposed to using the more principled mechanisms of Craft et al. (Craft et al., 1997). This method was similar to the segmentation mechanism in Sun & Fisher (2003) in the sense that it involved a more object-based selection without trying to address the occlusion and border ownership.

The model of Fazl et al. (2009) used the mechanisms of surface-based attention to address the problem of category learning from different viewpoints. Because the attended entity is a surface or *shroud*, it was possible for the system to differentiate between saccades within a surface and outside a surface, which allowed a library of different views of a single object to be established.

2.5.3 Object Recognition

The most successful object recognition algorithms to date have purely feed-forward architectures, in which the object is recognised without any prior selection stage. Many of these systems are based on local feature matching without any deeper structural description of the visual scene (the "bag-of-features" approach). The potential applications of a machine attention system are discussed in section 2.6.

Two families of systems have been particularly successful at object recognition: bio-
2.5. COMPUTATIONAL MODELS OF ATTENTION AND OBJECT RECOGNITION

logically inspired hierarchical models such as the HMAX system (Riesenhuber and Poggio, 1999; Serre et al., 2007), and Lowe’s more abstract scale invariant feature transform (SIFT) system (2004).

The genealogy of HMAX can be traced back at least as far as Fukushima’s neocognitron (1980), which provided the broad form of many subsequent machine vision object recognition systems. The basic structure of successive neural models for object recognition has remained similar since: a hierarchical structure of feature detectors and spatial sub-sampling units build a position invariant representation of the image features (Murase and Nayar, 1995; LeCun et al., 1998; Riesenhuber and Poggio, 1999; Swain and Ballard, 1991; Serre et al., 2007). Initial feature detectors are tuned to low-level prototypical features, such as edges, colours and orientations followed by the detection of more complex features such as conjunctions of colour and form in successive layers. By having receptive fields on many scales, the relative positions of local features are implicitly encoded. Large banks of these feature detectors are assembled and the response can provide enough certainty about the identity of the target object to allow it to be recognised, even in cluttered conditions.

Many different training methods have been used for tuning the feature detectors at high levels: for example, Serre et al. (2007) use a brute force approach by using several thousand randomly chosen features to produce a robust response to an object despite the presence of occlusion and clutter in the scene. It is because of the computational resources required to calculate so many feature detector matches that a lot of effort has been invested in developing efficient feature matching algorithms, for example, MSER provides a method of accelerating the search process through tree searching (Obdržálek and Matas, 2005). In a similar manner, hashing (Gionis et al., 1999) has been suggested as a useful approach to this problem and has recently been incorporated into the classification stage of the HMAX model (Lee et al., 2011).

Lowe’s SIFT algorithm (2004) is much less biologically plausible, and operates by identifying scale invariant features in an image that are also invariant to distortion and rotation. Such features are found at local luminance maxima and minima in scale space of the image. These features also have the property of being highly distinct, and the
2.6. VISUAL ATTENTION FOR BEHAVING MACHINES

An interesting exploration of the interaction between attention and object recognition was made by Walther, Rutishauser et al. (Walther et al., 2005) in which the authors showed that by using a window of attention to select relevant parts of the image, object recognition could be improved with the SIFT model (Lowe, 2004). In general, "bag of features" approaches are used in machine vision, but some algorithms also take relative feature positions into account (Leibe and Schiele, 2003). Forssén et al. (2008) have also shown that such a "bag of features" approach is highly dependent upon the viewing angle (Forssén et al., 2008).

2.6 Visual Attention for Behaving Machines

Given the constraints of physical volume and power budget with which the brain has evolved, the visual attention system has adopted the strategy of processing a small, selectable region of the visual scene in great detail while coarsely processing a surrounding portion and neglecting the remainder. Machine vision systems are subject to analogous constraints of computing volume and power consumption (especially for battery-operated systems) and both biological and robotic systems must produce appropriate behaviours in a timely fashion: otherwise the opportunity to behave advantageously will be lost.

These similarities suggest that a machine vision system may benefit from processes which are analogous to attention in primates, in which a region of interest is selected from the visual field for more detailed processing.

The name 'machine attention' suggests that such a system should have the goal of identifying behaviourally relevant subsets of relevant sensory information from the set of all available sensory information, even when the goal is simply to reduce the size of the window of processing (e.g. Walther et al. (2005)). Early surface extraction and perceptual organisation of the visual scene is also appealing from an image understanding perspective: while recognising objects at certain locations is useful, a structural description of the image which will facilitate the planning of future behaviour is also highly desirable (Biederman, 1987).
A contrasting approach might involve processing the entire visual field in parallel without any selection or segmentation, which might be achieved with an extension of HMAX. Serre et al. (2007) provide a feed-forward method of achieving recognition, this leaves further work regarding segmentation, since the borders of the object will be important in a reaching task, as well as an understanding of potential obstacles, such as occluders might provide.

There are also very many differences between biological and machine-attention systems. The primate visual system can easily identify closure, yet cannot rapidly detect the intersection of lines of two particular slopes (Wolfe and Horowitz, 2004): the opposite is true of a conventional computer, whose architecture does not lend itself to calculating closure efficiently. The log-polar arrangement of cones on the retina may be advantageous for the primate, but it is more computationally demanding to implement such an arrangement using a uniform array of photosensors. It would be surprising if such a fundamental change to the front-end of this system does not also have the behavioural consequence of generating different gaze patterns: the need to reorient the sensor in order to improve resolution will be removed, though it will still be necessary to move the eye in order to explore the entire scene.

Although a visual attention system in machines may search for different early attributes than visual attention system of mammals, this does not mean that an early selection process may not be also useful for machine vision. This raises the question of what level of biological detail is beneficial when addressing machine vision in general, and the role of the computational architecture in defining the features which can efficiently (i.e. ought to) drive selection.

### 2.7 Self-Organising Maps

The self-organising map (SOM) features prominently in one of the systems in this thesis. The SOM algorithm develops discrete, low-dimensional representations of a vector space of inputs, while preserving the topological relationship between the inputs. The global ordering of the space is achieved by purely local processes. The development of the topological map in the SOM is unsupervised and purely a function of pre-defined rules
2.7. SELF-ORGANISING MAPS

and the input vectors.

Initially developed by Willshaw and von der Malsburg (1976), the general form of the self-organised map was proposed by Kohonen in 1982 and has been very influential in a wide range of classification and dimension-reducing tasks.

The principle on which the SOM is trained is simple: a space of n-dimensional inputs is to be mapped onto an m-dimensional space. A m-dimensional mesh of nodes is defined, typically square or hexagonal. At each node there exists a neuron, with n-dimensional receptive fields. The Euclidean distance between an input vector and every node is calculated, with the closest node being defined as the best matching unit (BMU) at coordinate \( k \). A Hebbian learning rule is implemented in which the node vector values at the BMU and its neighbours (typically weighted with a Gaussian function) are then adjusted to become more similar to the input vector:

\[
\Delta W_i = G(i, k, \alpha, \sigma) d_i
\]  

(2.1)

Where \( G(i, k, \alpha, \sigma) \) is a Gaussian distribution centred at coordinate \( k \) with amplitude \( \alpha \) and spread \( \sigma \), \( d_i \) is the distance between the BMU and every other node in the mesh in the stimulus space. The values of \( \alpha \) and \( \sigma \) are slowly decreased over the course of training, initially causing the map to represent features coarsely, then allowing it to represent features in increasingly fine detail.

The main uses of SOMs are as data classifiers, in which a input vector is expressed as a position on the SOM (i.e. the position of the BMU). For this purpose, SOMs are typically trained offline prior to classification. Numerous variants of the SOM exits, such as the iSOM, to which points can be subsequently added without retraining, the data-structure preserving ViSOM and the temporal sequence learning recursive SOM (Yin and Allinson, 1999; Yin, 2002; Vogetlin, 2002). While these different varieties are tailored towards specific applications, the classical SOM was used in this thesis, though with an online training regime.
2.8 Summary

This review has presented the research context in which our framework for visual attention has been developed.

Visual processing in primates has been discussed, and the key functional stages of visual search have been described: early feature extraction, perceptual organisation, selection and attention. Each of these processes has been associated with an anatomical region of the primate brain and the interaction between the functional stages has been discussed.

Prior to attentional selection, the visual scene is organised into shape-based percepts, which form the channels which compete for attention. Certain visual features are capable of driving attentional selection, such as shape and colour, though feature conjunctions are not. Although shape information is available prior to attention, it is difficult to use it in achieving object recognition. Following attentional selection, activity related to unselected percepts was suppressed, allowing the selected object to be recognised.

The relevant psychology and physiology literature regarding segmentation and selection suggests that a great deal of processing is performed on the visual scene prior to selection. The nature of the attended entity and the effect of attention during object recognition tasks has been discussed.

Theoretical and computational models of visual attention have been reviewed and discussed in terms of their similarity to the psychological literature. Visual attention has also been discussed in the context of machine vision problems: the utility and form of an analogous system is perhaps not immediately obvious, due to a number of fundamental differences in the computing architecture.
3.1 Introduction: Framework for Guidance, Attention and Learning

This section contains an overview of our framework for visual attention, and a brief description of three system implementations. Details and analyses of these systems appear in Chapters 4 - 5.

3.2 Summary of Implemented Systems

Fig. 3.0 contains a more detailed representation of the same functional stages of attention and selection than were shown in the overview in Fig. 2.2. This system was influenced by the four stages of attention and object recognition, as presented by Knudsen (2007). Coloured boundaries indicate the aspects of the framework which were implemented in each of the three systems. Each of these systems addressed a different aspect of our framework:

System I (a & b): visual attention in the context of reward harvesting tasks.

System II: the interaction between scene segmentation and object recognition. In particular, the use of occlusion information in object recognition was investigated.
System III: shape-based attentional selection, again in the context of developing robust responses against environmental clutter.
3.2. SUMMARY OF IMPLEMENTED SYSTEMS

Figure 3.0: Schematic of the proposed framework for visual attention; (a) System Ia; (b) System Ib; (c) System II; (d) System III
System I was sub-divided into two parts. System Ia considered a reward harvesting task with online learning of both the target feature and the feature detectors, which were learned on a self-organising map (SOM). A parameter search was used to uncover the most successful strategies for reward harvesting and the effectiveness of online and offline training of the SOM was compared.

System Ib presented a robotic implementation of the reward harvesting task using seven-segment digit cues, SOMs were trained offline prior to the experiment. Successful reward harvesting was demonstrated. This system highlighted a number of problems regarding perceptual organisation, which were addressed in the subsequent systems. In Systems II and III, the self-organising map was not used to learn representations of feature space.

System II addressed attentional selection and object recognition under occluded conditions. Different representations of owned and unowned borders allowed the system to distinguish between partially occluded objects and 'mosaic' objects (i.e. those which owned all of their visible borders). This information was used to improve object recognition performance.

System III presented a synthesis of Systems I and II with a reward harvesting system, in which selection could be guided on the basis of shape. Shape representations were robust against occlusion prior to selection.

Experimental results from each system are discussed singularly within the relevant chapters, followed by a general discussion of the significance of the systems and their relationship with both other work and with biology in Chapter 7.

### 3.3 Structure

The structure of the proposed framework is described in this section, Fig. 3.0. Systems can be categorised as being active vision systems, in which attentional selection resulted in a saccade, and passive systems, in which all attentional shifts were covert. System Ia involved active vision in a virtual world and System Ib involved a robotic implementation of active visual search. Systems II & III used a static retinal image. In all systems, the scene consisted of cues on a neutral background, see Table 3.1.
The input to the system consisted of a 'retina', which viewed a stimulus image. In the first processing stage, basic features (colour, line orientation, line termination) were extracted from the retinal image, after which the visual signal was processed in two functionally separate pathways: a recognition-based \textit{what} channel and a selection-based \textit{where} channel.

Local features were extracted in the \textit{what} pathway. Selection gated access to higher recognition processes and could bias selection towards a particular feature. In System I, local feature detectors were learned on a self-organising map. Learning could take place offline prior to the experiment, or online, in which case the attended cues were used for training the SOM. In system Ia, the plasticity of the SOM weights was linked to the success rate over a window of time. Systems II & III used static sets of feature detectors which were trained prior to the experiment.

The perceptual organisation stage took place within the \textit{where} pathway. The visual scene was segmented into early surface representations, and border ownership was determined.

The third stage (selection) stage was dedicated to the formation of a retinotopic priority map for saccades. Saliencies across individual surfaces were pooled into competing channels, such that the surfaces described the image regions which competed with each other for attention (Systems II and III specifically addressed segmentation and the development of robust representations in clutter). The contents of each saliency channel competed to select the most behaviourally relevant entity, which then became the target of either an overt (saccadic) or a covert (non-saccadic) attentional shift.

Once the attentional target was selected, activity unrelated to that target was suppressed in the gating \textit{what} pathway, which allowed the target object to be recognised as though it has been presented in isolation. Object recognition took place in the highest

<table>
<thead>
<tr>
<th>System</th>
<th>Retina</th>
<th>World</th>
<th>Stimulus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ia</td>
<td>Active</td>
<td>Virtual</td>
<td>Point</td>
</tr>
<tr>
<td>Ib</td>
<td>Active</td>
<td>Real</td>
<td>Point</td>
</tr>
<tr>
<td>II</td>
<td>Passive</td>
<td>Virtual</td>
<td>Object</td>
</tr>
<tr>
<td>III</td>
<td>Passive/Active</td>
<td>Virtual</td>
<td>Object</td>
</tr>
</tbody>
</table>

Table 3.1: Summary of the framework implementations in this thesis
layers of the *what* pathway in the fourth stage (attention and learning).

The attended target gained access to working memory, and if a reward accompanied the attentional shift, then the attended features were stored in working memory and were used to bias the selection of subsequent targets. An inhibition of return (IOR) mechanism suppressed repeat saccades to the same target within a short period of time.

### 3.4 Implementation Details

The systems were implemented APRON software (http://apronsoft.co.uk), a simulation tool for massively parallel processor arrays (Barr and Dudek, 2009) and data analysis was performed in MATLAB (http://www.mathworks.com/).

The systems were implemented at a high level of neural abstraction, as a hierarchical structure of abstract neural layers, in which the magnitude of a ‘neuron’ approximately corresponded to the mean activity of a population of biological neurons. The implementations in this thesis do not attempt to model neural dynamics, and the level of abstraction is too great to predict the time-course of the attentive processes. Instead, the emphasis of this work is to investigate visual attention and reward harvesting in the context of machine vision and image understanding applications.
4.1 Introduction

This system was used to investigate the interaction between the learning of internal representations of stimuli and the learning of a policy for the guidance of target selection through reinforcement.

A reward was given for an attentional shift to a particular category of cue and the goal of the search task was to successfully bias selection in order to maximise the number of rewards which were obtained.

A representation of encountered sensory information was learned on a SOM (Kohonen, 1982). The learning of SOM weights took place online and reinforcement learning was used to create associations between positions on the SOM and expected rewards in a working memory. Reward expectation was encoded as activity in the working memory, which had the same topology as the SOM, i.e. activity at a position on the working memory indicated the reward expectation for the feature represented at the same position on the SOM. The topology-preserving nature of the SOM was essential for this reason: diffusely injected activity in the working memory map allowed associations to be made between rewards and similar features. The classical SOM was used in this system, though training took place online, using the attentional target. Because the system learned to look at rewarding cues more frequently than unrewarding cues, the SOM was found to over-represent rewarding cues, allowing fine distinctions to be made between these.
The sizes of both the working memory and the SOM were limited, such that it was necessary for the system to develop an efficient representation of the input space. In order to allow the system to adjust behaviour when the rewarding targets were changed, it was necessary for the system to contain a degree of plasticity. This plasticity was implemented in two ways: firstly, the activity of the working memory could be rapidly altered through reinforcement learning and secondly, the plasticity of the SOM could be increased in order to allow the SOM to relearn a more behaviourally relevant representation of the input space.

We tested the ability of the system to successfully harvest rewards for different biasing strategies in two different implementations: a virtual system which searched for colour cues (System Ia) and a robotic system which searched for seven segment digits (System Ib). This work has been partially published in (Brohan et al., 2010a,b).

4.2 System Description

4.2.1 Overview

An overview of the system is presented in Fig. 4.1 and a list of symbols is shown in Fig. 4.1. The world image $S^W$ was composed of a large number of cues on an invisible background.
4.2. SYSTEM DESCRIPTION

Figure 4.1: Network diagram for System I, see text for details. Maps beginning with R were retinotopic, with S were world maps, and with F were in featural coordinates, with the exception of $F^{SOM}$, which was of dimensions 768 x 768 pixels.

At any point in time, the camera viewed a sub-section of the visual world ($S^W$) as the retinal image ($R^W$). The retinal image was split into a luminance channel $R^L$, which was processed in the retinotopic where pathway, and into a retinotopic array of feature vectors $C$, from which activity at each retinotopic location was pooled into a feature-selective what pathway.

4.2.2 Low-level Feature Extraction

System Ia used stimuli which consisted of colour patches, and rewards were made available for selecting cues of a particular colour, which was changed throughout the course of an experimental trial. The retinotopic set of vectors $C$ contained three RGB colour...
4.2. **SYSTEM DESCRIPTION**

<table>
<thead>
<tr>
<th>Layer</th>
<th>Size</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>RW</td>
<td>128 x 128</td>
<td>Retinal image</td>
</tr>
<tr>
<td>SW</td>
<td>256 x 256</td>
<td>World image</td>
</tr>
<tr>
<td>C</td>
<td>128 x 128 x 3</td>
<td>RGB colour detectors</td>
</tr>
<tr>
<td>F_{SOM}</td>
<td>128 x 128 (6 x 6)</td>
<td>Retinotopic array of SOMs</td>
</tr>
<tr>
<td>F_{R}</td>
<td>6 x 6</td>
<td>Pooled SOM activity</td>
</tr>
<tr>
<td>R_{L}</td>
<td>128 x 128</td>
<td>Luminance detectors</td>
</tr>
<tr>
<td>R_{Th}</td>
<td>128 x 128</td>
<td>Threshold luminance</td>
</tr>
<tr>
<td>R_{D}</td>
<td>128 x 128</td>
<td>Threshold feature distance</td>
</tr>
<tr>
<td>R_{S}</td>
<td>128 x 128</td>
<td>Visual saliency map</td>
</tr>
<tr>
<td>R_{F}</td>
<td>128 x 128</td>
<td>Feature-based saliency</td>
</tr>
<tr>
<td>F_{S}</td>
<td>6 x 6</td>
<td>Feature saliency map</td>
</tr>
<tr>
<td>F_{M}</td>
<td>6 x 6</td>
<td>Working memory</td>
</tr>
<tr>
<td>S_{IOR}</td>
<td>256 x 256</td>
<td>World IOR map</td>
</tr>
<tr>
<td>R_{IOR}</td>
<td>128 x 128</td>
<td>Retinotopic IOR map</td>
</tr>
</tbody>
</table>

Table 4.1: Key to symbols in Fig. 4.1. The dimensions of each neural layer are shown.

channels \{R^R, R^G, R^B\}, which were processed in the *what* pathway. These vectors were represented as locations on a retinotopic map of SOMs (F_{SOM}). A schematic of the experimental setup is shown in Fig. 2.1.

**System Ib** was broadly similar to System Ia, but with the exclusion of online learning of the SOM weights and with the inclusion of a robotic front-end, for which a camera which was mounted on a pan-tilt unit to form a robotic eye that explored a visual world of seven-segment digit cues on a high-contrast background. An image of the experimental set-up is shown in Fig. 4.2. Due to difficulties in using colour cues under uncontrolled lighting, seven-segment digit cues were used instead, which were fixed on a uniform blue background. In humans, digit identity cannot drive attention (Wolfe and Horowitz, 2004), but it is treated as a pre-attentive feature in this model.
The cues were of identical size, defined by distance \( d \) in Fig. 4.3. At any point in time, the camera viewed a sub-section of the visual world (\( S^W \)) as the retinal image \( R^W \). Orientation information from the image was initially extracted using 5 x 5 edge detection kernels of orientation \( 0^\circ \) and \( 90^\circ \) to form two orientation feature maps: \( R^0 \) and \( R^{90} \). The kernels were tuned to detect maximum contrast for a single edge:

\[
R^0 = R^W \otimes k^0 \\
k^0 = \begin{pmatrix} -0.1 & -0.1 & -0.1 & -0.1 & -0.1 \\ 0 & 0 & 0 & 0 & 0 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0 & 0 & 0 & 0 & 0 \\ -0.1 & -0.1 & -0.1 & -0.1 & -0.1 \end{pmatrix}
\] (4.1)

\[
R^{90} = R^W \otimes k^{90} \\
k^{90} = \begin{pmatrix} -0.1 & 0 & 0.2 & 0 & -0.1 \\ -0.1 & 0 & 0.2 & 0 & -0.1 \\ -0.1 & 0 & 0.2 & 0 & -0.1 \\ -0.1 & 0 & 0.2 & 0 & -0.1 \\ -0.1 & 0 & 0.2 & 0 & -0.1 \end{pmatrix}
\] (4.2)

In the next layer, seven dimensional feature vectors \( C_i \) were generated for each retinotopic location by receiving projections from \( R^0 \) and \( R^{90} \). Features were extracted at 7 relative positions in the visual field, defined by the distance \( d \). These features were used...
4.2. SYSTEM DESCRIPTION

to recognise the digits, Fig. 4.3:

\[ C_i = C_{xy} = \{ R_{x-d, y-d}^{0}, R_{x,y-2d}^{0}, R_{x+d, y-d}^{0}, R_{x,y}^{0}, R_{x-d, y+d}^{0}, R_{x,y+2d}^{0}, R_{x+d, y+d}^{0} \} \]  \hspace{1cm} (4.3)

4.2.3 Selection

In both systems, the feature vector \( (C_i) \) for each location was classified using a self-organised map \( (F^{SOM}) \) with weights \( (W) \). For each retinal location \( i \), the best matching unit (BMU) was calculated as the location \( m \) on the SOM with the shortest distance.
between its preferred vector \((W_m)\) and the feature vector \((C)\) at the location \(i\). The resulting activity of \(FSOM\) was pooled for all retinotopic locations into a feature map \((FR)\) such that the activity of a neuron at location \(m\) in \(FR\) was equal to 1 if a feature was present and 0 if a feature was not present in the visual scene, regardless of the number different retinotopic locations that had activated the point \(m\) in the SOM:

\[
FR_m = \begin{cases} 
1 & \exists i : m = \arg\min_j \{\|C_i - W_j\|\} \\
0 & \text{otherwise}
\end{cases} \quad (4.4)
\]

The activity of the feature saliency map \((FS)\) represented the expectation that a feature would provide a reward if attended, which was calculated as the sum of the activities of the working memory \((FM)\) and the \(FR\) layer. In this way, each potential target had a baseline salience from \(FR\), which could be increased or decreased based on the activity of \(FM\):

\[
FS = FM + FR \quad (4.5)
\]

The \(FS\) activity was thresholded between \(-1\) and 1 and was mapped back to retinotopic space as \(RF\), such that the magnitude of the activity at location \(i\) in \(RF\) was equal to the magnitude of the activity in \(FS\) at position \(m\), which corresponded to the location of the best matching unit on the SOM for the corresponding feature vector \((C_i)\). This remapping was achieved through \(FSOM\), which allowed featural information to be associated with a retinotopic position. \(FSOM\) neurons encoded both position and featural information.

\[
RF_i = FS_m : m = \arg\min_j \{\|C_i - W_j\|\} \quad (4.6)
\]

The activity in the feature distance map \((RD)\) was a function of the distance between each retinotopic feature vector \((C_i)\) and its BMU in \(W\):

\[
RD_i = \exp \left( \min_j \frac{\|C_i - W_j\|}{\alpha_D} \right) \quad (4.7)
\]

where \(\alpha_D\) was a scaling constant.

A cue segmentation map \((RTb)\) was calculated from the thresholded luminance activ-
ity also modulated the saliency map, which had the effect of assigning the background a saliency of \(-\infty\):

\[
R_{i}^{Th} = \begin{cases} 
0 & R_{i}^{L} \geq Th_{L} \\
-\infty & \text{otherwise}
\end{cases}
\]

(4.8)

where \(Th_{L}\) was the luminance threshold.

The activity in the visual saliency map (\(R^{S}\)) represented the behavioural priority of an attentional shift to a given location in the retinal image, but with recently attended locations inhibited. The magnitude of the saliency depended upon an object being present in the visual scene (\(R^{Th}\)), having a large activity at the corresponding location in working memory (\(R^{F}\)), being well represented on the SOM (\(R^{D}\)) and upon not having been recently selected (\(R^{IOR}\)):

\[
R^{S} = R^{D} \cdot \left(1 + \frac{1}{2}R^{F}\right) - R^{IOR} + R^{Th}
\]

(4.9)

A noisy winner-take-all was implemented by multiplying the activity of each unit on the retinotopic saliency map (\(R^{S}\)) by white noise (\(N\)) and the maximum activity across the map was selected as the target location \(z\) for the next saccade:

\[
z = \arg\max_{i} \{R_{i}^{S}N_{i}\}
\]

(4.10)

4.2.4 Attention and Learning

Gaze was shifted such that the target cue was located in the centre of the visual field. The inhibition of return map contained traces of previously visited locations. After a saccade was made, activity was injected into the location surrounding the target in \(S^{IOR}\). Activity in this map decayed over time, which prevented gaze from returning to a previously selected target for a short period of time.

The saccadic target was classified by finding the BMU \(k\), for the feature vector \(C_{z}\),

\[
k = \arg\min_{j} \{||C_{z} - W_{j}||\}
\]

(4.11)

If a cue was rewarding (\(r = 1\)) (this was determined by an external control system),
Some activity was introduced to the location surrounding \( k \) in the learning layer \( F^M \) while if a cue was unrewarding, \( (r = 0) \) the activity at the point \( k \) was reduced. At each timestep \( (t) \), the working memory activity was calculated as follows:

\[
F^M_t = \begin{cases} 
\gamma_L F^M_{t-1} + F^L_t & Th_n \leq \gamma_L F^M_{t-1} + F^L_t \leq Th_p \\
Th_n & Th_n > \gamma_L F^M_{t-1} + F^L_t \\
Th_p & Th_p < \gamma_L F^M_{t-1} + F^L_t 
\end{cases} \tag{4.12}
\]

\[
F^L_j = \begin{cases} 
G(j, k, A, \sigma_L) & , r = 1 \\
-G(j, k, B, \sigma_L) & , r = 0, j = k \\
0 & , r = 0, j \neq k 
\end{cases} \tag{4.13}
\]

where the activity of \( F^M \) was thresholded at \( Th_n \) and \( Th_p \), \( G \) was a Gaussian function in feature map space centred on neuron \( k \) with amplitude \( A \) and full width at half maximum \( \sigma_L \).

During training, the SOM weights were modified such that the distance between each weight \( W_j \) and the input vector of the winning unit \( C_z \) was reduced. The degree to which a weight was modified \( \Delta W_j \) decreased as a Gaussian function of the topographic distance from the BMU in the map space. The strength of the change was modulated by the learning rate \( \alpha_t \) and the degree to which changes influences neighbouring unit weights was controlled by the width of the Gaussian function \( (\sigma_t) \):

\[
\Delta W_j = \| C_z - W_j \| G(j, k, \alpha_t, \sigma_t) \tag{4.14}
\]

In the classical SOM, global organisation is generated across the map by constantly decreasing values of \( \alpha_t \) and \( \sigma_t \) over time. This slowly 'freezes' the weights across the map, allowing increasingly more fine details to be represented by the SOM.

For System Ia, if the network was unable to predict rewards with sufficient success, it was assumed that the internal representation of that stimulus was not sufficiently resolved on the SOM and that it may be useful to 'unfreeze' the SOM weights and to attempt to generate a new representation of the stimulus. A small number of SOM nodes were
deliberately used in order to force the inputs to compete for representation on the map.

The time course of the variables associated with the modulation of the SOM during a typical trial is shown in Fig. 4.4.

The values of $\alpha_t$ and $\sigma_t$ were modulated by the reward history in the following way. The reward prediction error ($\delta(r)$) provided a measure of the surprise when an expected reward was undelivered or if a reward was received from an unexpected source. It was defined as the absolute value of the difference between the received reward $r$ and the expected reward, represented by the activity of the winning neuron in the working memory layer $F^M_k$:

$$\delta(r) = |r - F^M_k|$$  \hspace{1cm} (4.15)

Performance over time was measured with the leaky integral of a non-linear function of the reward prediction error:

$$\xi_t = (1 - \gamma_P) \xi_{t-1} + \gamma_P \left( \frac{\delta(r)}{1 - \delta(r) + \theta} \right)$$  \hspace{1cm} (4.16)

where $\theta$ was a small value to prevent the function from being undefined when $\delta(r) = 1$, and $\gamma_{PE}$ was the leak rate.

If the value of $\xi$ exceeded a predefined threshold ($T_\xi$), which occurred at a time $t_{th}$, the values of $\xi$, $\alpha_t$ and $\sigma_t$ were reset to their initial values of zero, $\alpha_0$ and $\sigma_0$. The values of $\alpha_t$ and $\sigma_t$ decreased linearly with decay rate $\lambda_P$ after the reset time $t_{th}$, and $\epsilon$ was a small positive number ($1E-5$) which prevented $\sigma_t$ from decreasing to zero, which would result in the value of $G$ being undefined:

$$\alpha_t = \max\{ -\lambda_P(t - t_{th}) + \alpha_0, 0 \} \quad , \quad \sigma_t = \max\{ -\lambda_P(t - t_{th}) + \sigma_0, \epsilon \}$$  \hspace{1cm} (4.17)

The time course of the variables associated with the modulation of the SOM during a typical experiment is shown Fig. 4.4.
4.3 Results: System Ia

The system was tested with a reward harvesting task, in which the goal was to maximise the number of times a particular colour was selected. The target colour was static in Experiments I and II, and changed over time in Experiment III. A reward was received for a saccade to the target colour, and the system was initially unaware of the target, and of changes in the target.

Figure 4.4: Use of reward prediction error to modulate $\alpha_t$ and $\sigma_t$ over time. (a) $\delta$ was the reward prediction error; (b) $y$ was a non-linear function of the error, which filtered out small errors; (c) $\xi$ was the leaky integral of $y$ over time, the dashed line marks the reset threshold, $T_\xi = 40$; (d) $\sigma/\sigma_0$ was the width of the neighbourhood function used during SOM training.
4.3. RESULTS: SYSTEM IA

4.3.1 Implementation Details

The world image ($S^W$) was of dimensions $256 \times 256$ pixels and contained $517 \times 1 \times 1$ pixel colour cues. The retina ($R^W$) viewed a $128 \times 128$ pixel region of the world, in which the mean number of cues on the retina at any time was $122$, with a standard deviation of $31$. Colours were drawn from the HSV distribution $HSV = \{u, v, 0\}$, $u = 0..2\pi$, $v = 0..1$ (Fig. 4.5), and were randomly assigned to the cue positions at the beginning of each trial.

![Figure 4.5: Colours used as stimuli in System Ia, and their locations in RGB space](image)

Experiment I

The goal of the first experiment was to investigate the effect of changing the number of discrete colours ($N$) and the width of the reward window on the rate of reward harvesting, in order to obtain the performance range of the system. The reward window was defined as the fraction of the $N$ discrete cues for which rewards were available.

Two cases were examined: when the SOM weights were static (i.e. were learned offline prior to the experiment), and when the weights of the SOM nodes could be modulated online. Trials lasted for 1000 steps and the parameter values used were $\{A = 0.25, B = 0.25, \gamma_L = 0.95, \sigma_L = 0.5\}$, Eqns. 4.12, 4.13. The width of the reward window was varied as a fraction of $N$ (i.e. $\{\frac{N}{32}, \frac{2N}{32}, \frac{3N}{32}, \frac{4N}{32}, \frac{5N}{32}\}$).

Fig. 4.6 contains two plots of the reward harvesting rate as a function of the width of the reward window and the number of discrete colours used in a trial, for online (a) and
4.3. RESULTS: SYSTEM IA

offline (b) learning of SOM weights.

The number of harvested rewards was consistently greater for the case with online learning of SOM weights, indicating that this allowed the system to develop a more behaviourally useful representation of the sensory world.

For both cases, when the reward window was narrow, the system was much more successful at extracting rewards from a low resolution colour space; this effect decreased as the width of the reward window increased. This indicated the point at which the system became unable to distinguish between neighbouring rewarding and unrewarding stimuli. Accordingly, the system failed to consistently associate a position on the SOM with a reward expectation value.

Standard deviations in the accumulated rewards were also smaller when the learning of SOM weights took place online: again indicating that the rewarding locations were more clearly represented with online learning. With offline learning, the standard deviations of the accumulated rewards were especially large for small reward window widths. Individual data points are shown with hollow markers for $N = 32$ in Fig. 4.6b. When the standard deviation was large, the data was bimodal: the task was only learned with certain SOMs in the offline case.

When $N > 32$, the reward harvesting performance increased greatly once the width of the rewarding region exceeded $\frac{2}{32}$ of the total input space.

From this experiment, it can be concluded that the width of the rewarding window played a critical role in the harvesting rate, and that performance could be improved by using the modulated SOM, presumably because it allowed the 'borderline' region of space, in which neighbouring colours are rewarding and unrewarding, to occupy a greater area on the SOM. Similarly, the number of discrete colours $N$ also affected the harvesting rate, though this remained large when the size of the rewarding region of space was greater than $\frac{2}{32}$ of the total input space.
Figure 4.6: Effect of the number of discrete stimuli $N$ and the size of the reward window on the fraction of rewards harvested; (a) online learning of SOM weights; (b) no learning of SOM weights

### 4.3.2 Experiment II

The goal of this experiment was to find the system parameters which resulted in the greatest reward harvesting rate. The learning rate for the working memory ($A = B$), the decay rate for working memory ($\gamma_L$) and the broadness of the injected activity in the
working memory ($\sigma_L$) were varied. To reduce the number of free system parameters, SOM weights ($W$) were learned offline prior to the experiment and did not change ($\alpha_t = 0$). For this experiment, the number of discrete colours was also fixed $N = 256$.

Trials lasted 1000 steps and there were 10 trials in each experiment. For the first 500 steps, a reward was given for selecting yellow cues in the range (RGB = {1, 0.55- 0.74, 0}), and for the second 500 steps, a reward was given for selecting green cues in the range (RGB = {0, 1, 0.06- 0.42}). The fraction of rewarding cues was 0.031 and 0.059 respectively. Appendix A.1.2 contains results for different frequencies of change in the rewarding target.
Fig. 4.7 shows the effect of the decay of the working memory activity $\gamma_L$ upon the reward harvesting rate for different values of $A(= B)$, $\gamma_L$ and $\sigma_L$. The harvesting rate was greatest for values of $\gamma_L = 0.95$, and was frequently poor when $\gamma_L = 1$: rewarding SOM positions which also represented distractors could be strongly inhibited, then never visited again. The was especially true when $\sigma$ has a small value: in this case there was very little communication between neighbouring positions on the working memory, and so there
4.3. RESULTS: SYSTEM IA

was no mechanism to disinhibit a location which had provided a single punishment.

Reward harvesting was more successful when there was some decay of working mem-
ory because it allowed locations which were unrewarding for a particular cue to be re-
assessed under different cues at a later point in time, which also lead to a greater variance
in the accumulated reward in this case: a good location could be judged as unrewarding
because a single cue which excited that location did not result in a reward. If there was
no decay of the working memory and a very large learning rate, then this location could
be permanently inhibited in working memory. Situations in which there was no visible
reward rarely occurred and did not affect these results.

When activity was more diffusely generated in $F^M$, the reward harvesting rate in-
creased and reward harvesting was especially poor for the least diffuse case, $\sigma_L = 0.25$.
This shows that a significant component of the success of this system lies in the topolog-
ical ordering of inputs on the SOM.

This experiment places a boundary on the performance of the system: the kind of task
which it can solve, and the strategies which are best employed when mapping a reward
onto a map of a continuous sensory space.

4.3.3 Experiment III

The third experiment tested the ability of the working memory to guide reward har-
vesting, again comparing static and online learning of SOM weights. Trials lasted for
10000 steps, with 10 trials per experiment and 256 cues. Four different reward epochs
were used, each lasting for 2500 time steps. For the first 2500 steps, a reward was pre-

dented for foveating to orange cues in the region RGB = {1, 0.74-0.81, 0}. For the second
epoch, a reward was given for foveating to green cues in the region RGB = {0, 1, 0.08-
0.25}. For the third epoch, a reward was given for foveations to blue cues (RGB = {0,
0.15-0.23, 1}). Finally, in the fourth epoch, a reward was given for foveating to magenta
cues in the region RGB = {1, 0, 0.47-0.63}. The reward window for epochs 2 and 4 was
twice as wide as for epochs 1 and 3, in terms of the number of rewarding cues which
could be selected.

The following three cases were compared:
4.3. RESULTS: SYSTEM IA

- Online learning of both the SOM and working memory
- Offline learning of SOM, online learning of working memory
- Offline learning of SOM, no learning of working memory

SOMs for each of these cases were trained for 200000 steps with constant linear decay throughout, and $\alpha_0$ and $\sigma_0$ values of 0.02 and 0.05 respectively.

Fig. 4.8 shows the number of saccades made to each cue colour in this experiment. Saccades were preferentially made to rewarding cues when $\alpha_w \neq 0$, though peaks were taller and sharper when there was online learning of SOM weights, indicating that the bias towards the rewarding feature was more precise. When there was no memory of rewarding features ($\alpha_W = 0$), saccades were made to all features without any bias.
Figure 4.8: Number of saccades to each colour over 10000 time steps; (a) working memory & modulated SOM; (b) working memory & static SOM; (c) no working memory & static SOM

Fig. 4.9 shows the fraction of saccades made to each colour in time bins of 100 steps.
The bias for saccades in the region of each of the rewarding cues was confined to its particular epoch (Fig. 4.9 (a, b)). In the case without working memory (Fig. 4.9c), the system showed no preference for any cue colour.

Figure 4.9: Saccades to each colour in bins of 100 time steps; colour indicates the fraction of saccades made to each colour index. In the control experiment, the rate of selection was similar for all cues.

The increased contrast of the peaks in Fig. 4.9 indicates that the system was able to make a much finer distinction between rewarding and unrewarding cues in the online learning mode, which resulted in a large increase in the number of rewards over the course of the experiment. This was possible because the modulated feature map preferentially represented regions of the input space that were in the vicinity to the rewarding feature. Fig. 4.10 shows examples of the modulated SOMs at the end of each of the four epochs: in each epoch, a large fraction of the area was dedicated to the rewarding cue. When the reward window was larger (Epochs 2 and 4), the degree to which the rewarding area was over-represented was greater.

This effect is analogous to plasticity observed in topological maps in the primate brain: areas which fell silent were observed to be ‘invaded’ by neighbouring regions (Merzenich
et al., 1984), while areas which were frequently activated developed more tightly tuned receptive fields, which allowed finer distinctions to be made between similar stimuli (Jenkins et al., 1990).

Figure 4.10: Examples of modulated SOMs preferentially representing rewarding regions of input space; (a) yellow, Epoch 1; (b) green, Epoch 2; (c) blue, Epoch 3; (d) magenta, Epoch 4

Figure 4.11: The total number of rewards received over time. The error bars mark one standard deviation across 10 trials. For clarity, every 250th point has been plotted

Fig. 4.11 shows the cumulative number of rewards over time for all three cases. The number of rewards was much greater when working memory was used. The use of the modulated SOM also resulted in a large increase in the mean reward harvesting rate, while the standard deviations were also reduced by the online learning of the SOM. The slope of the graph shows the rate at which rewards were harvested, which was greater when the
reward window was more broad.

The mean number of times a saccade was made to a cue as a function of cue distance from the reward in RGB space was measured (Fig. 4.12). Distances between cue and reward were assigned to bins of size 0.05 and the number of cues per bin was counted (the RGB cube was of length 1 on all axes). The number of saccades per bin was normalised by the number of different cues that were assigned to each distance bin. In the modulated SOM, saccades were more tightly tuned towards rewarding units, indicating that the online system was better able to distinguish between rewards and distractors.

Figure 4.12: Number of selections as a function of the Euclidean distance between the colour of the selected cue and a reward in RGB space

4.4 Results: System Ib

Experimental results from System Ib have been presented in this section. In a single experiment, rewards were given for making saccades to certain digits, and the ability of the working memory to learn rewarding locations was tested.

4.4.1 Implementation Details

The robot consisted of a V-UAS14 camera (Logitech), a pan-tilt unit PTU-46-17.5 (Directed Perception, California) and a controller PT-D46r5C14S (Directed Perception,
California). The world image contained three instances of each digit (Fig. 4.2). The
camera captured a 128 x 128 region of the visual scene at any time. Retinotopic saccade
coordinates were linearly mapped to the servo controller for the pan-tilt unit and the next
saccade was actualized and the next image was recorded from the camera. The camera
viewing angle was 28°, and cues centres were located between pan angles ±15° and at tilt
angles ±12°.

The SOM had 16 nodes in a 4 x 4 square topography with initially random weights.
Prior to the experiment, the weights \( W \) of the SOM were trained in the usual manner for
1000 trials with linear decay of the learning rate and the influence. During the training pe-
period, the camera was fixed at the cental viewing position and feature vectors were chosen
from pre-identified points of the map, which lay at the centres of the digits. The trained
SOM is shown in Fig. 4.13.

![Image of the digit representations at each of the 16 SOM nodes after train-
ing. The grayscale intensity of the bars indicates the magnitude of the weight for each
dimension of the feature detector. The subscript shows the node index.]

**4.4.2 Experiment IV**

This experiment tested the ability of the system to successfully harvest rewards by
associating locations on the feature map with rewards. The experiment consisted of three
ePOCHs, each containing 200 saccades.
In the first epoch, the robot explored the visual world without receiving any rewards, in the second epoch, rewards were given for making saccades to the digit 2, and in the third epoch rewards were given for making saccades to the digit 6. In both rewarding epochs, the system learned to preferentially direct saccades to the rewarding digit, (Fig. 4.14).

![Figure 4.14: The cumulative number of saccades made to each digit over time](image)

In the first epoch, there was a small systematic difference in the number of saccades to the different cues, which was attributed to the position of the cues on the board, though this was not investigated further. In the second epoch, the number of saccades made to the rewarding digit (2) increased greatly due to the reinforcement learning of the rewarding features. There was also an increase in the number of saccades made to the digit 3, which activated nearby neurons on the memory map $F^M$ (Fig. 4.13) and became biased by the diffuse stimulation of $F^M$ activity in equation 4.13. Saccades to each of the other digits decreased in frequency. In the third epoch, saccades were preferentially made to the rewarding digit, 6. Again, due to the diffuse activity in $F^M$, there was a slight increase in the number of saccades made to digit 5 in this epoch.
4.4. RESULTS: SYSTEM IB

Fig. 4.15 shows the $F^M$ activity for each neuron as a function of time. In first epoch, no feature developed a positive reward expectation and reward expectation was generally negative. In the second epoch, there was sustained activity at positions 0 and 1, which are associated with the rewarding digit 2. In the third epoch, the activity in positions 0 and 1 quickly died away and activity was instead maintained at positions $I2$ and $I3$, which were associated with the new rewarding digit: 6.

Fig. 4.16 shows the fraction of saccades to features which activated each position on the SOM: in the first epoch there were slightly more saccades to positions around the edge of the SOM (Fig. 4.16a). In the second epoch (Fig. 4.16b), the vast majority of saccades were made to features which activate positions 0 and 1 in feature space and in the third epoch, the majority of saccades are made to positions $I2$ and $I3$ (Fig. 4.16c). This demonstrates that the system learned to saccade to features which it expected would provide it with a reward, and that the system has the plasticity to adapt its expectations as the reward scenario changes.
4.4. RESULTS: SYSTEM IB

Figure 4.16: The fraction of saccades which activated each region on the SOM; (a) Epoch 1: no working memory; (b) reward for saccades to the digit 2; (c) reward for saccades to digit 6

It is notable that far more rewards were harvested during Epoch 2 than in Epoch 3 (Fig. 4.14). Fig. 4.17 shows the distance to each of the locations on the SOM for each of the test stimuli \{2, 3, 4, 5, 6\}. The representation of the digit 2 on the SOM appears much more distinct from other digits than the representation of 6. This lead to greater confusion between digits 5 and 6, which can be seen in Fig. 4.15.

Figure 4.17: Normalised Euclidean distances in RGB space to each point on the SOM for ideal stimuli. This provides an indication of the quality of the representation of a digit on the SOM. (a) Digit 2; (b) Digit 3; (c) Digit 4; (d) Digit 5; (e) Digit 6
4.5 Discussion

Experiments I, II and III have demonstrated that System Ia was capable of reward harvesting in a changing environment though two concurrent learning mechanisms: a slow unsupervised learning of feature detectors and a rapidly changing working memory which learned through reinforcement. In Experiment III, it was shown that the interaction between these two learning systems resulted in a greater number of harvested rewards than when the learning of representations took place offline. Modulating the learning rate of the feature detectors with a reward prediction error resulted in a greater number of feature detectors being dedicated to resolving between similar rewarding and unrewarding features. This allowed rewards to be harvested more precisely.

A parameter sweep of the system was used to investigate the robustness of the system to changes in the number of stimuli, the width of the reward window, the number of changes in the rewarding target and the internal parameters of the system. The use of diffuse injections of activity in the working memory was shown to improve the performance in reward harvesting.

In Experiment IV, associations between feature representations and rewards in working memory were again used to successfully bias saccades towards rewarding objects. Although the objects are not represented by individual neurons, the system was still able to learn to interact with the SOM representation in order to successfully bias selection towards the target.

Because the working memory acted on the same topological map as the feature detectors, diffuse activity in the working memory provided a mechanism for exploring regions of feature space close to those which have previously been found to be rewarding. This encouraged the system to explore the rewarding region of the feature space completely. This method of exploring similar stimuli for rewards relied on the assumption that the map of feature space was continuous, as in Experiments I, II and III. In experiment IV, this was not the case: unlike the continuous, 1-dimensional data which was used to train the SOM in Experiment I, II and III, the feature data in Experiment IV was 7-dimensional and not continuous, but clustered around the representations of the 5 digits in 7D space. As a result, the neurons in locations between the best digit representations on the SOM
4.5. DISCUSSION

encoded locations on the paths between the digits, which were not possible attention targets (for example, position 8 in Fig. 4.13). This is also clear from Fig. 4.16a, in which 1/2 of all locations on the SOM appear to almost never become active in the unbiased condition. Experiment IV indicated that the SOM method, in its present form, does not work well for high-dimensional data.

In Experiments I, II and III, the data was also located at a single point in retinotopic space and no segmentation between object and background is necessary, while in Experiment III, segmentation was avoided by searching for a single high-dimensional feature which could identify the entire object, essentially making object identity an early feature.

The key contribution of System I was that it demonstrated a method of learning novel features from the sensory world online, and reorganising limited resources towards a behaviourally useful representation of external world. In terms of more general tasks, the system is currently unable to address temporal actions, though in principle, the SOM could be used to search for positions in multimodal sensory spaces: for example, if the reward was contingent upon searching for a point with a particular velocity in 2D space, then a 4D SOM could be trained upon position and velocity information.

In the primate visual system, selection cannot be guided by object identity (Wolfe and Horowitz, 2004), though there is evidence that the scene is segmented into perceptual surface prior to selection (Rensink and Enns, 1998). It is also likely that surface shape is an attribute which can guide attention (Wolfe and Horowitz, 2004).

A key problem with this system was that it contains in-built mechanisms for treating the stimuli it encounters as single points, thereby avoiding the issue of percept formation. It is more usual to encounter objects which have an irregular spatial component, and the recognition process suffers from problems of crowding and occlusion.

System II was developed to address the problem of early segmentation by creating representations of percepts in the visual scene prior to the competition for attention. These representations were robust against occlusion and fragmentation and were used to inform the selection and recognition processes.
CHAPTER
FIVE

SYSTEM II: OBJECT OCCLUSION IN SELECTION AND RECOGNITION

5.1 Introduction

The system presented in this chapter examines the perceptual grouping and segmentation process, and its effect upon target selection and object recognition in machine vision applications. The system was heavily inspired by the early segmentation and completion processes in primates.

A key property of this system was that it used different representations of the owned and unowned borders of perceptual surfaces. Similar representations have previously been employed by Grossberg (1994), in which the author also describes the use of both perceptual surfaces and owned and unowned border representations to explain a variety of optical effects. While this was achieved with binocular cues in Grossberg’s system, the system presented in this thesis uses the local cue of a phenomenal T-junction to perform image segmentation. Chapter 7 contains a discussion of the different possible cues for surface segmentation and border ownership assignment. The system presented in this chapter focuses in detail upon the effect of percept formation on object recognition, for which quantitative results are presented.

By allowing the system to distinguish between potentially occluded objects regions and objects which were fully visible, a target could be distinguished from environment-
tal clutter. This border ownership information was used to improve object recognition performance.

The system implemented object-based visual attention in the sense that the object was the entity which was selected, while activity related to nearby or overlapping distractors was suppressed prior to recognition, i.e. the effect of attention was to segment the object from the background, reducing the SNR for the recognition system.

5.2 System Description

The system acted in four stages, Fig. 5.1.

Figure 5.1: Overview of System II. Unimplemented connections were shown with dashed lines

In the first stage, the luminance of a stimulus image provided the input to the network
in the retinal layer ($L$). Basic features of line orientations ($O$) and line terminations ($E$) were extracted in 12 and 24 retinotopic orientation selective channels respectively:

\[
O_{\theta i} = \exp \left[ -\beta_O \sum_{j \in \mathcal{N}} W^O_{\theta j} \left( G^O_{\theta j} - R_{ij} \right)^2 \right] \tag{5.1}
\]

\[
E_{\theta i} = \exp \left[ -\beta_E \sum_{j \in \mathcal{N}} W^E_{\theta j} \left( G^E_{\theta j} - R_{ij} \right)^2 \right] \tag{5.2}
\]

Where $\mathcal{N}$ was the span of the receptive field, a 7 x 7 square in this case. $R_i$ was the region of $L$ centred at retinotopic location $i$, $\beta_O = \beta_E = 20$. $G^O_{\theta}$ and $G^E_{\theta}$ were the orientation and line termination features while weights ($W^O_{\theta}$ and $W^E_{\theta}$) modulated the contribution from each retinotopic position $\mathcal{N}$. The features and weights are shown in Table 5.1.

$G^O_{\theta}$ was calculated as a Gabor function and $G^E_{\theta}$ as a modified Gabor function, with a sinusoid modulating the major axis (Table 5.1):

\[
x_{\theta} = x \cos \theta + y \sin \theta \tag{5.3}
\]

\[
y_{\theta} = -x \sin \theta + y \cos \theta \tag{5.4}
\]

\[
G^O_{\theta} = \exp \left[ \frac{- \left( x_{\theta}^2 + \gamma^2 \cdot y_{\theta}^2 \right)}{2\sigma^2} \cos \left( \frac{2\pi x_{\theta}}{\lambda} \right) \right] \tag{5.5}
\]

\[
G^E_{\theta} = \exp \left[ \frac{- \left( x_{\theta}^2 + \gamma^2 \cdot y_{\theta}^2 \right)}{2\sigma^2} \cos \left( \frac{2\pi x_{\theta}}{\lambda} \right) \sin \left( \frac{2\pi y_{\theta}}{\sigma} \right) \right] \tag{5.6}
\]

Where $\theta = \left\{ 0, \frac{\pi}{N}, \frac{2\pi}{N}, ..., \frac{(N-1)\pi}{N} \right\}$, $x = \{-3, -2, \ldots, 2, 3\}$, $y = \{-3, -2, \ldots, 2, 3\}$, $\gamma = 0.01$; $\lambda = 6$; $\sigma = 10$, $N^O = 12$, $N^E = 24$.

$W^O_{\theta}$ and $W^E_{\theta}$, the weights for each position $j = \{x, y\}$ in the grid, $\mathcal{N}$ were calculated as follows. Pixels which were useful in uniquely identifying an orientation had a greater weight than pixels which were less informative (e.g. the pixels at the very centre of the receptive field contained very little information about the orientation of a bar, and so they were assigned a small weight value). Pixels to the sides of orientation detectors also had a small weight, such that any intersecting lines did not reduce the response of the detectors. Line termination detectors had a large weight in the region opposite the line end to ensure that responses to continuous lines were reduced.
\[ H^O = \sum_\theta \exp \left[ -\left( \frac{x_\theta^2 + \gamma^2 y_\theta^2}{2\sigma^2} \right) \right] \cos \left[ \frac{2\pi x_\theta}{\lambda} \right] \] (5.7)

\[ W^O_\theta = \sqrt{\left( \frac{(H^O - G^O_\theta)}{(N^O - 1)} - G^O_\theta \right)^2} \] (5.8)

\[ H^E = \sum_\theta \exp \left[ -\left( \frac{x_\theta^2 + \gamma^2 y_\theta^2}{2\sigma^2} \right) \right] \cos \left[ \frac{2\pi x_\theta}{\lambda} \right] \frac{1}{1 + \exp \left[ -\tau y_\theta \right]} \] (5.9)

\[ W^E_\theta = \sqrt{\left( \frac{(H^E - G^E_\theta)}{(N^E - 1)} - G^E_\theta \right)^2} \] (5.10)
Table 5.1: Features and weights (G & W) for equations 5.1 and 5.2
5.2. SYSTEM DESCRIPTION

T-junctions (T) were identified at locations where a line continuous line intersects a terminated line, which was calculated as a function of the product of O and E at each retinotopic position i:

\[ T_{\phi_O,\phi_E} = \begin{cases} 
1 & \text{if } \argmax_{\theta_O,\theta_E} \{O_{\theta_O,i} \cdot E_{\theta_E,i}\} = \phi_O, \phi_E \quad \text{and} \quad O_{\theta_O,i} \cdot E_{\theta_E,i} \geq \gamma \\
0 & \text{otherwise} 
\end{cases} \] (5.11)

where \( \gamma \) is the threshold for an intersection (\( \gamma = 0.70 \)).

In the second grouping and segmentation stage, a surface (S) was defined as the set of points within a border line. Surfaces could be joined through a grouping process, in which two set of points were judged to belong to a single object which appeared fragmented due to the presence of an occluder.

Fig. 5.2 (a) contains an illustration of the grouping process on an example stimulus, which consists of a rectangle occluded by a trapezoid. Naïve borders were initially identified by tracing the outlines of each surface contour without regard for ownership (S_1, S_2 and S_3, Fig. 5.2a). Border ownership was then calculated from the orientation of T-junctions: the unoccluded trapezoid in Fig. 5.2b owns all of its border lines, while the rectangle does not own any of the shared border lines between it and the trapezoid.
5.2. SYSTEM DESCRIPTION

Figure 5.2: Illustration of the perceptual grouping process; (a) the stimulus consisted of a trapezoid, which occluded a rectangle. The scene was initially parsed into three surfaces $S_1$, $S_2$, $S_3$; (b) ownership was assigned to the borders of each surface, based on the orientation of T-junctions (circles). (c) By extending the terminated lines of the T-junctions and searching for intersections, $S_1$ and $S_3$ were linked as a single percept to form two new surfaces, $S'_1$ and $S'_2$. (d) The borders of the final surfaces. Unowned borders are shown in grey.

Grouping involved a search for a continuation of the borders of the inferior surface on the other side of the occluder. The terminated border lines are extended beyond the T-junctions and a search took place for intersections of the extended lines, Fig. 5.2b, 5.3. Activity relating to different line continuations was summed and an intersection was identified as a cluster of activity above a threshold ($L_{Th}$) and with an area above a second threshold ($A_{Th}$). If the angle of intersection was close to perpendicular, then the area of the intersection was small, but if the angle of intersection was close to $180^\circ$, then the area of the heightened activity was large, Fig. 5.3.
Figure 5.3: Illustration of the grouping mechanism, which was used to unite surfaces into single percepts. Stopped lines were extended beyond the occluding border (light grey). (a) No intersection of the extended lines; (b) area of the intersection (dark grey) was small for perpendicular lines; (c) as the angle approaches $180^\circ$, the area of the intersection increased; (d) the largest common area was obtained for the apparent continuation of a line behind an occluder. Ellipses were shown for clarity in this figure: diffuse projections were used in the implemented system.

In Fig. 5.2b the two pre-surfaces ($S_1$ and $S_3$) were connected to form a single entity in 5.2c. Fig. 5.2d shows the final borders $B_1$ and $B_2$.

In the selection stage, each surface was represented by a saliency channel ($C_k$), which funnelled salient elements at different spatial locations across the percept to form a single competing entity, though the calculation of visual saliency was not addressed in this system (see section 6).

Competition took place between the saliency channels, resulting in the selection of the attentional target, $S^W$. The effect of attention was to alter the activity of neurons in
L (and in subsequent layers) such that the attended object produced the same patterns of activity as if it has been presented in isolation. A contained the activity of L, modulated by the winning surface, S^W:

\[ A_i = \lambda_i L_i \quad (5.12) \]

\[ \lambda_i = \begin{cases} 
1 & i \in S_w \\
0 & i \notin S_w
\end{cases} \quad (5.13) \]

In the final stage, the target was recognised by a local feature matching process, (Table. 5.2). Prior to the experiment, a set of feature detectors (\( G^F_k, k = 1 \ldots K \)) was trained. Local feature matches were computed for the attended surface:

\[ F_{k,i} = \exp \left[ -\beta F \sum_{j \in \mathcal{M}} \left( G^F_{k,j} - U_{ij} \right)^2 \right] \quad (5.14) \]

Where \( G^F_k \) was the pre-learned feature, \( U_{ij} \) was a square 10 x 10 tile from A centred at retinotopic location \( i \).

This recognition method was similar to the HMAX models of Serre et al. (2006), in the respect that recognition was dependent upon the identification of local features. However, a number of differences also existed: feature matching was confined to the attended (i.e. segmented) object in this system, while in the HMAX models, recognition takes place across the entire visual scene. Another difference is that the relative position of features plays an important part in the recognition process in our system, especially in the determination of potentially occluded features. Feature position is only implicitly used in the HMAX models.

The recognition algorithm used both the identity of local features, and their relative positions. Border ownership made two contributions to the recognition process: features along unowned borders could be rejected (i.e. not contribute to the recognition process, and features which were likely to lie behind occluders were able to make a small contribution to the activity of the recognition neurons). Recognition neurons pooled inputs from the relevant feature detector neurons \( F^G_{k} \) from pre-learned relative locations \( d \), Table 5.2. If a feature was expected at a location which was occluded, then an occlusion neuron
at that position increased the activity of the recognition neuron by $\Delta O$. $I$ represented the quality of the match for each set of features, which was calculated from the sum of the relevant feature detector neuron activities:

$$I_{li} = \sum_{k,d \in \Lambda} F_{k,i+d} + \Delta(i + d)$$  \hspace{1cm} (5.15)

where $\Delta = \begin{cases} 
\Delta_O & i + d \in \text{occluding surface} \\
0 & \text{otherwise}
\end{cases}$

Where $\Lambda$ was the set of feature indices ($k$) and their corresponding relative locations ($d$) that identified the object $l$, $\Delta_O$ was the increase of activity caused by a potentially occluded feature.

The identity of the target was given by the recognition cell with the greatest activity:

$$z = \arg\max_l \{ I_l \}$$  \hspace{1cm} (5.16)
5.3 Results

Training  Local features consisted of a 10 x 10 pixel patch of activity from L. In Experiment I, in which the basic functionality of the system is demonstrated, local features were hand-selected. In Experiment II, the positions at which local features were extracted were calculated as the local maxima of the sum of the activities of the line termination detectors $E_\theta$. The activity of the line terminations was greatest at corners and areas of high curvature. Examples are shown in Table. 5.6.

5.3 Results

These experiments addressed border ownership and occlusion in object recognition tasks. Experiment I contains two demonstrations of the processes of grouping and recognition through border ownership assignment. Experiment II contains statistical results for the recognition of occluded objects and shows that the treatment of ownership and occlu-
sion significantly improved the object recognition rate when compared with a situation where occlusion information was not considered.

For these experiments, two objects were presented in the visual scene, one occluding the other. Attention was externally driven to select the lower object, which was then recognised. Because the analysis focuses on occluded figures, the upper object was not processed.

5.3.1 Experiment Ia

![Table 5.3: Stimuli and local features used in Experiment Ia. Side image: combined rectangle, trapezoid and small triangle](image)

This experiment investigated two aspects of System II: the ability to create a perceptual grouping from an object which appeared as two separate fragments because of the presence of an occluder, and the ability of the system to use border ownership information to correctly distinguish between a partially occluded object and a mosaic object by correctly determining border ownership. Stimuli are shown in Table 5.3.1 and consisted of a large triangle, a rectangle, a trapezoid and a small triangle. The trapezoid, the rectangle and the small triangle could be arranged to have the same appearance as the large triangle occluded by the rectangle. Black squares in the images in Table 5.3.1 show the positions of the local features which were used for object recognition (4 per object). In this case, feature locations were hand-selected at positions which increased the difference in response when grouping was successful and when it was not successful. Two local feature positions on the large triangle were chosen such that they would be hidden by the rectangle in the occluded condition (Fig. 5.4).

The results for grouping and border ownership assignment are shown in Fig. 5.4. Three cases were examined: when border ownership was used to create groupings (A),
when it was not used to create groupings (B), and a control where border ownership was unambiguous (C). An image of each stimulus is shown in column 1 of Fig. 5.4. Columns 2 – 5 show the border ownership activity for each percept in turn. When border ownership was not considered (B), border ownership was incorrectly assigned, while it was correctly assigned in cases A and in the trivial control case C.

In A, the fragments of the large triangle were correctly grouped and the ownership of the shared border between them and the rectangle was correctly ascribed to the rectangle A2, A3. In case (B), where border ownership information was not considered (i.e. \( T = 0 \) in Eqn. 5.11), the large triangle was incorrectly classified as two separate percepts: a trapezoid (B2) and a small triangle (B5). In the control case, each object was correctly calculated to own all of its borders.

Figure 5.4: Grouping and border ownership assignment for the three test conditions. The leftmost column (1) shows the stimulus and the other figures in each of the rows (2 – 5) show the border ownership for each of the figures prior to selection. Positive activity indicated that border ownership was assigned to the attended surface and negative activity indicated that border ownership was not assigned to the attended surface; (A): border ownership when an occluder was present; (B) border ownership when no occluder was present; (C) border ownership for an unambiguous arrangement.
Figure 5.5: Activity of each recognition neuron \( I_i \) when each surface was attended (indicated by trace colour). The shapes to which the recognition neurons were trained are shown at the top of the figure; A: response of each neuron to each shape when border ownership information was used; B: when border ownership information was not used; C: unambiguous control case.

Each percept in Fig. 5.4 was attended and the activity of the recognition neurons \( I \) is shown in Fig. 5.5. When border information was considered \( (A) \) then the large occluded triangle was successfully recognised, (black trace). Activity was also high in the trapezoid and small triangle recognition neurons as these share common features with the large triangle. When no border ownership information was considered \( (B) \), the grouping was incorrect and the large triangle was treated as two separate entities: a trapezoid and a small triangle. The activity of the large triangle recognition neuron was low for all stimuli. The lower portion was mistakenly recognised as a trapezoid and the top of the large triangle was misclassified as a small triangle. \( C \) shows the recognition neuron responses when there was no occlusion of the targets: all recognition neurons respond strongly to the stimuli to which they were trained, and weakly to the others, except where there were some common features.

This experiment demonstrated the importance of correctly parsing the scene prior to
the object recognition process and illustrated a basic need to consider perceptual organisation in machine vision tasks. The next part of this experiment investigated with effect of border ownership in the object recognition process. Again, a specific example is used, while results for a larger set of objects are shown in Experiment II.

5.3.2 Experiment Ib

The goal of this experiment was to investigate the effect of border ownership on the object recognition process. A single pair of objects were used. Stimuli are shown in Table 5.4 and local features positions in Table 5.5.

The experiment took place in four trials, during which the local features which were used to recognise the triangle. Positions of the local features were hand-selected. Two recognition neurons were trained prior to the experiment: one for the square and one for the triangle.

<table>
<thead>
<tr>
<th>Trial</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stimulus</td>
<td><img src="image1" alt="Stimulus" /></td>
<td><img src="image2" alt="Stimulus" /></td>
<td><img src="image3" alt="Stimulus" /></td>
<td><img src="image4" alt="Stimulus" /></td>
</tr>
</tbody>
</table>

Table 5.5: Local feature positions in Experiment Ib

![Local feature positions](image5)

Recognition neuron responses were compared when border ownership was considered ($\Delta_O = 0.6$) and when it was not considered ($\Delta_O = 0$). Fig. 5.6 shows the activity of both recognition neurons when the triangle was attended for each image in Table 5.4. When border ownership information was used, the response of the square recognition neuron remained much lower than when it was not considered (a). This was because the corner features which appeared on the border of the occluded triangle were not considered to be...
part of the triangle. These features produced strong activation of the right angle feature
detectors and consequently the square recognition neuron.

The ratio of the activity of the triangle recognition neuron to the square recognition
neuron is shown in Fig. 5.7, for values of $\Delta_O = 0$ (a) and $\Delta_O = 0.6$ (b). The ratio
was consistently higher when border ownership was considered, which indicated that the
distinctness of the objects was increased by the use of border ownership information in
the case where $\Delta_O = 0$. The distinctness was not increased when $\Delta_O = 0.6$, which
indicated that automatically providing activity for a potentially occluded feature was not
always useful.

Figure 5.6: Activity of the triangle recognition neuron (black) and the square recognition
neuron (gray) while the triangle in Table 5.4 was attended; recognition neuron activity
when border ownership information was used (BO) and when it was not used (NBO)
during the recognition process. (a) $\Delta_O = 0$; (b) $\Delta_O = 0.6$
5.3. RESULTS

Figure 5.7: Ratio of the activity of the triangle recognition neuron to the activity of the square recognition neuron for each shape in Table 5.4 when border ownership information was used (circle) and when border ownership information was not used (square). (a) $\Delta_O = 0$; (b) $\Delta_O = 0.6$

This experiment demonstrated the importance of correctly parsing the scene prior to the object recognition process and illustrates a basic need to consider perceptual organisation in machine vision tasks. Hand-crafted examples were used to illustrate the potential use of a system which can distinguish between owned and unowned borders. The next experiment generalises this result with a more generic set of objects.

5.3.3 Experiment II

Object recognition was tested on a larger set of objects when border ownership was considered $\Delta_O \neq 0$ and when it was not considered $\Delta_O = 0$. The effect of changing the value of $\Delta_O$ ($0 - 1$), the contribution of the occlusion neurons to the recognition neuron activity, the number of local features ($1 - 8$) and the number of objects ($20 - 50$) was also investigated. The set of objects and the positions of the local features are shown in Table 5.6. All objects were composed of combinations of the same four corner features such
that no object could be recognised from a single feature, but instead recognition relied on the successful identification of a number of features at different relative locations.

Table 5.6: Objects (white) and local feature positions (black) used in Experiment II.

Two objects were randomly selected and superimposed such that one object occluded the other. Cases in which the occluder obscured all of the local features of the underlying object were discounted (recognition would be impossible with this system), as were cases in which poor T-junction resolution provided insufficient information for reliable border ownership assignment. Examples of counted and discounted stimuli are shown in Table 5.3.3.
The lower object became the target of attention and the recognition response was measured for all recognition neurons (I). As a control, the response of the recognition neurons was measured when border ownership information was not considered. During the recognition task, features along the unowned borders of the attended object did not contribute to the activity of the recognition neurons: e.g. the black lines in the second row of Table 5.3.3 were not recognised as potential features.

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Border ownership</th>
<th>Control (no border ownership)</th>
<th>Discounted stimuli</th>
</tr>
</thead>
<tbody>
<tr>
<td>![stimuli image]</td>
<td>![border ownership image]</td>
<td>![control image]</td>
<td>![discounted stimuli image]</td>
</tr>
</tbody>
</table>

Table 5.7: Examples of stimuli and the borders in Experiment II. White borders were owned and black borders were unowned by the attended surface. Discounted stimuli are shown in the bottom row.
Figure 5.8: Fraction of correct recognitions as a function of the number of local features used during recognition. Trace colour indicates the number of objects used in the experiment. (a) $\Delta_O = 0.5$: a contribution of 50% activity was made by the occlusion neurons for potentially missing features; (b) $\Delta_O = 0$: the occlusion neurons had no effect on the recognition process;
5.3. RESULTS

Figure 5.9: Ratio of recognition neuron responses (border ownership vs. no border ownership); (a) for different numbers of local features; (b) more explicit comparison of (a) for 6 and 7 local features.

Fig. 5.8a, (b) show the fraction of correctly recognised objects as a function of the number of local features used during recognition for different numbers of objects.

We compared the case in which border ownership information could be used to exclude features which were formed by owned borders with the case where border ownership was not considered (continuous line vs. dashed line) for $\Delta_O = 0.5$ and $\Delta_O = 0$. In both cases, the fraction of matches was greatest when the smallest set of objects was used, though the recognition rate did not decrease monotonically as the number of objects.

111
was increased, Fig. 5.8a. The effect of considering border ownership during recognition was significant and was not sensitive to using very small data sets. The greatest response was obtained from the use of six local features, which may reflect the fact that most of the objects used in this experiment had approximately six corners. As the number of objects was increased, the case where $\Delta_O = 0.5$ led to a greater recognition rate compared with $\Delta_O = 0$, i.e. the contribution of a potentially occluded feature became more important. For 8 local features, the significance of the improvement for $\Delta_O = 0$ was measured with an independent two-sample t-test ($p = \{.92, .95, .94, .97\}$ for $N = \{20, 30, 40, 50\}$).

This is examined in more detail in Fig. 5.9. Fig. 5.9a shows the ratio of correct recognitions when border ownership was used to the number of correct recognitions when border ownership was not used during recognition for $\Delta_O = 0.5$ and $\Delta_O = 0$. The effect of $\Delta_O$ increased when 50 objects were used and decreased as the size of the object set decreased, Fig. 5.9b.

Similar responses were obtained for values of $0 < \Delta_O < 1$, and performance severely degraded when $\Delta_O \geq 1$ (i.e. when a feature which was potentially behind an occluder made at least the same contribution to the activity of the recognition neurons as if it had been positively identified).

The measured SNR for these objects may be surprisingly small, however this test is particularly severe for the system: the test stimuli contain the same local features, and can only be recognised by the relative positions of these features. Frequently, occlusion was too great to allow an object to be unambiguously identified.

### 5.4 Discussion

This system was inspired by the functional stages of visual processing and attention in primates. The processes of scene segmentation, selection and recognition are instrumental in preparing appropriate behaviours in a world in which objects are rarely encountered in isolation.

This system demonstrated that a significant advantage can be gained in the robustness of object recognition when border ownership was considered. Object recognition was achieved through local feature matching at a set of relative positions. Because occluded
regions were explicitly represented in the object recognition process, it was possible to compensate for local features which were likely to be present at occluded locations. This approach resulted in a reduction of the SNR when matching a small number of local features. If the scene was incorrectly segmented before the selection stage, then the incorrect object may be recognised due to poor grouping while, if border ownership was not calculated, then the partially visible object will be treated as an entire mosaic figure.

This system represents a proof of concept for the relevance of border ownership information in object recognition. While there are pre-existing implementations of percept-based object recognition systems, the inclusion of border ownership information in this system is novel.

While previous models of object recognition have successfully recognised an object in a purely bottom-up manner, one of the main differences of this system was that it produced an understanding of the scene in terms of objects, their border locations and the locations of potential occluders.

While the recognition system has been demonstrated to work on geometrical figures, a number of steps would be necessary before a more general system could be developed on this principle. The perceptual organisation processes are conceived of as a generic front-end for an object recognition system, such as H-MAX of SIFT, though the improvement from such a system has not yet been quantified. While T-junctions are excellent cues for uncovering the order of geometrical figures, they are not as useful in real scenes, where binocular disparity and motion-based segmentation may be of more use (Grossberg, 1994). An extended discussion of the generalisation of this model is contained in Section 7.
6.1 Introduction

The goal of this system was to investigate the role of the shape-based guidance of visual attention in machine vision applications. System III addressed the shape-based guidance of attentional selection. As in System II, border ownership information was used to increase the robustness of responses under cluttered conditions.

The shape of a percept can drive the attentional selection. Though the shape of an object may be instrumental in recognising it, some shape information is available prior to explicit recognition. This is equivalent to the distinction made between recognition and seeing by Grossberg (1994). In humans, completion of occluded object regions takes place rapidly, and a distinction can be made between an occluded and a mosaic shape during the selection process (Rensink and Enns, 1998).

Surface vectors, which described the relative locations of the border points of a surface were used to describe the shape. A surface-centric working memory represented target and distractor shape features, and selection was guided towards the target though the modulation of the activity across a visual saliency map. Information about the identities of local features was not used to guide selection, but only information about the relative locations of the border points across a surface.
The system was tested with both passive and active visual search tasks. During active visual search, targets which lay across the edge of the visual field were also treated as though they were occluded along the border. By treating the edges of the visual scene as though they are unowned by any surface, more robust representations of these objects could be obtained.

### 6.2 System Description

![Flow diagram for System III](image)

**Figure 6.1: Flow diagram for System III, see text for details**

An overview of the system is shown in Fig. 6.1. The system implemented three of the four stages from the visual attention framework in Fig. 3.0. Early features of line orientations, line terminations and T-junctions were extracted in the first stage. Perceptual surfaces were formed, border ownership was calculated and surface shape descriptor vectors were computed in the second stage. In the third stage, shaped-based guidance and inhibition of return were used to bias selection towards a particular shape. The behavioural relevance of different objects was represented on a visual saliency map, and a
competition for attention took place between different surfaces. An object recognition stage was not included in this implementation and both active and passive search processes were implemented in this system. The substantial difference between System II and System III lay in the addition of a shape-based guidance mechanism.

The early feature extraction and grouping processes were the same as in System II. A virtual ‘eye’ was free to explore a larger visual scene \(W\). Perceptual surfaces \(S_j\) were extracted and border ownership \(B_j\) was assessed of each item \(j\) in the visual scene, as in System II. When active vision was used, any border which lay at the edge of the visual field was not owned by the corresponding surface.

Surface shape was described by the vector \(d_j\), which contained a list of the relative locations of the most prominent points within each surface. These points were selected from activity maxima of the map \(P\), which was calculated as the sum of the activities of the line termination detectors \(E\), with a minimum spatial separation \(m\) between points. This measure tended to prioritise locations with high spatial frequencies, such as corners:

\[
P = \sum_n E_n
\]  

Fig. 6.2 shows the locations of the points in \(d_j\) for two objects: a rectangle and a notched rectangle. The representation of the surface as a list of relative point locations allowed shapes of different sizes to be compared, though the property has not been tested.

![Figure 6.2: Black dots indicate the locations of the points in \(d_j\) on a pair of stimuli consisting of a rectangle and a notched rectangle](image)
The activity of the working memory map (M) distinguished between the shapes of targets and distractors. The topology of M was surface-centric: activity at a point represented the relative feature locations across a surface and did not contain information about the location of the surface in retinotopic space. Activity was positive at locations which were associated with distinct rewarding features and activity was negative at locations which were associated with distinct punishing features. In this implementation, the working memory only represented the difference between targets and distractors. Table 6.1 shows the activity of M when the system was biased towards different targets and away from different distractors.

<table>
<thead>
<tr>
<th>Target</th>
<th>Distractor</th>
<th>Working Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="a" alt="Image" /></td>
<td><img src="b" alt="Image" /></td>
<td><img src="c" alt="Image" /></td>
</tr>
<tr>
<td><img src="d" alt="Image" /></td>
<td><img src="e" alt="Image" /></td>
<td><img src="f" alt="Image" /></td>
</tr>
<tr>
<td><img src="g" alt="Image" /></td>
<td><img src="h" alt="Image" /></td>
<td><img src="i" alt="Image" /></td>
</tr>
<tr>
<td><img src="j" alt="Image" /></td>
<td><img src="k" alt="Image" /></td>
<td><img src="l" alt="Image" /></td>
</tr>
</tbody>
</table>

Table 6.1: Working memory maps different objects and distractors: positive working memory activity is shown in white and negative in black; (a) diamond target and notched diamond distractor; (b) notched diamond target and diamond distractor; (c) rectangle target and notched rectangle distractor; (d) notched rectangle target and rectangle distractor

The working memory activity was computed offline prior to the experiment as the difference between the target and any distractors. The use of diffuse activity introduced some invariance to the precise location of the surface borders, while still making the system selective to the same overall shape. Activity was positive for the corresponding target locations and negative for the distractor locations:

The saliency of each surface ($C_j$) was calculated as the sum of the working mem-
ory activities at the locations $d_j$. Saliency was calculated as the sum of matching border points along the target shape, and was reduced by each instance of a point along a distractor shape. In this implementation, objects were of a single size, and the issue of size invariance was not addressed, though see Section 6.4.

For each perceptual surface ($j$) the saliency was calculated as the sum of the working memory activities at the location of each point ($l$) on $d$. Potential features in occluded locations also contributed to the saliency ($\Delta$), and inhibition of return ($R_{IOR}$) reduced the saliency of recently attended objects.

$$C_j = R_{IOR}^j \cdot \sum_l K_{jl}$$  \hspace{1cm} (6.2)

$$K_{jl} = \begin{cases} 
\Delta_OM_{d,jl} & d_{jl} \in \text{occluding surface} \\
M_{d,jl} & \text{otherwise} 
\end{cases}$$

where $\Delta_O$ was the gain of the saliency contribution from occluded regions.

The surface points $d_j$ were mapped onto the working memory in a manner which was robust against occlusion. This was achieved by using one of the corners of the bounding box for the surface as the origin for the mapping process. The choice of corner depended upon the location of any occluded regions on the surface. Table 6.2 contains four examples of this process, in which different choices of origin were necessary in order to correctly map the surface shape $d_j$ onto the working memory. When border ownership was not used, then the choice of origin was frequently incorrect. For example, if the occluded rectangle in column of Table 6.2 (a) was mapped onto the working memory using the top-left or bottom-left corners, then the matching feature (the strongly inhibited bottom-right corner) would not have been correctly positioned.
Table 6.2: The choice of origin for the mapping process depended upon the occluded regions of the surface; top row: images of a set of surfaces; second row: borders for each surface when border ownership information was used: unowned borders are shown in black; third row: position of the origin for a suitable map when border ownership information was used; fourth row: borders when no ownership information was used; fifth row: position of the origin when no border ownership information was used. The choice of origin depended upon the location of unowned borders on the surface; (a) the notched rectangle was unoccluded and owned all of its borders: the default axis (top-left) was chosen for mapping the shape onto the working memory in both ownership cases; (b) the rectangle appeared at the left edge of the visual field and the left border was unowned: the top-right was used as the origin for mapping the shape onto working memory in the case with border ownership, while the origin was incorrect placed when no border ownership information was used; (c) the upper half of the rectangle lay outside the visual field: the bottom-left axis was used as the origin for mapping to working memory when border ownership information was used. When no border ownership information was used, the surface was incorrectly mapped onto the working memory; (d) the right half of the rectangle lay outside the visual field: the top-left corner was used as the origin for mapping the shape to working memory in both cases

Points which were potentially occluded also contributed to the saliency of an object, in the same manner as in System II (see Eqn. 5.15). Occluding surfaces were identified, and the magnitude of $\Delta$ was the sum of the working memory activity across the surface, with a gain of $\Delta_O$. Points which lay within a short distance ($l_{\text{min}}$) of an unowned border did not contribute to the calculation of attentional saliency: this prevented the system from treating the mosaic shape as a likely shape for an occluded surface.

Inhibition of return was implemented as a leaky integrator in a world map space
6.3 Results

This section presents three experiments, which tested the ability of the system to correctly assign saliency to partially occluded shapes. Experiments I and II took the form of passive vision tasks and Experiment III took the form of an active visual search task.

The dimensions of the visual field were 256 x 256 pixels and in Experiment III, the visual world was of size 512 x 512 pixels. The minimum distance \((m)\) between points on \(d\) was 10 pixels.

6.3.1 Experiment I

The first experiment tested the ability of the system to correctly calculate the saliency of different shapes in a cluttered arrangement, such that unowned borders were considered to provide no information about the shape of an object. Two sets of objects were used in this experiment: the first set consisted of a diamond and a notched diamond, and the second set consisted of a rectangle, a notched rectangle, and a circle, shown in Figs. 6.3a & 6.4a. The system was biased towards different targets, and attentional saliency was measured for each object in the visual scene. Both sets contained ambiguous cases, in which the mosaic image of a partially occluded object had the same shape as another object. In these cases, it was desirable that the system should be biased towards the completed object and not towards the mosaic shape. Such an interpretation reflects a more reasonable understanding of the scene, in terms of what is normally experienced:
while it is possible that the mosaic shape is the exact object shape, in such instances it is generally the case that the object continues behind the occluding surface and it would be erroneous to assign the occluding border to the inferior object. Results were compared with a control case in which no border ownership information was available.

Figure 6.3: First set of objects; (a) the arrangement of the objects; (b) normalised saliency of each object in (a)

Fig. 6.3b shows the normalised saliency of each numbered object in Fig. 6.3a. Objects 1 to 4 and 7 to 10 were either unoccluded, or occluded in a manner which did not make
their identities ambiguous, and the system successfully prioritised these objects regardless of the use of border ownership.

Objects 5 and 6 represented ambiguous cases in which the occluded mosaic image was the same as another object. When border ownership was used, these objects were treated as completed diamonds, while the magnitude of the saliency of the mosaic shape (notched diamond) was low. The unowned border did not contribute to the shape-based saliency, resulting in a reduction in the magnitude of the saliency in these cases. When border ownership information was not available, the objects were treated as notched diamonds: the shared borders with objects 2 and 8 were treated as though they belonged to objects 5 and 6 respectively.

It can be inferred from cases 3 and 9, which contained partially occluded objects whose identity was unambiguous, that a simple increase in saliency for an unowned border would not be sufficient to achieve the same distinction between mosaic shapes and partially occluded shapes.
The second set of objects consisted of rectangles and notched rectangles. A circle was also used as a distractor, but the system was not trained to avoid this shape. Ten cases were examined, both when selection was biased towards the rectangle and towards the notched rectangle, Fig. 6.4b. The shapes of objects 2, 3, 4, 8, 9 & 10 were unambiguously resolved by the system, and saliency was successfully assigned to the correct shapes, regardless of the use of border ownership information. The magnitude of the saliency was low for the
circle in case 1, indicating that it had little similarity with either of the shapes to which the working memory had been trained. Object 5 was occluded to the extent that its identity was completely ambiguous, which resulted in it having a saliency of small magnitude. When no border ownership information was used, the magnitude of the saliency was zero, since no part of the object was mapped onto an active region in working memory.

In case 6, the identity of the object depended upon the successful use of border ownership information to infer that the object was actually a rectangle, with an occluding circle hiding the identifying shape features, which was only achieved in the condition where border ownership information was available. When no border ownership information was available, shape 6 was treated as the mosaic shape of a notched rectangle.

Case 7 also resulted in an incorrect response when the system did not use border ownership information. The notch in the mosaic shape caused by object 3 caused the system to interpret object 7 as a notched rectangle when no border ownership information was used.

A limitation of the system can be seen from case 9 in Fig. 6.3b and in case 5 in Fig. 6.4b. In both of these cases, the presence of an occluder interfered with the calculation of the correct magnitude for the saliency. The influence of the occluded portion of the image (Δ), reduced the magnitude of the saliency of the notched diamond in case 9 from Fig. 6.3b. Shape 5 of Fig. 6.4b was treated as a rectangle as a result of the contribution of Δ to the saliency. In this instance, this was due to the proximity of the upper border of shape 8 to the region which represented the notch in the working memory: regions which were within a short distance of an occluding border did not contribute to the saliency.

6.3.2 Experiment II

This experiment tested the ability of the system to guide selection when objects were located across the edge of the visual field. Stimuli consisted of a square and a notched square, Fig. 6.5a. Any common border between the edge of the visual field and an object was treated as being unowned by the object.

The saliency of each object was again calculated when the working memory was bi-
ased towards one object and away from the other. Results were compared with a case in which the borders at the edge of the image were considered to be owned by the objects.

![Image](image_url)

**Figure 6.5:** Stimulus and priorities for Experiment II; (a) stimulus image with partially visible shapes at the edge. Numbers indicate the order of the objects in (b); (b) saliency of each object

Fig. 6.5b contains a plot of the saliency for each object in Fig. 6.5a. Cases 1, 2 and 12 presented unoccluded or almost unoccluded shapes, and the calculated value of their saliency was only slightly affected by the use of border ownership information.
In cases 3 and 4, the system was successfully biased towards the shape when border ownership was used, and a saliency of zero was calculated when no border ownership information was used: in both cases, the shape had no overlap with the working memory activity.

The identities of objects 7 and 8 were ambiguous, and very small portions of objects 5 and 6 indicated that they were notched squares. Regardless of the use of border ownership, the system was poor at resolving these shapes, resulting in their priorities having a small magnitude. Saliency was small for cases 5 and 6 because of the large square-bias contribution of the occluded portion of the image ($\Delta$).

In cases 10 and 11 the shapes were biased incorrectly when no border ownership information was used: both objects were treated as rectangles and points along the longest line were mapped to active positions on the working memory associated with the notch. A similar situation occurred with shape 9, though there was little overlap with the working memory in this case. In each of these cases, saliency was correctly calculated when border ownership information was used.

### 6.3.3 Experiment III

The ability of the system to direct search towards a particular shape was tested in the final experiment. Again, stimuli consisted of squares and notched squares, Fig. 6.6. Cues were spaced so that 16 cues appeared at the periphery of the visual field. The scene was explored for 100 saccades though eye movements between different objects and exploration was facilitated by inhibition of return.
Four cases were examined in this experiment: in case 1, selection was guided towards the notched square and border ownership information was used during the biasing process. In the case 2, selection was again guided towards the notched square, but no border ownership information was available. In case 3, selection was biased towards the square, and border ownership information was available. In case 4 selection was again biased towards the square, but with no border ownership information available. In all cases, the system was unbiased towards the circle.

Fig. 6.7 shows the number of saccades to each object during the four cases. In both cases (i.e. case 1 vs. case 2 and case 3 vs. case 4), there was a greater number of saccades to the target object when border ownership information was used than when it was not used. In the case 4, the target was selected < 50% of the time. When the square was the target (case 3 and 4), erroneous saccades were not made to the notched square, but rather to the circle.
6.3. RESULTS

Figure 6.7: Plot of the total number of saccades to the notched square and the square for each vase; the system was biased to select the notched square in cases 1 & 2, and the square in cases 3 & 4.

Fig. 6.8 shows the scan paths for the first 10 saccades of case 1 and case 2. In Fig. 6.8a (case 1, with border ownership), distractors were rarely selected. In Fig. 6.8b (case 2), occluded squares were selected three times. All of these errors involved the mosaic image of a square with a circle occluding the bottom-right corner being mistaken for the target of a notched square.
6.3. RESULTS

(a) Case 1: with border ownership

(b) Vase 2: no border ownership

Figure 6.8: Images of the scan paths for cases 1 and 2; (a) when border ownership information was used; (b) when border ownership information was not used
6.4 Discussion

This chapter has proposed a machine vision algorithm for the guidance of attentional selection by surface shape.

It was demonstrated that the system could be used to bias selection towards a particular shape in a manner which was robust to occlusion. Surface shape was represented as a set of feature positions in a set of working memory maps. Border ownership information, which was inferred from T-junctions, was used to distinguish between owned edges (which contributed the saliency of a target) and unowned borders (which did not contribute to the saliency of a target).

The first experiment produced a qualitatively similar result to Experiment I from Rensink & Enns (1998) (Fig. 2.5). The system used surface shape representations that were robust against occlusion and the system successfully distinguished between mosaic and unoccluded surfaces when calculating saliency.

In Experiment II, we demonstrated that this system could also be applied to selection tasks in which objects were partly visible at the edge of the visual field. By treating the edge of the visual field as an unowned border for all objects, shape-based guidance information could still be used to drive selection towards the target shape. This property of the system will be relevant to machine vision applications, in which objects are commonly partly visible at the edge of the visual field.

The system was tested with an actual search task in Experiment III, where it was demonstrated that performance in shape-based guidance was improved when border ownership information was used.

The contribution of this system lies in the way in which it uses robust percepts to guide attentional selection. It is through the use of border ownership information that an invariant representation of shape can be generated, and then used to drive the search process. It was demonstrated that shape-based guidance did not become confused by clutter in this system.

The guidance of selection by shape was purely based upon the difference between the target and distractors, and so the system should perform poorly if unfamiliar distractors were introduced. While this was sufficient for the cases shown, in which there are clearly
specified targets and distractors, a more developed system would be necessary for cases in which there are several distractors.

A second way in which a greater degree of generality could be achieved by the system is in the inclusion of a stronger segmentation system. These are discussed in detail in Section 7.

Future work could include a mechanism by which the increase in saliency is gated by the condition that a shape matches the target shape to some degree, i.e. that the minimum distance between a point on a shape $d_j$ and a location on the ideal target does not exceed a certain threshold. By using $d_j$ to describe surface shape, it would be possible to extend the system towards scale invariant search tasks. The present scheme is only applicable to a single situation, in which targets and distractors are clearly known. It would also be desirable to have a more non-linear saliency accumulation function: very small details in the shape should be capable of causing a shape to be treated as either a target or a distractors. Psychophysical evidence also indicates that shape representations are 3 dimensional (Enns and Rensink, 1990).

For such a selection system to be truly useful, it must be demonstrated that it improves search efficiency. At present, the implementation takes several seconds per stimulus, which a more naïve template-matching routine may succeed in recognising the target object faster than a two-stage guided approach. This remains a possible course for future investigation.

This work presented in this chapter united System I, which addressed the guidance of selection towards rewarding targets in a visual search task, with System II, in which early object representations which were robust against occlusion were developed for object recognition.
This chapter contains a summary of the research presented in this thesis. The work will be compared with similar models of attention for machine vision applications. The major contributions of this research are detailed and possible courses for future work are outlined.

### 7.1 Summary

Given that the accurate recognition of objects is crucial for guiding intelligent behaviour, and that these objects are very frequently encountered in clutter, it seems unsurprising that the primate visual system has developed an early perceptual grouping strategy which provides robust object representations. Such an observation contributes to a view of primate vision in which representations are intimately linked with future behaviour, even at some of the earliest processing stages.

Functionally, attention allows sensory information related to a single entity to be processed without interference from the surrounding world. One of the immediate effects of attention is that it allows the target object to be recognised, something which is usually impossible prior to an attentional shift.

This thesis has presented a framework for the generation of visual search behaviour in machine vision applications. This framework was based on functional stages of processing in the primate pathway for visual search. In particular, the perceptual organisation
stage and its role in the selection and recognition processes has been heavily inspired by the current psychology and physiology literature. Three systems have been implemented to examine different aspects of this framework.

System I addressed the problem of generating behaviourally useful representations of sensory information, upon which reinforcement learning can operate. Internal representations of the sensory world were learned online, and a mapping was learned between representation and rewarding behaviour. It was demonstrated that by allowing behavioural goals to modify the internal representation online, system resources could be used to increase the reward harvesting rate.

System II addressed the interaction of perceptual grouping mechanisms and attention to generate object representations with increased robustness to occlusion from clutter in the environment. The use of border ownership information was shown to significantly improve object recognition, which was achieved with an object based position-sensitive local feature matching process.

System III examined the use of a shape based guidance mechanism in the selection process. Using the same selection mechanisms as System II, this was demonstrated to guide selection towards shapes in a manner which was robust against occlusion, a process which would be impossible without both early shape descriptions and a representation of border ownership.

### 7.2 Comparison of Models with Previous Work

Few models of object-based visual attention for machine vision applications exist. Among the most significant of those have been the systems of Sun & Fisher (2003; 2008). These models have been reviewed in Section 2.5.1. One similarity which exists between our system and theirs’ is that the competition for attention took place at the level of segmented surfaces. By using an implementation of the biased competition hypothesis, competition took place at many different levels of representation in their system, and selection converged upon the most distinct entity. Selection was purely driven by bottom-up saliency, and there was no consideration of border ownership or the possible role of occlusion when calculating visual saliency.
The systems in this thesis have differed in the respect that competition took place on a single map, in which activity represented the perceived behavioural priority of attending to a target (in System I & III). The systems in this thesis have focused on generating purposeful behaviour, and visual search in a taskless condition has not been addressed (i.e. except as a control). The second substantial difference between our systems and theirs lay in our inclusion of border ownership representations in the guidance of selection, and in the object recognition process.

The systems of Sun & Fisher have not been extended beyond the selection of an attentional target, while our framework also includes the high-level processes of object recognition. The final result in their system was in the generation of scan paths. Selection has not been the end point in our systems, but rather the use of attention beyond the directing of gaze has been considered in the object recognition task of System II.

A model by Walther & Koch (2006) also included a grouping mechanism, though this was achieved through feedback after the competition for attention has taken place, Sec. 2.5.1. Their system has provided a precedent for the object recognition experiments in System II. The authors questioned the purpose which visual attention can serve in machine vision systems, and demonstrated an improvement in object recognition performance when an object-based attentional window was used. We have provided a similar result in System II, though our model has the advantages of utilising occlusion and border ownership information to distinguish between occluded and mosaic objects. Their model has the significant advantage of operating on natural images, which are unsuitable for the systems presented in this thesis. This was due to the fact that objects were manually segmented in their systems, and there was no border ownership processing. The requirements to make the systems presented in this thesis operate on real images is discussed in Section 7.4.

Many machine vision methods address the problem of recognition with purely feed-forward architectures which bypass the need for attentional selection. By using a great number of feature detectors, a redundant object recognition signal can be obtained, even when large portions of the image are absent. Such methods (e.g. SIFT and HMAX) have been hugely successful in object recognition tasks, but the scene is represented with little
reference to the physical world or to future behaviour. Instead, this thesis proposes that future work in behaviour generation and learning could benefit from the use of a more complex perceptual organisation of the visual scene.

The approach presented in this thesis brings another benefit in machine vision applications: during active visual search tasks with a Cartesian retina, part of an object is often visible at the side of the visual field. Prior to selection, potential targets may be partially visible at the edge of the visual field. There is a clear argument that it is beneficial to interpret regions of the image which do not appear as being occluded (i.e. that the border of the object surface which is caused by the edge of the sensory is unowned), in order to prevent the half-object from being interpreted as a new and distinct shape.

7.3 Contributions

This work has made a number of contributions to the field of artificial intelligence, image processing and scene understanding. These systems represent early work on creating coherent representations for the generation of goal-directed behaviour. To summarise the novel contributions of this work:

- the first system to use self-organising maps to encode a representation of visual feature in a visual search system, and in which a working memory with the same topology encodes behaviourally relevant locations on the SOM.
- the first implementation in which selection can be biased towards shape, while distinguishing mosaic and occluded shapes
- the first implementation of an object recognition system which explicitly addresses occluded regions of the image

7.3.1 Self-Organising Internal Representations

With System I, a reward harvesting system was developed which made overt attentional shifts to cues on an invisible background. The SOM learned a behaviourally useful representation of sensory information without external supervision, which avoided the
need to create hard-coded recognition cells for feature detection before the system was initialised. Instead, this method allowed the system to develop its own representations, by training the SOM with attended stimuli. Stimuli which were rewarding were preferentially represented on the SOM and it was demonstrated that this method of online learning improved the reward harvesting rate in a simple task.

The topological mapping of feature space, which the SOM provided, was exploited during the reward harvesting task. A working memory layer with the same topology encoded the expectation that attending to a location on the SOM would result in a reward. By diffusely updating activity in the memory layer after each attentional shift, the system was driven to investigate whether rewards could be obtained in nearby regions of feature space.

### 7.3.2 Reward Harvesting with Limited Resources

System I was used to investigate a general strategy for successful reward harvesting by performing a parameter search, with the goal of determining and interpreting the parameter values which resulted in the greatest reward harvesting rates. The SOM was trained online with the attentional target, which caused the map to preferentially represent the rewarding cues, as these were the most frequent targets of saccades. If the system became unable to successfully predict rewards, then the plasticity of the map was increased. We have shown that the system successfully converged upon a stable solution which allowed it to successfully harvest rewards from the environment.

### 7.3.3 Perceptual Organisation

System II, which was heavily inspired by possible functional roles for border ownership selective cells in visual processing, was used to investigated the possible interaction between object-based attention and object recognition. Different representations of owned and unowned borders allowed the system to perceptually group regions of an object which appeared separate due to the presence of an occluder. This grouping process took place prior to the competition for attention.

The effect of representing border ownership upon the object recognition process was
also tested. Using this information, it was possible to infer the existence of occluded parts of the target object. This allowed features which were not visible, but which were likely to be occluded, to contribute to the recognition process. The system was also shown to distinguish between occluded and mosaic objects.

7.3.4 Guidance by shape

By creating early surface representations which were robust to occlusion, it was possible to allow a surface to complete for selection as a single entity. This was more compatible with primate visual processing than a more usual saliency map model in which competition took place between many discrete locations in the visual field.

System III extended the border ownership system to create stable representations of object shape, despite the presence of occluders. Shape was described with a list of the relative locations of the most salient in-object features. Selection could be successfully biased towards a particular shape in a manner which was robust to occlusion using this method.

7.4 Future Work

At present, the systems presented in this thesis have been restricted to operating on simple stimuli, such as digits and geometrical figures. It is highly desirable to expand these systems such that they can perform on natural images, and to create a more generic reward harvesting system, which can learn to organise large quantities of multimodal information. Although the systems in this thesis have been heavily inspired by biology, they are not sufficiently detailed to allow predictions to be made about behaving systems, though this also represents a highly desirable path for future development.

7.4.1 Natural Images

The system for surface-based competition and attention was only useable with line-diagram stimuli. It is desirable to expand the model to be able to deal with real images. While T-junctions have been established as a visual cue which provides information about
the order of objects in the visual scene, there are other cues available to assist this process. A number of approaches are possible:

**Continuing the T-junction approach:** this could be accomplished by providing a greater range of T-junction detectors on different scales which are able to reliably extract T-junctions from real images. This could be used in conjunction with an advanced segmentation algorithm, such as the EDISON system (Christoudias et al., 2002).

**Binocular disparity:** this would provide another cue through which border ownership could be determined (Nakayama et al., 1989). A similar system which used binocular input could be used to help segment the visual scene. The largest disparity in position should exist for the closest surfaces, which will own all their borders, and so on.

**Segmentation from motion:** this could also provide cues about the order of different surfaces in the visual scene. Such a system would involve a moving agent, which could be embedded in a virtual world. This would also extend the potential behaviours of the system, such as choosing where to move next etc.. Segmentation could be aided by considering portions of the visual scene with undergo similar transformations while the agent moves.

**Colour and texture:** though both of these properties provide very strong segmentation cues, they do not provide information about border ownership. While the experiments presented in this thesis operate on strict geometrical figures, a more general recognition system would require a strong segmentation process.

### 7.4.2 Completion Mechanisms

While border ownership information was used to improve performance during an object recognition task, we did not investigate the possibility of also including a completion mechanism. Such a system would allow the shape of an occluded feature to be inferred using Gestalt-like principles. For example, in Experiment I of Rensink & Enns’ 1998 paper, the suggestion is that the partially occluded square is actually treated as a square
by the visual system, and so it does not pop out as being any different from its unoccluded neighbours. The use of a more complete perceptual organisation system has been addressed by Grossberg (Grossberg, 1994). This system includes an illusory contour completion system, and method of segmentation through binocular disparity. A method of creating robust borders is essential for any system which wishes to use surface-based percepts, since it is from these that the limits of the surface can be identified.

7.4.3 Category Learning

Because the system is capable of object-based attention, it is possible to use the attended entity to train object and category recognition systems, in a similar manner to Fazl et al. (2009). The purpose of such a system would be to generate invariant object representations which can be used to drive more complex behaviours than the generation of eye movements, such as object categorisation. Information about the locations of potential occluders could also be interpreted as carrying information about potential obstacles during reaching tasks.

7.4.4 Representations of Complex Sensory Input

System Ia operated upon 3D sensory input, while System Ib operated on 7D input. An interesting expansion of this system would be to use it on many more channels of sensory input, e.g. position, colour, velocity, etc.. By representing the relevance of each channel with a weight, it would be possible to allow the SOM to represent only the relevant channels. Such a system would be capable of solving more complex tasks, in which the target mode changes, e.g. Wisconsin card sorting (Milner, 1963).

The perceptual organisation methods from System II and III has not yet been integrated with a continuous feature-learning scheme, such as the SOM provided in System Ia. The representation of more complex percepts on the SOM may be possible through an appropriate hierarchy of feature detectors.
7.4.5 Biological Detail

At present, the system is implemented at a high level of neural abstraction, with an emphasis on the functional stages of processing of the visual signal, and not on achieving a neurally plausible implementation of the system. Another desirable course of future research is to implement the system as a rate-based or spiking neural network. By developing the framework into a predictive model for visual behaviour, it would be possible to quantitatively test it against psychophysical data. Qualitative results for early robustness to occlusion were obtained in System III, which were comparable in this respect with the results of Rensink & Enns (1998) (see Fig. 2.5).

Implementing the model with a neural network would allow time-based predictions to be made, for example the time required to compare within-object features and between-object features (Behrmann et al., 1998). A neural implementation of part of this network has been published by Mihalas et al. (2011), in which the authors hypothesise a neural circuit through which attention can select a percept. Their model does not explore the interaction between attention and object recognition.

A second area opportunity exists for modelling neural responses in IT during the deployment of selective attention. By creating populations of prototype shape tuned features which can be modulated by attention, it may be possible to predict the time-course of gist-based judgement and object recognition.

7.4.6 Machine Vision

In order to develop the systems in this thesis towards useful machine vision applications it is important to demonstrate that a pre-attentive search stage brings significant benefits, not only in terms of search performance, but also in terms of efficiency. For this, it would be necessary to demonstrate that the system is significantly faster in locating target objects, or at least that the cost in terms of increased computation is less than the cost of making selection errors. At present, this is particularly a problem for Systems II and III, which both take several seconds per image.

A second way in which the systems could be used in machine vision is for object interaction and learning: one advantage which Systems II and III have over more usual
recognition algorithms is that they allow the location of the object to be clearly understood, while providing information about the structure of the scene, and any occluders. This representation could be useful for interaction, in the sense that there is additional information present to allow the system to plan reaching movements etc., or for learning, in the sense that it will be possible to extract visual features that purely belong to the target object.

7.5 Conclusion

The systems presented in this thesis emphasise a close relationship between scene understanding, visual attention and object recognition, which is likely to emerge in the early stages of visual processing.

Machine vision algorithms have frequently taken inspiration from the perceived function of the mammalian visual system, though such systems are often based upon models of classical receptive fields. This thesis proposes that the integration of non-classical receptive field models, perceptual organisation mechanisms and attention-like processes may also be beneficial for the generation of intelligent machine behaviour.
A.1 System I

This section contains extra graphs from System I.

A.1.1 Experiment I

When there was no negative activity in the working memory $B = 0$, then there was also no bimodal distribution in the accumulated reward when static SOM node weights were used, Fig. A.1. This was because selection was not biased away from any region, but only towards them. If a location occasionally provided a reward, but was generally unrewarding, it could be expected to be strongly inhibited when $B = -1$, but would generally have a value of zero or slightly more when $(B = 0)$. 
A.1. SYSTEM I

A.1.2 Experiment II

Extra graphs from Experiment II of System I, which show the effect of the frequency in the change of the rewarding target \(F\) on the reward harvesting rate. Figs A.2 to A.6 plot the harvesting rate as a function of the decay rate for working memory activity \(\gamma\). The location of the rewarding cues was varied throughout the experiment \(F = \{0, 1, 2, 3, 4\}\). We also compare the online learning of the SOM weights, with offline learning of the SOM weights.

Figure A.1: Graph showing the effect of the resolution of the colour space and the size of the reward window on harvesting for different values of \(N\) when \(B = 0\), no online learning of SOM weights
Figure A.2: Reward harvesting rate for $F = 0$
Figure A.3: Reward harvesting rate for $F = 1$
Figure A.4: Reward harvesting rate for $F = 2$
Figure A.5: Reward harvesting rate for $F = 3$
A.1. SYSTEM I

Figure A.6: Reward harvesting rate for $F = 4$

The reward harvesting rate slowly decreased as $F$ was increased. Fig. A.7 shows the cumulative number of rewards harvested over time for the most successful parameter sets for $F = 0 \ldots 4$. The slopes of the traces were similar in regions of the graph that did not follow changes in the target (with the exception of the case $F = 4$), which indicated that the decrease in the total number of harvested rewards occurred as a result of the time required for the system to relearn the new target, but that it recovered to the same harvesting rate afterwards. The case $F = 4$ provided a very severe test of the system, as this did not allow enough time for the SOM to reset and converge upon a stable new representation.
Figs. A.8 to A.12 show the reward harvesting rate as a function of the learning rate for working memory activity, $A$. Except in the case where the injected activity was not diffuse (the most pale trace), the harvesting rate varied little for values of $A \geq 0.25$. This indicated that the most significant effect on the harvesting rate was caused by $\sigma_L$, the diffuseness of learning on the SOM, followed by the decay rate of working memory activity $\gamma_L$ and finally, the learning rate for working memory activity $A$. The most generally successful set of parameters involved diffuse learning $\sigma = 1$, a small, but nonzero decay rate $\gamma = 0.95$ and a small, but nonzero learning rate $A = 0.25$. 

Figure A.7: The cumulative number of rewards received over time
Figure A.8: Reward harvesting rate for no changes in target, $A = B$; (a) $\gamma = 1$; (b) $\gamma = 0.95$; (c) $\gamma = 0.9$; (d) $\gamma = 0.85$; (e) $\gamma = 0.8$;
Figure A.9: Reward harvesting rate for 1 change in target, $A = B$; (a) $\gamma = 1$; (b) $\gamma = 0.95$; (c) $\gamma = 0.9$; (d) $\gamma = 0.85$; (e) $\gamma = 0.8$;
Figure A.10: Reward harvesting rate for 2 changes in target, $A = B$; (a) $\gamma = 1$; (b) $\gamma = 0.95$; (c) $\gamma = 0.9$; (d) $\gamma = 0.85$; (e) $\gamma = 0.8$;
Figure A.11: Reward harvesting rate for 3 changes in target, $A = B$; (a) $\gamma = 1$; (b) $\gamma = 0.95$; (c) $\gamma = 0.9$; (d) $\gamma = 0.85$; (e) $\gamma = 0.8$;
Figure A.12: Reward harvesting rate for 4 changes in target, $A = B$; (a) $\gamma = 1$; (b) $\gamma = 0.95$; (c) $\gamma = 0.9$; (d) $\gamma = 0.85$; (e) $\gamma = 0.8$;


