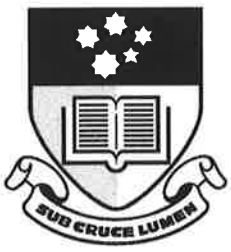
A Neural Framework for Visual Scene Analysis with Selective Attention

by

Eric Wai-Shing Chong, B.E.(Hons)

A thesis submitted in fulfilment of the requirement for the degree of

Doctor of Philosophy



The University of Adelaide

Faculty of Engineering, Computer and Mathematical Sciences

Department of Electrical and Electronic Engineering

June, 2001

Contents

Abstract

Attention is essential to the analysis of visual scenes that consist of multiple objects, especially in cases where the objects are embedded in complex and cluttered backgrounds. Despite its importance, few artificial neural network models of object and pattern recognition have incorporated attentional mechanisms. Recent advances in cognitive neuroscience have provided important information on the neural mechanisms of attention. Significantly, attentional processes involve the modulation of neuronal signals, and are clearly influenced by memory related processes via feedforward-feedback interactions.

This thesis proposes an architectural framework based on neural networks for visual scene analysis with attentional mechanisms. The core of the framework is based on an adaptive resonance theory architecture, which is a self-organising neural network for stable learning of recognition codes. The proposed model exploits the computational role of attention in visual object recognition by modelling the dynamics of attentional processes for perceptual grouping and selective processing. As a result, the proposed model is capable of performing translation, rotation, and distortion invariant 2D object recognition in the presence of background clutter and occlusion. The model is shown to be flexible to extensions by incorporating an elementary motion detection architecture for recognising moving objects. Furthermore, the use of feedforward-feedback modulation has enabled partial or incomplete familiar objects to be recognised in a variety of visual conditions. Biologically, such feedforward-feedback interactions can be used to explain the phenomenon of visual completion.

Simulation studies undertaken demonstrate the effectiveness of the proposed model in recognising 2D objects in many non-ideal visual conditions. The practical feasibility of the neural architecture is demonstrated through its application to real-world images. Despite difficult visual environments, including severe distortion, the simulation results indicate the model can detect, locate and recognise the learned objects from the simulated images.

From the research presented in this thesis, it is concluded that the use of attentional mechanisms can enhance artificial vision systems to cope with difficult visual conditions. It is shown that feedforward-feedback interactions with synaptic modulation are a versatile and powerful mech-

anism for performing many useful functions such as transformations, filtering, gain control, and selective processing in neural network based vision systems.

Declaration

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.

Signed: Date: 5/6/2001

Acknowledgments

I am greatly indebted to my supervisor, Dr. Cheng-Chew Lim, for his tireless guidance, encouragement and assistance throughout the course of this research. I would like to thank him for his availability, and his thorough and detailed reviews of my work.

I would like to express my heartfelt gratitude to Dr. Peter Lozo, of Defence Science Technology Organisation (DSTO), who has provided me with many insightful suggestions and ideas that have formed the backbone of this thesis.

Many thanks must go to the staff of the Department of Electrical and Electronic Engineering at the University of Adelaide for providing a stimulating and benign work environment. In particular, my appreciation goes to Prof. R. Bogner for kindly giving me his neural network notes on Adaptive Resonance Theory, Dr. A. Moini for providing the LaTeXthesis templates, and Kiet To for helping me with various computing queries. Thanks to my colleagues Hong Gunn Chew, Alex Lin, Jee Gah Lim, Qiang Fu, and Peter Celinski, for their constant encouragement and friendship.

Special thanks must go to my friends Thanh Chung and Carmine Pontecorvo for proofreading my thesis, and their contributions are gratefully acknowledged.

Finally, I would like to thank my parents for all of their support and guidance; their support has made this thesis possible.

This work was supported by an Australian Postgraduate Award with stipend from the Australian Government, which I gratefully acknowledge.

List of Publications

The following is a list of publications which were published or submitted during the PhD candidature by the author.

1. **E. W. Chong**, C. C. Lim, and P. Lozo. Modelling of a neural motion detection filter for attentional modulation. In *International Workshop on Image Analysis and Information Fusion*, pages 311–321, Adelaide, 1997.

2. **E. W. Chong**, C. C. Lim, N. Atsikbasis, and P. Lozo. Design of a 2-D neural motion detection filter. In *IEEE Region 10 Annual Conference*, pages 667–670, Brisbane, 1997.

3. **E. W. Chong**, C. C. Lim, and P. Lozo. Neural model of visual selective attention for automatic translation invariant object recognition in cluttered images. In *Proceedings of the Third International Conference on Knowledge-Based Intelligent Information Engineering Systems, KES'99*, pages 373–376. IEEE Press, 1999.

4. **E. W. Chong** and C. C. Lim. Elementary motion detection with selective attention. In *Proceedings of the Third International Conference on Knowledge-Based Intelligent Information Engineering Systems, KES'99*, pages 365–368. IEEE Press, 1999.

5. **E. W. Chong**, C. C. Lim. Neural model for distorted and cluttered scene analysis. *Submitted* to *Image and Vision Computing*, 2000.

6. **E. W. Chong**, C. C. Lim. A self-organising neural architecture for parts recognition in occlusion. *Submitted* to *Neural Computation*, 2000.

List of Principal Symbols

A	passive decay rate of postsynaptic cellular activity
A^T	transpose of matrix A
$\bar{A}$	passive decay rate of inhibitory interneurons
B	upper saturation level for postsynaptic cellular activity
$\bar{B}$	charging rate of inhibitory interneurons
C	lower saturation level for postsynaptic cellular activity
C_{ij}	central representation element (i, j)
D	passive decay of postsynaptic potential
$\vec{D}$	preferred direction
F_i	facilitatory signal
G	postsynaptic gain
$\bar{G}$	lateral feedback inhibition gain
$\bar{G}_{ij}$	inhibitory Gaussian receptive field
H_{ij}	high activity field activity
J_i	excitatory synaptic input
K_u	gain of the postsynaptic feedback induced depletion of the stored transmitter
$M_{ij,r}$	memory field element (i, j) of the rth stored model
$\mathcal{M}$	band transformed memory
S_i	presynaptic signal
$\mathbf{S}$	shape attracted pattern
T_{ij}	top-down field element (i, j)
W_{ij}	attentional capture Gaussian receptive field
Y	threshold for transmitter release
$\|.\|$	Euclidean norm or L_2 norm
$f : \mathbb{R}^n \rightarrow \mathbb{R}^m$	function f maps $x \in \mathbb{R}^n$ onto $f(x) \in \mathbb{R}^m$
$f(.)$	threshold function
$g(.)$	threshold function
$\vec{b}$	directional bias

n	number of neurons in a PMSCNL
$\mathbf{r}$	reset vector
u_i	stored transmitter
v_i	excitatory postsynaptic potential
v^+	excitatory synaptic channel
v^-	inhibitory synaptic channel
$\bar{v}_i$	lateral feedback inhibition
w_{ij}	weight for the connection between nodes i and j
Δw	change in weight
x_i	postsynaptic cellular activity
$x_{\vec{D}}$	direction-selective cellular activity
y_i	mobilized transmitter
z_i	long-term-memory, transmitter production rate
α_u	passive storage of the produced transmitter
β_u	passive decay and mobilisation of the stored transmitter
β_y	passive transmitter mobilisation
ψ	a threshold level
ϕ	rotational template orientation
$\varphi(.)$	threshold function
γ_y	passive decay of the mobilised transmitter
θ	a threshold level
ρ	vigilance parameter
ς	secondary vigilance parameter
σ	standard deviation of Gaussian receptive field
ϱ	standard deviation of inhibitory Gaussian receptive field

List of Abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
ANN	Artificial Neural Network
ART	Adaptive Resonance Theory
BT	Band Transformation
BU	Bottom-Up
CR	Central Representation
CSAART	Complementary Selective Attention Adaptive Resonance Theory
ERP	Event-Related Potential
LGN	Lateral Geniculate Nucleus
LTM	Long-Term-Memory
NN	Neural Network
PMSCNL	Presynaptically Modulated Shunting Competitive Neural Layer
SA	Selective Attention
SAART	Selective Attention Adaptive Resonance Theory
STM	Short-Term-Memory
TD	Top-Down
WTA	Winner-Take-All

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

Visual attention is essential for analysing visual scenes consisting of many objects, especially for scenes with complex and cluttered backgrounds. Attention is required to determine where one should focus and what input stimulus features should be selected to form meaningful objects. Without attention, one would not be able to detect changes in one's immediate surroundings. Furthermore, failures in the attentional system may result in seeing meaningless visual patterns of elementary features instead of familiar objects.

Despite its importance in visual perception, the incorporation of attentional phenomena and mechanisms in vision models is rare. This can be attributed to three main factors. The first is our insufficient understanding of the underlying neural mechanisms that are responsible for attentional processes. The second is that attentional phenomena are diverse, serving a great number of computational purposes, thus difficult to unify by a single theory. Lastly, it is generally agreed that there is some intelligent force or agent that is controlling the attentional system; such a higher-level decision making unit cannot be defined or modelled easily [94].

Recent advances in cognitive neuroscience and neurophysiology have unearthed some of the mysteries surrounding selective attention. Findings in these fields have provided information on the timing and sequential order in specific anatomical locations that are affected by attentional processes; the modulation of the responsiveness of neurons that encode colour, form, and spatial information during attention [50, 54, 198]. Moreover, the use of event-related brain potential has allowed us to gain insight into the levels of processing at which different kinds of visual information are selected for further analysis [120]. These findings have enabled us to postulate theories and develop models and algorithms for attentional functions. In particular, studies

in electrophysiology have provided knowledge on the implementation of attentional processes using neuronal circuits and neural systems.

Of special interest to modelling the computational properties and dynamics of attentional mechanisms is the feedforward-feedback interactions between bottom-up processes (stimulus driven) and top-down processes (memory driven). It is known that attention may regulate access to memory based on the fact that spatial attention limits the amount of information processed by suppressing unattended stimuli, thereby denying them access to memory [131]. The reverse may also occur, in that memory guides attention [51]. In a visual search, top-down signals have been found to bias cells that are related to the search, and activation of these cells is enhanced when the right stimulus occurs. Often such a visual search is achieved by memory-guided attention, in which the representation of a target in memory is used to guide the search of a visual scene.

1.2 Research Objectives

The primary goal of this thesis is to acknowledge the importance of attention in visual perception and to incorporate attentional processes into a model for visual scene analysis. The objectives of this study are therefore:

- to *explore* the role of attention in visual perception, in particular, its relation to object recognition and visual scene analysis;
- to *review* the current status of artificial vision systems in the context of object recognition;
- to *model* the computational properties and dynamics of attentional processes, namely parallel-preattentive and serial-attentive processes, using feedforward-feedback interactions and biological principles and mechanisms;
- to *develop* a neural architecture that utilises attentional mechanisms for analysing visual scenes under a variety of operating conditions;
- to *establish* a general framework that allows additional visual functions to be incorporated with ease; and
- to *investigate* the effectiveness of the proposed neural architecture through simulation studies on synthetic and real-world complex and cluttered image scenes.

In visual scene analysis, we expect the proposed system to be able to detect, locate, and recognise from a visual scene any familiar objects that may be shifted in position and orientation,

distorted in shape, and with minor changes in size, as well as in the presence of occlusion and background clutter.

Figure 1.1 is intended to provide readers an overview of the scope covered in this thesis. It shows the processes required for an input scene to be analysed. Basically visual scene analysis as proposed in this thesis is a result of three major processes involving selective attention, object recognition and memory. Figure 1.1 also shows that these three processes are built on a neural architecture that learns and stores object representations from an input image.

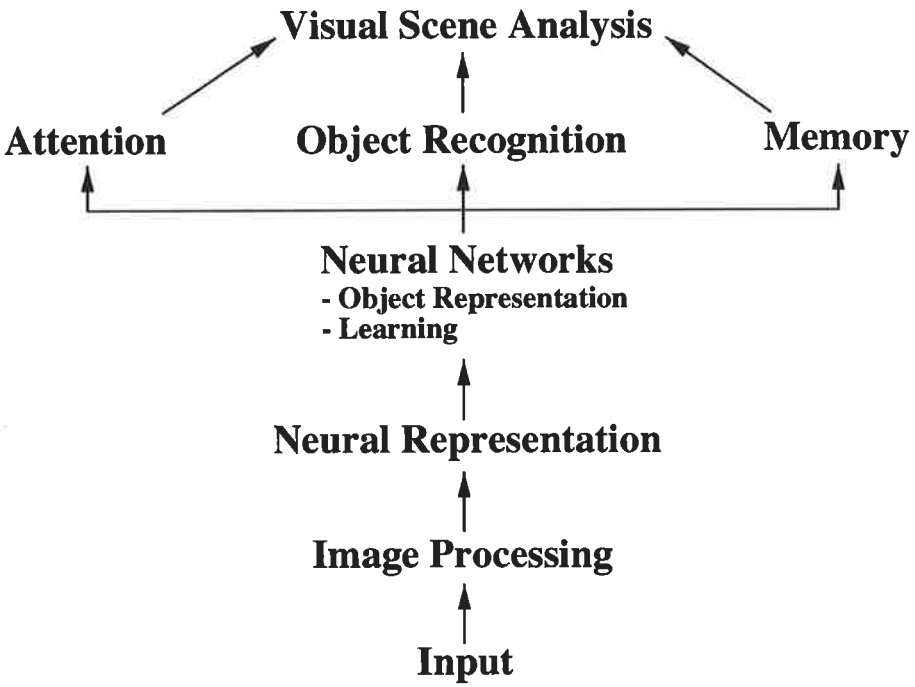


Figure 1.1: Scope of the thesis.

1.3 Research Methodology and Approach

The field of vision is vast and diverse. Scientists from disciplines such as psychology, neurophysiology, and cognitive neuroscience are interested in discovering scientific facts about various aspects of the visual system. On the other hand, engineers, computer scientists and mathematicians are more concerned with formulating theories and developing models for visual functions. In order to unifies knowledge from the various disciplines of vision, a suitable research approach must be devised.

We formulate our approach based on the so-called top-down modelling proposed by Marr [122]. In top-down modelling, there are three levels of information processing:

1. **Computational theory** - specifies a function to be computed, which matches the input-output nature of neuropsychology. Examples are higher-level visual functions such as visual search, object recognition, and attentional shift and capture.

2. **Algorithm and representation** - concerns the computational steps that are required in order to transform an input representation to an output representation of that function. Neurophysiology has provided us with insights into neural representations at various stages of the visual pathway. Moreover, direct measurements of neural activity during visual tasks in electrical and magnetic recordings have allowed computational relationships among various neural substructures to be established. Top-down attentional modulation is an example of an algorithmic step.

3. **Implementation** - determines the underlying mechanisms and structure that are responsible for carrying out the computational steps. Neural network models are well suited for implementation.

From these three processing levels, we derive a three-level approach to modelling our proposed neural architecture, which we simply refer to as *psychological, neurophysiological*, and *implementation* levels. To begin modelling, we first attempt to develop a framework for our proposed model from psychological theories and models using a "blackbox" strategy, where the details are hidden and the emphasis is entirely on the input-output relationship of the system concerned. Models and concepts of neurophysiological studies are used to devise neural representations, as well as establishing computational relationships and connections among various neural substructures within the framework. Finally, the framework is implemented using fundamental building blocks such as chemical synapses, synaptic connections, and neural layers.

1.3.1 Shape-Based Representation

Having established the modelling approach for our proposed visual scene analysis system, we need to consider the sources of information and its representation in the system. The central issue here is how to represent an object. The retinal projections of real-world objects are rich in information, containing many different object features and properties such as colour, shape, texture, depth, shading, surface curvature and reflectance, and motion. We may recognise an object visually based on one or any combinations of these features. Nonetheless, one can generally recognise an object according to its characteristic shape alone [191].

In this thesis, we will be concerned primarily with shape-based recognition in our proposed system, because for the recognition of many objects, non-geometric features play only a secondary role [166]. Furthermore, the shape of an object tends to dominate over other visual cues [68].

For example, a green dog, despite its unusual colour, can readily be recognised. The use of shape information for object and pattern representation has been pursued by many authors, see [110] for a review.

1.3.2 Self-Organising Neural Architectures

Besides object representation, we need a suitable choice of neural paradigm to form a basis for the proposed system. Biological systems are known to be self-organising in nature, where useful information is extracted from the input data without the need for a "teacher". An important family of neural architectures that are capable of preserving existing knowledge at the same time maintaining its plasticity are *Adaptive Resonance Theory* (ART) neural networks by Carpenter and Grossberg [28, 29, 30]. ART is a self-organising network that embraces the cognitive concepts of attention, vigilance, top-down priming and bidirectional learning. However, ART alone is insufficient to address the problem of real-time learning and recognition in complex and cluttered environment as pointed out by Lozo [113]. It led to the proposition of a neural architecture called *Selective Attention Adaptive Resonance Theory* (SAART) [113], which implemented the concept of top-down selective attention via synaptic modulation. In this thesis both ART and SAART will be used to model the proposed visual analysis system. Detailed discussions on ART and SAART for pattern and object recognition are provided in Chapter 3.

1.4 Major Contributions of the Thesis

The principal contributions made in this thesis are as follows:

- A review of the development of neuropsychological and neurophysiological aspects of vision, in particular the role of attention in visual perception and its relation to object recognition.
- A critical review of the computational and neural approaches to the high-level visual function of object recognition.
- An in-depth study of Adaptive Resonance Theory and Selective Attention Adaptive Resonance Theory for pattern recognition.
- The modelling of the computational properties and dynamics of attentional processes for high-level visual functions.

- The development of a biologically inspired neural framework for invariant visual scene analysis with selective attention.
- The investigation of the performance of the proposed neural system under a variety of visual conditions in the presence of occlusion, noise, and background clutter.
- The development of a neural architecture for elementary motion detection with attentional modulation mechanisms.
- The integration of the motion detection module allowing the framework to detect, locate, and recognise both static and moving objects.
- The development of a self-organising neural architecture for parts recognition using complementary feedforward-feedback modulatory pathways.

1.5 Outline of the Thesis

In Chapter 2, a brief introduction of vision is provided. We define visual perception and examine two of its critical components, namely visual selective attention and object recognition. We explore the psychological and neurophysiological basis of visual perception, and consider the neural mechanisms that constitute the visual system, in particular those responsible for visual attention. The chapter summarises cognitive data on attentive vision, serving as background information for modelling in later chapters.

Chapter 3 provides the mathematical and theoretical foundations for our research. The core of this chapter is a literature review on the various approaches to object recognition. We highlight the use of artificial neural networks for vision systems, and discuss some of the related issues such as network architectures and learning paradigms. A detailed study of Adaptive Resonance Theory and its extension Selective Attention Adaptive Resonance Theory is presented.

Chapter 4 presents a number of neural architectures for visual object recognition under a variety of visual conditions. The visual functions considered in this chapter include translation invariance, rotation invariance, distortion invariance, recognition in the presence of occlusion and background clutter, and attentional shift and capture. We show how ART and SAART are embedded and integrated into the overall architecture. Specifically, the concept of top-down presynaptic facilitation is implemented to model the effects of memory-guided selective attention for the recognition of familiar objects in cluttered background. This chapter forms the main body of the thesis. It is an attempt to combine cognitive data, theories and models together into a neural framework for vision scene analysis.

In Chapter 5, we present a computer simulation study for all the visual functions modelled in Chapter 4. The simulations contain both synthetic and real-world imagery scenes. This study demonstrates the practicability and effectiveness of the proposed framework in real-world applications. Discussions on the design of system parameters and limitations are provided.

Chapter 6 provides a demonstration of the extendibility of the framework by including elementary motion as a bottom-up visual cue for capturing spatial attention. The chapter begins by proposing a neural architecture for elementary motion detection, which is then incorporated into the framework. An implementation of directional bias is presented. The chapter also contains an analysis of the motion detection architecture under various operating conditions.

Chapter 7 presents some advanced features of the framework. An extension of ART and SAART is proposed using feedforward-feedback modulatory pathways for bottom-up gain control of top-down signals. The resultant neural architecture is capable of recognising incomplete and occluded objects in cluttered images. Ways in which the robustness of the automatic attention stage may be improved are discussed. A proposal for size invariance is also presented.

Chapter 8 provides the overall conclusions of the thesis and discusses possible avenues of future research.

Chapter 2

Vision

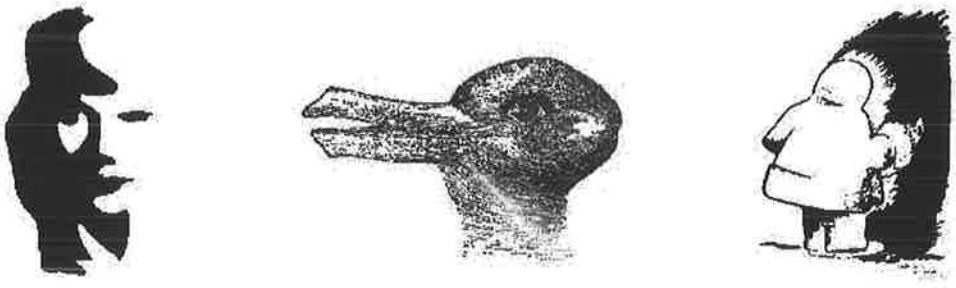
This chapter provides a brief overview of visual perception in terms of object recognition and visual attention. In addition, we examine the mammalian visual pathway that underlies these visual functions. The reader should note that this chapter serves to illustrate some of the findings that are relevant to the modelling of our proposed visual scene analysis system. As such, a complete review of the visual system is beyond the scope of this thesis.

2.1 Visual Perception

Visual perception is a process of reaching an understanding and awareness about objects and events in our immediate surroundings visually. It involves the use of knowledge and memory, and is guided by our attentional system [23, 68, 146, 164]. One must be able to recognise all important objects in the visual scene before such an understanding can be reached, therefore visual scene analysis can be regarded as the first stage to visual perception.

Two essential components of visual scene analysis are object recognition and visual attention. While there are numerous models for object recognition, models equipped with selective attention are less common. To illustrate the importance of attention, several ambiguous figures are depicted in Figure 2.1. In each case, if one was told the figure's identity, then a top-down expectation would be generated to match that description. For example, if told Figure 2.1(b) was a duck, then most people would see it a duck. On the other hand, most would see a rabbit if so suggested. This simply shows the strong influence top-down attention has on object recognition.

In the next section, we review the underlying nervous system that is responsible for visual information processing. This is followed by discussions on higher perceptual functions of object



(a) (b) (c)

Figure 2.1: Ambiguous figures. Can be seen as either (a) a man playing horn or a woman silhouette; (b) a rabbit or a duck; and (c) a face or an Eskimo.

recognition and visual attention.

2.2 The Visual System

Vision begins with our eyes as sensors for light, but in terms of neural activity the visual system begins processing at the retina, located at the back of the eye. The primary function of the retina is to convert the image formed by the eye into neural activity that can be understood and processed by our nervous system. The retina is formed by millions of light-sensitive neurons. Each neuron receives stimuli from only a fraction of the visual scene, and produces a response with magnitude according to the light intensity of that fraction. In this way, the visual scene is transformed by a photo-sensitive neural layer, and represented as discrete neural responses. These responses are further processed by other neural layers (ganglion cells) in the retina, each having its own distinct receptive field properties, to produce contrast signals of the input stimulus before proceeding to the lateral geniculate nucleus (LGN). It is believed that the LGN is where the segregation of visual streams begins, as the LGN consists of six layers and it projects to several other areas of the cortex. However the great part of that projection arrives at the primary visual cortex (area V1) which is the starting point of specialization because the receptive fields of neurons in this area are qualitatively different from those that produce contrast responses in the early stages. In particular the receptive field properties in area V1 define two categories of cortical neurons, called simple cells and complex cells [87]; the former have oriented receptive fields, and hence they respond to stimuli in some orientations better than others; the latter's receptive fields are direction-selective, that is the cells respond only to stimuli moving in the preferred direction of the receptive fields.

A simple illustration of the processing in the visual system is shown in Figure 2.2.

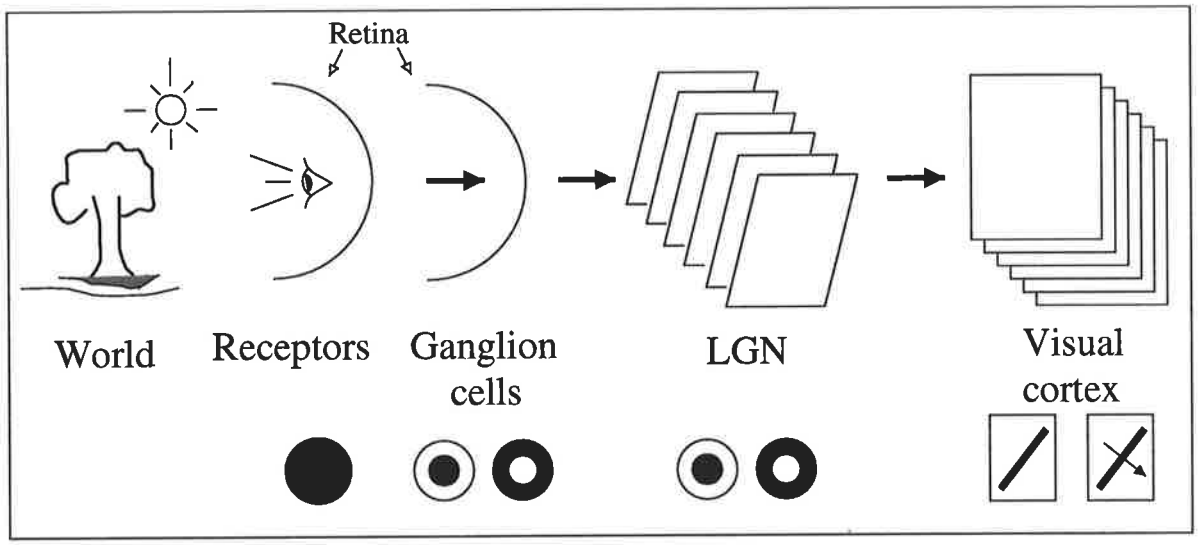


Figure 2.2: Outline of the visual system.

2.2.1 The Neuron

Neurons are the basic processing units of the brain. Each neuron can only perform an elementary computation. It receives information from other neurons that are connected to it, makes a decision based on that information, and transmits the decision to some other neurons. A typical neuron has several major components, consisting of a cell body or *soma*, a *nucleus*, two types of tree-like structures: the *axon* and the *dendrites*, and the *synapses* at the axon terminals. The dendrites of a neuron are its receivers. They receive signals from the synapses of nearby neurons. These signals allow the cell body and its nucleus to generate a decision signal which is passed along the axon for transmission to other neurons. The axon eventually branches into strands and sub-strands called the axon terminals which are the transmitters of the neuron. At the end of the terminals are the synapses, they act as connectors between the two tree-like structures. The axon and its branching terminals are said to be *presynaptic* because they are located before the synapses with respect to information flow, whereas the dendrites are *postsynaptic*. A diagram of a biological neuron is given in Figure 2.3.

The cerebral cortex, the outer shell of the brain, contains approximately 10^{11} neurons, and each neuron is connected to 10^3 and 10^4 other neurons, giving about 10^{14} to 10^{15} connections. These massively parallel networks provides enormous processing power for the brain.

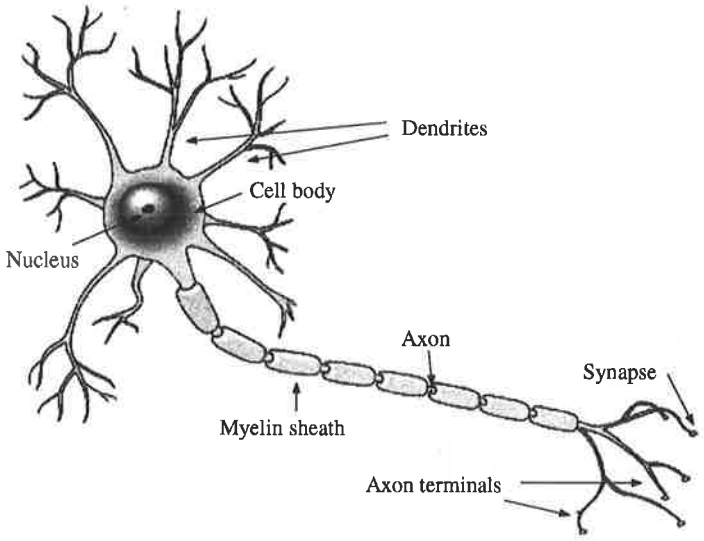


Figure 2.3: A typical neuron.

2.3 Object Recognition

Studies in *visual agnosia* have provided insights into human object recognition. Gazzaniga *et al.* [68] gave a detailed account of a patient who had lost the ability to recognise everyday objects, yet he was still able to perform all the fundamental visual functions such as identifying colour and shape. It seems that he had difficulties in linking features together to form a coherent object. This example illustrates that seeing is meaningless unless object recognition can be achieved.

Object recognition is about retrieving identity information for a set of conjoint features from our memory. The set of conjoint features to be identified depends on the circumstances. Object recognition is often viewed as a computational problem due to Marr's influential theory of vision [122]. From computational point of view, object recognition is a process in which images are compared to stored representations of objects for the purposes of identifying the objects that give rise to the images.

Despite continual efforts, object recognition remains a challenging problem. Much of the difficulty associated with object recognition arises from the fact that objects captured in images do not always provide sufficient resemblance to their stored representations. One obvious reason is real-world objects are 3D in nature, thus an object may give rise to a large number of different 2D projections. Furthermore, real-world objects are often embedded in a scene that contains many other objects. Those objects that are located close to, or occluding, the target object can cause object recognition to fail. Ullman [191] suggests there are two more sources of variability in object appearance, besides the two mentioned above. One is the result of varying illumina-

tion conditions, which can drastically alter the performance of a recognition system. The other is caused by changing shape in non-rigid objects, or objects with movable rigid sub-parts, such as a pair of scissors.

Variability in sensory information is the fundamental reason why object recognition is difficult. Therefore object recognition problems are usually formulated as *object constancy* problems - the ability to recognise an object in countless situations. Central to object constancy is the debate whether recognition occurs in an object-centred or viewer-centred frame of reference.

2.3.1 Object-Centred versus Viewer-Centred Representations

Object-centred representation describes objects using view independent properties and has a frame of reference (coordinate system) that is inherent to the object concerned. Recognition with object-centred representation is usually achieved through invariant transformations [37, 147] such that a single stored model is used to recognise the object from all possible views. In viewer-centred representation objects are described according to their appearance as viewed [13, 86, 162, 165], and has a frame of reference defined in terms of the direction and distance from the observer's viewpoint to the object. A viewer-centred model is typically a set of views of an object covering a restricted range of views of the object. Therefore, a number of models are required to represent the object from all possible views.

An obvious weakness of the viewer-centred approach is the need for an extremely large memory of stored models. In contrast only one model is required for each object in object-centred representation. Nonetheless, viewer-centred representations are easier to acquire and manipulate than object-centred ones. Furthermore, some studies in physiology and psychophysics have provided experimental results that suggest viewer-centred representation may be employed by the visual system [109, 148, 183]. Computational models and approaches of object recognition are further discussed in Chapter 3.

2.4 Selective Attention

Attention is usually regarded as a complex mental process that performs selective processing. In many cases, authors avoid defining attention directly, rather it is explained or illustrated in terms of phenomena that are associated with attention such as the "cocktail-party" effect. The reluctance to define attention can be attributed to our lack of understanding of the underlying neural mechanisms and processing. Unlike object recognition which can be defined as a computational process, attention is bordering on consciousness and intelligence, thus difficult to

define adequately. Review papers on attention can be found in [2, 3, 50, 54, 94].

Attention is most commonly conceptualised as the consequences of limited or insufficient processing resources or processing capacity of the brain [22]. There are two aspects to this conception: the first is the limited processing resources which means attention must act as a resource allocator in terms of mental effort and time; the second is the limited processing capacity which requires attention to serve as a scheduler, thus its selectivity processing nature.

So how does attention relate to object recognition, and what role does it play in visual scene analysis? Given that limited capacity is the basis of attentional processes, our visual system must utilise the system resources in the most effective and efficient manner. Therefore, visual attention can be described as the "processes that enable an observer to recruit resources for processing selected aspects of the retinal image more fully than nonselected aspects" [146].

For survival purposes, the visual tasks of detection, location, searching, learning and recognition must be performed effectively in the shortest possible time, despite system limitations. To achieve this, our visual system must choose to process information that is most relevant to our needs.

Accordingly, attention first handles the priority of processes. It requires detection and location of regions of interest based on elementary feature strengths. Our everyday life can attest this, for example, if an unknown object was flying towards us, understandably, most of us would try to move out of its path. For this to happen, the moving object must first attract our attention - this amounts to target detection. We then need to know where the object is coming from and its flight path, i.e., its location. From which we can determine if we are in danger of being hit, thus reach an appropriate decision to counter the situation, and finally act upon it. Notice how during the entire process, it was not necessary to identify what the actual object was, because situations of this nature require urgent attention. Recognition of the object may come afterwards. This example shows that attention tells us which objects or events we need to understand and be aware of first and how much system resources should be allocated for processing. While object recognition is required to identify all objects involved. Attention is also involved in higher level perceptual functions such as segmentation and grouping. It helps the selection of related features to form meaningful objects, and assists in analysing and interpreting stimulus events using memory and knowledge. Thus attention is essential to semantic encoding and analysis. Failures in the attentional system can lead to recognition failures.

In the above, we have described the "where" and "what" of visual perception. Not only must we recognise what we are looking at, but we also need to know where it is in order to respond appropriately. The existence of the what-where pathways in higher cortical layers is widely supported by experimental evidence in neurophysiology [90, 156].

Clearly, effectiveness and efficiency are two conflicting and competing requirements of attention. A balance between the level of detail processed and processing speed must be maintained. It appears that the attentional system achieves this by allocating system resources, in terms of mental effort and time, according to the situation. As the level of analysis increases, more interaction with memory is required, and the longer it takes to process. Detection, classification and recognition differ in their level of abstraction, which can be attributed to the amount of resources allocated in each case. It is interesting to note, as pointed out by Ullman [191], that classification is more demanding than recognition in artificial systems contrary to their biological counterparts.

2.4.1 Psychology of Attention

One of the first psychological models on selective attention is the "filter theory" by Broadbent [20], which proposes that the sensory system has a limited-capacity channel and all the information passing the channel is screened to let the most important portion through. This filtering mechanism has been described as a gate that can be selectively opened for attended information and closed for ignored information. This theory has since been modified to suggest that unattended information is degraded and attenuated rather than gated out completely [21].

Studies in psychophysics have suggested that visual perception is a two-stage process, namely parallel *preattentive* (or distributed attention) and serial *attentive* (or focussed attention). Experimental observations on time course of attention [198] have been able to show bimodal distributions of attention shifts. The first mode is a quick, effortless, automatic process which operates in parallel over the entire visual field. In contrast, the second mode operates serially over a limited portion of the input, and is a slower, effortful, controlled process. This has since been supported by experimental data in human electrophysiological and neuromagnetic recordings [118, 121].

Theories based on the two-stage model have suggested the preattentive mode locates regions of interest from the visual scene by processing simple features such as colour, direction of motion, orientation of edges, and luminance contrast in parallel rapidly. These regions are then processed by the spatially limited serial attentive stage [97, 186, 202, 203]. Experimental results on visual search tasks involving single feature and multi-feature targets have supported the theory [187]. It was found that response times for single feature searches were independent of the set size. In contrast, response times for conjunction searches increased with the set size.

Visual search is guided by the two modes of attention. In a search, there must be a target which may be partially or completely known. This information is used to guide the search of the visual

scene. Generally speaking, there are two types of search tasks [187]. The first is the so-called "pop-out" tasks, where the target stands out from its background on the basis of a strong featural cue. This may be achieved by having a target with a unique feature from distractors (e.g., yellow target among blue distractors). Reaction times for search tasks of this nature are not affected by increase in size set. In contrast, reaction times for the second type, *conjunction* search, increase as a function of set size, indicating the scene is searched region by region. Here the target is defined by a conjunction of features shared with distractors. Reaction times for these two search types suggest that pop-out searches are performed in parallel whereas conjuncture searches are performed by a serial processing stage, corresponding to the preattentive and attentive modes, respectively.

In our everyday life, we encounter complex visual scenes that require us to perform these search tasks repeatedly and frequently. Not only are the two search types not independent, in fact, they complement each other. Together they allow visual search to be performed in an efficient and effective manner - with one being fast but coarse, and the other sophisticated but slow. In searching for a particular object from a visual scene, the preattentive stage is first employed to single out one or a few potential regions that are most likely to contain the target object. This selection is based on the strength of some pop-out feature associated with the target (e.g., a red hat from a variety of items, then red regions would first be located). These regions are further analysed one by one by the attentive mode to verify the identity of each of these stimuli (e.g., checking one at a time if any of the red stimuli is a hat). Being able to perform one type of search without the other is useless to the visual system, as without pop-out search we would not know where to start the search; whereas lacking the ability to do conjuncture search means we lose the ability to make sense out of basic features, thus the ability to relate to known models in memory, i.e., failure in semantic encoding.

2.4.2 Neurophysiology of Attention

Psychophysical experimental results have provided us with behavioural information and input-output relationships of the attentional system. However, they fall short of revealing the underlying neural mechanisms that are involved in attentional processes. To gain insights into neural mechanisms of attention and find support for psychological theories, researchers have turned to physiological methods in humans and animals to monitor neural events during attention. We provide a brief account of the development and findings in neurophysiology of attention based on the following review papers [50, 120, 123].

Neurophysiological experiments are commonly conducted using electric and magnetic recordings of brain activity and neuro-imaging techniques. Early neurophysiological studies in spatial

attention by measuring brain waves that are directly related to stimulus processing, referred to as *event-related potentials* (ERPs), have been able to detect changes in ERPs. When a stimulus appears at an attended location, the corresponding ERP is enlarged in amplitude. This implies that spatial attention occurs in part via the modulation of sensory processes in the visual cortex. Further experimental results show that ERPs are enhanced in both automatic and voluntary attention, indicating that automatic and voluntary attention to locations involves a common mechanism with regard to the effect on cortical stimulus processing.

Luck *et al.* [118] show from recordings of ERPs that sensory processing is modulated in a spatially restricted manner during visual search. This provides support to the idea that focal spatial attention is required to analyse conjunction targets.

Modulations in extrastriate cortical regions specialised for processing colour, form and motion during visual attention have been detected using neuro-imaging techniques, hence providing support that selective attention alters the perceptual inputs prior to completing feature analysis.

In short, neurophysiological findings have provided support for some psychological theories on attention such as the concept of early selection. To a certain extent the data revealed that the attentional system achieves its objectives by altering incoming visual signals when stimuli having relevant physical features are encountered. The major implication of these data is that **descending** projections (feedback pathways) from attentional control systems affect the excitability of neurons coding the features of the attended or ignored stimuli.

2.5 Summary

The information presented in this chapter forms the basis for the construction of the proposed visual scene analysis system. In explaining visual perception, two essential perceptual functions: object recognition and selective attention have been described.

Object recognition plays a significant role in visual perception. When we see, we do not describe the physical world with simple, meaningless features such as colours, shapes, lines or curves. We see the world as a recollection of familiar objects and events that are informative, and can be understood and comprehended. We have explained why object recognition is a difficult problem. Variability in sensory information and stored object representation are two computational problems that must be solved for successful object recognition.

We have also discussed the importance of attention in visual perception and its relation to object recognition. Basically, attention can be regarded as a controller, controlling both the system resources and the selection of information to be processed. Without attention one would lose

the ability to detect changes in the environment, thus the ability to switch focus. Furthermore, failures in the attentional system can cause object recognition to fail because linking of simple features to form a coherent percept requires attentional processing.

Psychological theories on selective attention have suggested that the attentional system is analogous to a limited-channel that allows attended information through unchanged or amplified while unattended information is passed in an attenuated form. Psychophysical experimental results further indicate that attention operates in two distinct modes in order to maintain a balance between alertness and focussed processing.

Neurophysiological experimental findings have enabled us to understand the neural mechanisms of attention. Some of the important results are:

- spatial attention occurs in part via the modulation of sensory processes in the visual cortex;
- modulation is required to analyse multi-feature targets;
- modulation may occur to features at an early stage; and
- descending projections from attentional control systems affect a neuron's excitability in feature coding. Feedback pathways have been found to boost activities in LGN cells.

Chapter 3

Object Recognition Approaches and Models

Visual processes are often classified into low- and high-level. Low-level processes are concerned with the manipulation of input signals for preliminary analysis of the image. For example, contrast and brightness adjustment, and the detection of lines, edges, colour, direction of motion and orientation of edges. High-level processes are involved in the interpretation of the image, i.e., the where and what of vision. They typically include shape extraction, object recognition and classification, visual attention and some intermediate processes such as feature grouping, segmentation and figure-ground separation.

In this chapter, we provide a literature survey of several commonly employed approaches to the high-level visual task of object recognition. In particular, we emphasize the approach of neuro-vision systems, and introduce two neural network models that are essential in our proposed neural architecture for visual scene analysis. Learning in neural systems is also considered. This chapter lays the theoretical and mathematical foundations of the research.

3.1 Computational Approaches

Methods and recognition systems developed using computational approaches are well established, and are traditionally from the fields of machine and robot vision, computer science, and artificial intelligence. Object recognition in these approaches is viewed as the establishment of a correspondence between an input image and its stored object representation. Generally, no considerations are given to the neural mechanisms that perform visual recognition, and the phenomena of visual attention are ignored. Furthermore, object models are usually not learned,

rather they are stored as some featural representations, and memory activation is not necessary.

In the following, we present a review of two major approaches of computational object recognition: model-based and appearance-based. These two approaches differ in their internal object representations. Model-based object recognition uses geometric features to construct object descriptions, whereas appearance-based representations are formed by large sets of images without any need for knowledge on the geometric structure of the objects. The review is primarily based on several survey papers and books [4, 14, 40, 154, 180, 191, 211].

3.1.1 Model-Based Methods

The central idea behind model-based recognition methods is the construction of a model using features extracted from 2D, $2\frac{1}{2}$D or 3D sources [40]. Recognition is achieved through matching input object features with stored models. There are three major components in a model-based approach. The first is the extraction of *features* that can adequately describe an object's physical properties and their spatial relations. The second is the construction of an object *model* based on the extracted features such that all objects in the same class can be recognised using the same model. Lastly, a *matching* process is required to establish correspondence between image features and object models in order to achieve object recognition.

The type of model constructed is categorised by the features used in the construction. 2D models are viewer-centred representations, i.e., viewpoint dependent, constructed using mainly 2D geometric features such as shape, edge, corner, line, curve, hole, and boundary curvature. Since these features are 2D in nature, they belong to an "image space". Full 3D description of an object using 2D features is rarely achieved, thus these models are limited by the number of viewpoints used in feature extraction.

$2\frac{1}{2}$D models are also viewer-centred representations, but instead of boundary features, surface features are used, e.g., depth and surface orientation. Thus $2\frac{1}{2}$D models are defined in a "surface space". An example of $2\frac{1}{2}$D representation is the $2\frac{1}{2}$D *sketch* by Marr [122]. These models generally provide a more accurate representation of the object than 2D models but they are also limited by the viewpoints used.

3D models are object-centred representations and volumetric in nature, providing full descriptions of objects from an unconstrained viewpoint. Thus 3D models are defined in an "object space". They represent exact specifications of objects using surface patches, spines, and volume primitives such as generalised cylinders, cubes, spheres, and rectangular blocks. An obvious difficulty is the need to define an efficient method for correspondence between 2D images and 3D models.

The choice of features used in model construction is an important factor in determining the success of a recognition system. A clever selection of features can enhance recognition rate and speed. A single feature is rarely enough to uniquely describe a class of objects, most often a combination of features is required. Geometric features can be classified into three types: global, local, and relational features. Global features are physical attributes of an overall object, examples of which are perimeter, centroid, area, curvature, contour points from centroid, and moments of inertia. Local features are physical attributes of a part of an object, some examples are line segment, corner, arc segment, and area of a salient region. Relational features are usually distance or orientation measurements that provide information about neighbourhood relations between local features or substructures of an object.

Depending on the situation, it is not always necessary to have all types of feature for object modelling, but in complex cluttered scenes all three types of feature are usually required. Global features are deceptive to occlusion and varying illumination condition, so models with local features can provide additional robustness. Relational features provide important information regarding the arrangements of object parts, without which an object can be disassembled into its parts, may yet still have the same global and local features.

Matching in model-based systems is a model-driven process in that the features used in model construction must be used in the matching process, therefore the choice of object features dictates the recognition algorithm. The literature for model-based recognition methods is vast and for illustrative purposes we restrict our discussion to 2D recognition methods only. Generally, 2D models constructed from global features use statistical pattern-recognition schemes, local feature based models employ syntactic and hierarchical matching methods, while relational feature models use graph-matching techniques. In the following, three classes of method are discussed: (i) *feature spaces method*, (ii) *parts and structural method*, and (iii) *alignment method*.

Feature Spaces Method

The most critical part of feature spaces methods is the choice of features for object model construction. Ideally, invariant features that are common to all views are selected. Object moments and Fourier descriptors are examples of invariant features that have been suggested [147, 149]. Recognition is achieved by matching features of an object with those of the model.

However, invariant features are difficult to acquire. Different objects may share some features but not all, or a feature extracted is invariant for a limited range, thus it is possible to define each object uniquely using a combination of features, i.e., individually each feature may not be invariant but together as a whole it is.

If a view of an object is described by n features, it can be represented by a vector with n elements, so that the view is a point in an n-dimensional space, R^n, referred to as the *feature space*. It follows that an object which is completely described by its views is a cluster of points or a subspace in R^n. The features are relatively invariant if the subspace spanned by the views is compact. Furthermore, objects can be identified uniquely if the subspaces are non-overlapping. Recognition of an object usually involves statistical measures to determine which subspace that the point representing the object view is closest to. Alternatively, objects are represented as linearly separable subspaces such that an object is uniquely identified by the subspace that it occupies.

Parts and Structural Method

Instead of using global physical properties and attributes, this method describes objects in terms of their local geometric features such as arcs, lines and corners, or higher-level generic components that are composed of the geometric features, for example boxes and cylinders. Models are constructed by decomposing objects into parts with semantic information and geometrical relations between components stored implicitly as an ordered list or explicitly as a graph.

Recognition involves locating those local features and parts first, and then the structural and relational information is used to verify the identity of an object. Since recognition is carried out at a local level, this method is less affected by occlusion and illumination conditions. Having said that, structural description methods are limited in several ways. While man-made objects can usually be decomposed into simple geometric parts, this is not so for natural objects. Even if an object is decomposed into some generic parts, it is often insufficient to uniquely identify the object. A well known example of recognition by components is discussed in [16].

Alignment Method

This method assumes that there is a solution to the spatial correspondence problem in which the view of an object, V, can be mapped to its object model, M, through a series of transformations, or vice versa [189, 191]. Each of the set of transformations, such as changes in position, orientation or scale, may be represented by a matrix, T. These transformations are applied explicitly to either the incoming image or the stored model.

Recognition of a viewed object becomes a process of locating a model and a transformation matrix that can maximise the matching between the model and the viewed object. The search for a transformation matrix and a model is a two-stage process of hypothesis generation and verification. To find a suitable transformation matrix, hypotheses are generated for all possi-

ble matches between the image and the model. Followed by the verification stage, in which small errors are allowed in the matching between the image and the model according to some thresholds. Occlusion can be handled by generating hypotheses using local feature sets from non-occluding parts of the model and the image.

3.1.2 Appearance-Based Methods

In recent times, an object recognition approach based on an object's appearance rather than its shape has received considerable attention [86, 133, 154]. The *appearance-based* approach has been successfully applied to 3D recognition of a large collection of complex objects [138], as well as recognition in the presence of clutter and occlusion [165], with robustness and efficiency.

The success of the approach can be attributed to the power of the appearance representation which is both compact and descriptive. Besides shape information, the appearance representation includes both intrinsic and extrinsic visual information such as surface reflectance properties and illumination conditions. Since an appearance representation is a viewer-centred representation, each object is represented by a large number of views of the object. Such views can be simple 2D, image-like representation of the object. These can be acquired easily without any prior knowledge of the object, allowing direct training from visual data. Therefore the approach is relatively general and can be applied to a variety of object types. To achieve compactness, the object views are usually transformed into a representation in a low-dimensional space.

In contrast, model-based methods require construction of models manually, which is often difficult and time-consuming. The type of model constructed is based on the features extracted, so object modelling is case dependent, requiring experience and knowledge for effective feature selection. Thus the process of object modelling cannot be performed without human assistance.

The appearance of an object is not restricted to shape or reflectance information, spatial frequency descriptions such as discrete cosine transform, Fourier descriptors, wavelets, or eigenimages can also be used as long as the primary visual features are captured by the representation.

For 3D object recognition, the appearance of an object is the combined effect of its shape, surface reflectance properties, pose in the scene, and the illumination conditions. Nayar, Murase and Nene [137] acquire pose and illumination information from an image sensor using two robot manipulators; one for rotation, while the other varies the illumination direction. As a result, a large set of object images with high correlation among them is generated. To efficiently search for the corresponding object representation for a particular scene, the large set of training images is compressed into a low-dimensional representation of object appearance.

Principal Component Analysis

One image compression method that has been successfully applied in appearance object representation is the Karhunen-Loéve transform or *principal component analysis* [61, 142]. In this method, all images are projected to an orthogonal space, they are then reconstructed by using their principal components only.

Consider a set of n images of size $N = p \times q$ represented as N-dimensional vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$. For highly correlated images, the image vectors form a cluster in the N-dimensional space. In order to reduce the dimensionality of the image space, all the images are projected onto a lower dimensional space by minimising the mean squared error between the images and their projections. The average of the images, also the centre of the cluster, $\hat{\mathbf{x}}$, is an important reference point and is given by

$$\hat{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i, \tag{3.1}$$

and the corresponding covariance matrix, $\mathbf{C}$, can be computed:

$$\mathbf{C} = \frac{1}{n-1} (\mathbf{x}_1 - \hat{\mathbf{x}}, \mathbf{x}_2 - \hat{\mathbf{x}}, \ldots, \mathbf{x}_n - \hat{\mathbf{x}})(\mathbf{x}_1 - \hat{\mathbf{x}}, \mathbf{x}_2 - \hat{\mathbf{x}}, \ldots, \mathbf{x}_n - \hat{\mathbf{x}})^T. \tag{3.2}$$

$\mathbf{C}$ is an $N \times N$ matrix that allows us to determine the associated eigenvectors $\mathbf{e}_i$ and eigenvalues λ_i through the identity

$$\lambda_i \mathbf{e}_i = \mathbf{C}\mathbf{e}_i. \tag{3.3}$$

The projection of the images to an M-dimensional space or *eigenspace*, such that $M < N$, is given by

$$\mathbf{y}_i = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_M]^T (\mathbf{x}_i - \hat{\mathbf{x}}) \tag{3.4}$$

where $\mathbf{y}_i$ is the projection of $\mathbf{x}_i$ and is a point in the eigenspace.

In principle, all N eigenvectors $\mathbf{e}_i$ are needed to completely describe the input image set, but usually a small number of eigenvectors, ($M \ll N$), is sufficient to capture the significant appearance characteristics of the set.

Parametric Eigenspace Representation and Recognition

The projection of the images yields a set of discrete points in the eigenspace. With high correlation among the images, the points are expected to be closely located. One can construct a manifold for a continuous appearance function based on these points. Nayar *et al.* [137] used a standard quadratic B-spline interpolation algorithm [157] to construct such a manifold.

Each class of object will have a different manifold. Object identification is according to the distance between the projection of an image and the closest manifold in the eigenspace. The simplest distance measure is the Euclidean distance

$$d = \sqrt{(\mathbf{x} - \mathbf{y})^T(\mathbf{x} - \mathbf{y})} \tag{3.5}$$

where $\mathbf{x}$ is the projection of an input image and $\mathbf{y}$ is a manifold in the eigenspace. The matching problem is simply to find the minimum distance, d_{min}, between $\mathbf{x}$ and some manifold $\mathbf{y}_i$:

$$d_{min} = \min_i ||\mathbf{x} - \mathbf{y}_i|| \tag{3.6}$$

If d_{min} is within a certain threshold, we can conclude that the input image $\mathbf{x}$ belongs to the manifold $\mathbf{y}_i$, and the object class i.

Since the appearance representation is acquired from global visual data, occlusion and variations in scene illumination can cause problems for matching. Various approaches have been proposed to deal with such problems, in particular, Huang *et al.* [86] use principal components of segmented regions for indexing; Rao [154] tackles the occlusion problem based on the memorisation of the responses of a set of steerable filters, and Nelson and Selinger [140, 141] use contour patterns in keyed context regions.

3.2 Neuro-Vision Systems

Neuro-vision systems can be described as artificial neural machines that are designed to see and perceive the visual world. Motivated by the advances in the fields of biology, psychology, physiology and cognitive neuroscience on the understanding of various aspects (behavioural, functional, structural, and computational) of visual processing, and the lack of significant progress in traditional computer and machine vision. The approach seeks to use artificial neural network paradigms that are inspired by, or based on, the biological visual system for the development of high-level vision systems.

From an engineering point of view, an artificial vision system must be effective, efficient and robust under a variety of visual conditions. While the field of machine vision has made good progress on low-level visual tasks such as feature detection, high-level functions such as object recognition in general conditions is still a challenging problem. By realising some of the strengths of the biological visual system, it is hoped that we may overcome some of the shortcomings of machine vision systems that have hindered the progress of artificial vision.

The goal of neuro-vision is not to emulate the precise physiological mechanisms and structure of the visual system, rather it is more important to replicate the neuronal computational structures

and emulate the neuronal computational principles for processing, representing, storing, and interpreting visual information. Therefore neuro-vision systems are expected to have some or all of the strengths derived from the biological neural networks.

Examples of neuro-vision systems are numerous, tackling a variety of problems, ranging from low-level image processing and compression, texture analysis, and frequency domain analysis [48, 78, 119, 150, 151] to high-level visual functions such as pattern recognition [65, 64, 105, 201], visual perception [1, 159], preattentive vision [75], colour analysis [130, 192], 3D object recognition [106, 177], stereo vision [136], automatic target recognition systems [181, 197], and visual search models [203], to name a few. In the following sections, we review the computational basis for artificial neural networks and learning in neural nets in general.

3.2.1 Artificial Neural Networks

Artificial neural networks (ANNs) are highly simplified models of their biological counterparts, intended for exploiting the computational benefits of their massively parallel networking architecture. Some of desirable characteristics of ANN systems include massive parallelism, distributed representation and computation, learning ability, generalisation ability, adaptivity, inherent contextual information processing, fault tolerance, and low energy consumption. Multilayer feedforward type of neural networks are particularly popular due to their ability to approximate any nonlinear function [67, 83, 84, 85, 101, 200]. There are other neural network paradigms [91], some of which will be discussed in a later section.

An ANN is a highly interconnected network of simple processing units called *neurons* or *cells*. There are three basic elements for a neuron, as follows:

- *Synapses.* Neurons are connected together via synapses or more correctly synaptic connections. They act as information and communication channels. Each synapse is characterised by its *weight*. For example, an input signal x_j to neuron k is modulated by the synaptic weight w_{kj}. Positive weights represent excitatory connections, while negative weights represent inhibitory.

- *Adder.* A neuron is connected to a number of synaptic links, each carries an input signal, therefore an adder is needed to sum up all the input signals to the neuron.

- *Activation function.* A neuron is only activated if the summed input, u_k is greater than a certain threshold, θ_k. The activation function serves to limit the amplitude of the output of the neuron, y_k.

Mathematically, a neuron k with p synaptic inputs, x_j $(j = 1, \ldots, p)$, a weight vector, w_{kj}, and an activation function φ, can be described by the following pair of equations:

$$u_k = \sum_{j=1}^{p} w_{kj} x_j = \mathbf{w}_k^T \mathbf{x} \tag{3.7}$$

and

$$y_k = \varphi(u_k - \theta_k). \tag{3.8}$$

There are three basic types of activation function:

1. Threshold function

$$\varphi(v) = \begin{cases} 1 & \text{if } v \geq 0 \\ 0 & \text{if } v < 0. \end{cases} \tag{3.9}$$

2. Piecewise-linear function

$$\varphi(v) = \begin{cases} 1 & \text{if } \quad v \geq A \\ v & \text{if } -A < v < A \\ 0 & \text{if } \quad v \leq -A \end{cases} \tag{3.10}$$

   where $-A < v < A$ defines the linear region.

3. Sigmoid function

$$\varphi(v) = \frac{1}{1 + \exp(av)} \tag{3.11}$$

   where a is the slope parameter of the sigmoid function.

3.2.2 Learning in Neural Nets

One of the most powerful properties of a neural network is its ability to learn from its environment. Performance of a neural network system can improve through learning over time. The learning process has been accurately described by Mendel and McLaren [126]:

> Learning is a process by which the free parameters of a neural network are adapted through a continuing process of stimulation by the environment in which the network is embedded. The type of learning is determined by the manner in which the parameter changes take place.

Mathematically, the synaptic weight w_{kj} at time $(t+1)$ is determined by its value and adaptation in weight Δw_{kj} at time t:

$$w_{kj}(t+1) = w_{kj}(t) + \Delta w_{kj}(t) \tag{3.12}$$

3.2.3 Network Architectures and Learning Paradigms

ANNs can be regarded as directed graphs, consisting of nodes with interconnecting synaptic and activation links. Based on the graph connection pattern, two network architectural categories can be defined:

i. *Feedforward* networks, in which graphs have no loops, and the data flow from input to output is strictly feedforward.

ii. *Feedback* or *recurrent* networks, in which loops exist between nodes. Contrary to the feedforward type, dynamic properties of the network are important.

Examples of feedforward networks are multilayer perceptron and radial basis function networks, while major models for the feedback type include competitive, Self-Organising Map, Hopfield, and Adaptive Resonance Theory networks [31, 57, 79, 107, 125].

Besides architectural differences, neural networks are typically distinguished by their learning paradigms, which can be sorted into two classes:

- *Supervised* or *associative* learning, in which the network is trained by an external teacher, providing it with matching input and output patterns. Weights are updated to allow the network to produce an output as close to the correct training output as possible for a given input.

- *Self-organisation* or *unsupervised* learning, in which an output node is trained, without a teacher, to respond to a cluster of patterns from the input, by exploring the underlying structure or correlations between patterns in the data statistically. The resultant output nodes represent categories extracted from the data.

Under both learning paradigms, there are four basic types of learning rules (principles under which weights are updated): error-correction learning, Hebbian learning, competitive learning, and Boltzmann learning. Each learning rule can be performed in one or more ways, which we refer to as learning algorithms. For example, error-correction can be implemented using the back-propagation learning algorithm or the perceptron learning algorithm.

Some of the common neural architectures and their associated learning algorithms are briefly described below.

Multilayer Perceptron

The most well known and popular class of neural networks is multilayer perceptrons, whose structure consists of an input stage, one or more hidden layers, and an output layer of nodes successively connected in a feedforward manner with no interconnections between nodes in the same layer and no feedback connections between layers.

Multilayer perceptrons are trained in a supervised manner with the error-correction learning rule, which is commonly implemented using the back-propagation learning algorithm [158]. Learning using the back-propagation algorithm is a two phase process: a forward phase and a backward phase. In the forward phase, an input pattern is applied to the input stage and is allowed to propagate through the network, thereby generating a set of outputs. Comparing the actual and desired outputs produces an error signal which is used in the backward phase to adjust the synaptic weights so as to minimise the differences between the actual and desired outputs. As such, the back-propagation algorithm is a generalisation of the least-mean-square algorithm. The error cost function most frequently used in the back-propagation is given by

$$E = \frac{1}{2}\sum_{i=1}^{p} \|\mathbf{d}_i - \mathbf{y}_i\|^2 \tag{3.13}$$

where p is number of training patterns, $\mathbf{d}_i$ is the desired output response, and $\mathbf{y}_i$ is the actual output response.

Hopfield Network

The Hopfield network is a recurrent network that stores information in a dynamically stable configuration [82]. It acts as a nonlinear associative memory that can retrieve a stored pattern in memory upon presentation with an incomplete or noisy version of that pattern. The Hopfield network learns in an unsupervised manner according to *Hebb's postulate of learning* using associative memory learning. Weights in the Hopfield network are symmetrical and non-self-feedback, i.e., if w_{ij} is the synaptic weight, then $w_{ij} = w_{ji}, \forall i, j$, and $w_{ii} = 0, \forall i$.

Self-Organising Feature Map

Kohonen's self-organising feature map (SOFM) is another unsupervised neural architecture [100]. Basically, a self-organising feature map is a topographical map of the input patterns, in which the spatial locations of the neurons in a lattice structure correspond to intrinsic features of the patterns. Learning is based on the competitive learning rule or *winner-take-all* (WTA) in which only one output node is active at any given time. The algorithm used in SOFM

learning simply transforms an input pattern of arbitrary dimension into a 1D or 2D discrete map by computing the Euclidean distance between the input and the stored weight.

3.2.4 Related Neural Architectures for Object and Pattern Recognition

The neural architectures introduced so far are more for general purposes and are applicable to a wide range of problems. In this section, we review several specialised neural architectures for object or pattern recognition that are related to the theme of this thesis.

The Neocognitron

The neocognitron by Fukushima [64, 66] is a neural architecture that is closely related to the one proposed in this thesis. It is loosely based on known properties of the mammalian visual system for visual pattern recognition. It was motivated by the need to overcome the inability of the cognitron [62] to recognise position-shifted or shape-distorted patterns.

Figure 3.1 shows the basic topological structure of the neocognitron. It consists of a number of functionally equivalent stages, shown as U_{sl} and U_{cl}, where $l = 1, \ldots, n$ for an n layer neocognitron. Each U_{sl} layer is composed of *S-cells* and each U_{cl} of *C-cells*. There are also two other cell-types, called V_s and V_c. These cells only serve to normalise the activities of the S- and C-cells, thus are not required for understanding the basic operations of the neocognitron.

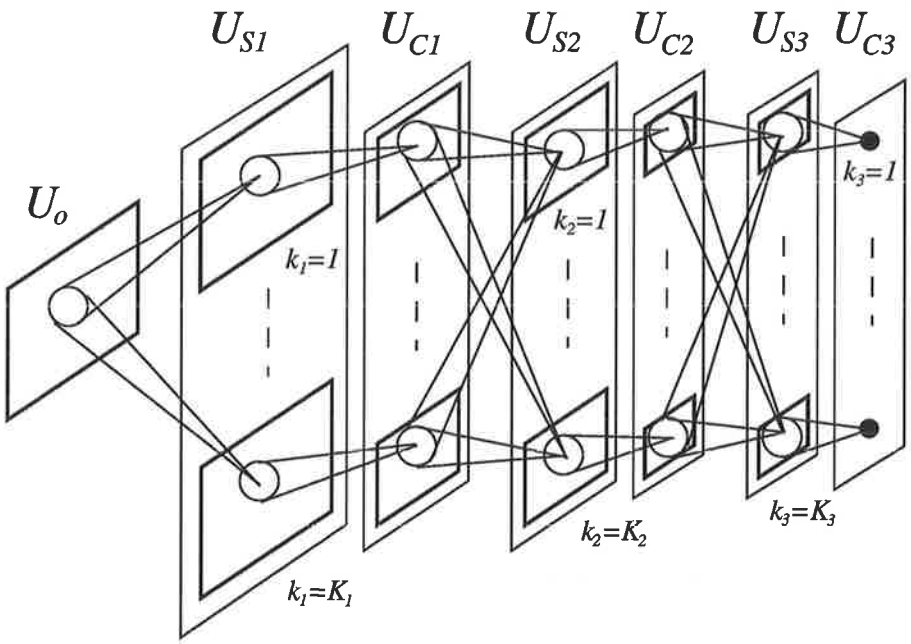


Figure 3.1: Topological structure of the neocognitron. Adapted from [66].

The S-cells, denoted by u_s, are feature extracting cells. Connections to u_s are variable and reinforced by learning. The C-cells, denoted by u_c, are for positional errors in the features of the input pattern. Connections from u_s to u_c are fixed and invariable. Cells in each layer are divided into K subgroups or S- and C-planes, indexed by $k_l = 1, \ldots, K_l$, corresponding to K different features. Within each S-plane, cells receive input connections of the same spatial distribution. This has the effect of detecting a particular pattern of activities in the preceding C-layer by correlation. Each S-plane is followed by a C-plane that contains fewer cells. The C-plane forms a compressed representation of the features extracted in the S-plane. Each u_c cell receives inputs via fixed connections from a group of u_s cells that extract the same feature. The u_c cell is activated if any of those u_s cells is active.

The entire network of alternate U_{sl} and U_{cl} layers allows individual features of the input pattern to be extracted and compressed repeatedly. This ensures the features are detected wherever they lie in the input layer, and the successive stages of compression provides tolerance to changes in size and shape.

An extended model with backward connections added to the neocognitron was proposed [63, 65] in order to recognise two or more patterns in the input layer. This extended model performs many functions that are similar to our proposed neural architecture, for example, selective attention, pattern segmentation, and associative recall. Although the neocognitron has many important properties and functions shared by our proposed neural architecture, both the architecture and implementation differ significantly. We have chosen to use the adaptive resonance theory neural architecture as a basis to model selective attention because its ability to handle the *stability-plasticity dilemma* [28]. Moreover, the degree of translation in the neocognitron is controlled by the number of layers [9]. Menon and Heinemann [127] reported that the neocognitron did not perform satisfactorily when it had to deal with large shifts in a 128×128 pixels image. Therefore, the neocognitron is suitable for applications with small shifts such as digit and character recognition (good performance on digit recognition has been demonstrated with small shifts in a 16×16 image [66]). For visual scene analysis, the number of layers required is exceedingly large, and in most cases impractical to implement. Most of our simulated scenes in Chapter 5 are larger than 200×200 pixels. Also, the original neocognitron lacks the mechanisms to handle occluded and arbitrarily rotated patterns. Rotation invariance in neocognitron has since been proposed by Satoh *et al.* [160, 161].

Adaptive 3D Recognition from Multiple Views

Seibert and Waxman [162, 163] developed a hierarchical real-time neural architecture for 3D object recognition from multiple 2D view sequences. The approach takes a viewer-centred

representation for 3D objects based on appearance. 3D object representations are constructed from both views and view transitions using the *aspect graph* concept [99]. An aspect graph consists of a number of nodes or aspects representing characteristic views of an object. Each characteristic view is formed by a number of invariant 2D appearance descriptions clustered together using the ART2 [29] learning algorithm. Between two connected nodes is an arc representing the visual event (aspect transition) that is linking the two aspects. Each aspect provides a partial description of the 2D appearance of one or more objects. Together with the aspect transitions the nodes form a complete description of a 3D object.

This work has introduced many novel concepts, in particular the use of aspect transitions as part of the 3D representation modelling can significantly improve the real-time recognition accuracy. However it has not considered more realistic visual scenarios, where objects are often in partial occlusion and embedded in cluttered backgrounds.

VIEWNET

This is another neural architecture for 3D object recognition from multiple 2D views. In fact, VIEWNET [12, 74] was inspired by the work of Seibert and Waxman [162]. The simplest VIEWNET consists of three parts: a preprocessor for generating compressed invariant 2D representation of an image, a self-organising pattern recognition network based on the Fuzzy ARTMAP [32, 34], and a working memory to accumulate evidence over multiple views. As in the neural architecture of Seibert and Waxman, VIEWNET also uses log-polar transform to achieve size and orientation invariant 2D aspects of 3D objects.

While good recognition performance has been achieved with noisy and clean images, VIEWNET does not address recognition of incomplete objects due to occlusion or cluttered images. Both VIEWNET and the approach of Seibert and Waxman assume the pattern to be recognised has been separated from its background.

3.3 Adaptive Resonance Theory

Of all the neural network models described so far, none can be used to explain biological cognitive data and processes better than the Adaptive Resonance Theory (ART) network. For example, ART has been used to explain data on visual perception, speech perception, and neural substrates of learning and memory [73]. Besides, ART has a number of useful properties that have enabled it to be applied in a variety of applications [11, 35, 89, 93, 102, 210]. For these reasons, ART has been chosen as the basis for our proposed visual scene analysis system.

ART was first proposed by Grossberg [71, 72] as a theory of human information processing. The theory was motivated by the *stability-plasticity dilemma* in competitive learning. A neural network must be stable enough to preserve significant past learning, and yet remain adaptible enough to incorporate new information should it arise. Therefore, ART allows for the learning of new input patterns on an incremental basis, while preventing the erosion or corruption of past memories. In contrast, the multilayer perceptron must be retrained using the entire set of inputs each time a new input is presented.

An ART network, shown in Figure 3.2, is a feedback neural architecture that is capable of self-organising stable pattern recognition codes in response to arbitrary sequences of input patterns. It consists of two neural fields: F1 and F2, which are interconnected by a pair of adaptive filters containing long-term-memory (LTM) weights. The architecture is the result of two complementary subsystems. Familiar patterns and top-down expectations are processed within an attentional subsystem, which consists mainly F2 and the top-down adaptive filter. An orienting subsystem is also required to process unfamiliar patterns and reset the attentional subsystem when an unfamiliar pattern appears. Learning is performed in an unsupervised manner, based on the competitive learning rule such that synaptic weights are adaptively changed in an approximate match phase or a *resonant* state. It can self-stabilise in learning while maintaining plasticity. An ART network also has three useful properties: *normalisation, contrast enhancement,* and *short-term-memory (STM) reverberation loops*. Normalisation has the benefits of becoming adaptive to large changes in input patterns, and allowing direct access to category representation without search after learning stabilised. Contrast is enhanced through nonlinear feedback processes and normalisation in the STM loops, thus noise is separated and suppressed from the input signal.

The critical property of self-stabilisation is achieved by the introduction of top-down pathways and matching mechanisms. The signals from the top-down pathways can be regarded as learned expectations. They enable the network to perform attentional priming, pattern matching, and self-adjusting parallel search, all of which help stabilise learning in response to arbitrary sequences of input patterns.

Figure 3.3 illustrates a typical ART search cycle. A STM pattern X is generated across F1 when an input pattern I is presented to an ART network. A signal pattern S is sent from F1 to F2 through the bottom-up adaptive filter. As a result, S is transformed to T, which in turn activates a STM pattern Y in F2. Y represents an internal category. The output of F2, U, is transformed to V through the top-down filter into F1. X becomes a new STM pattern formed by the common features between I and V. X is then compared with I. If the two patterns are sufficiently close, then learning may proceed, otherwise a reset signal is triggered suppressing the currently active F2 pattern Y. After Y is inhibited, the top-down pathway is eliminated, and

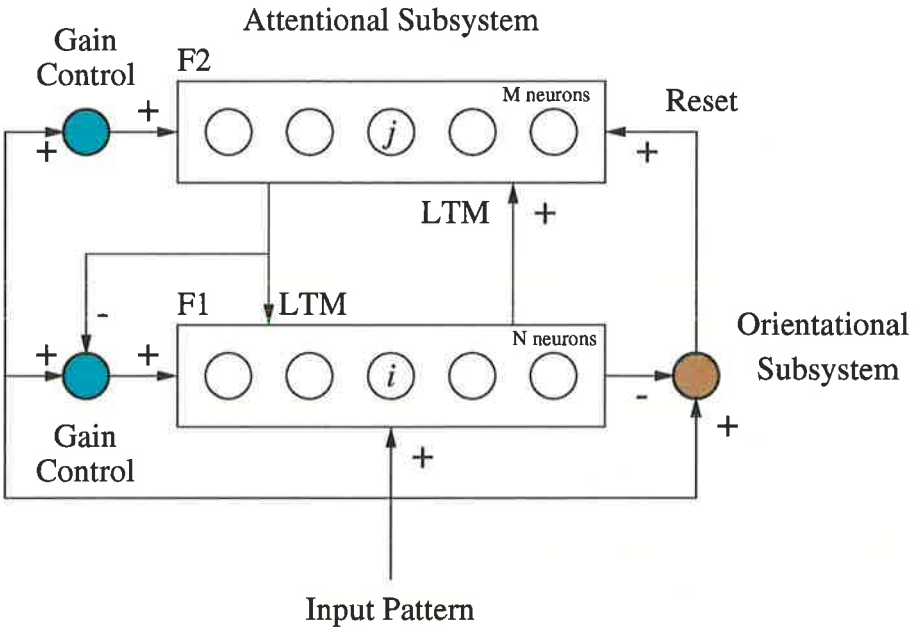


Figure 3.2: An ART1 architecture. Adapted from [28].

the original X is restored. X is once again transformed into T, which leads to the activation of a different Y in F2. This search cycle continues until a match is found or all established categories are inhibited, in which case the pattern I is established as a new category.

A family of ART based neural architectures have since been proposed, including: ART1 for binary inputs [28], ART2 for both binary and analog inputs, ART3 for hierarchical neural architectures [30], ARTMAP and FuzzyARTMAP for supervised self-organisation of memory codes [32, 33, 34], and other variants of the ones mentioned here. For our research, we focus mainly on the ART2 and ART3 architectures.

3.3.1 ART2 and ART3

ART2 is the second generation of the Adaptive Resonance Theory networks. The primary intention of ART2 is to overcome some of the shortcomings of ART1 such as noise deception, category proliferation, non-distributed representation, and non-analog input representation. The major structural difference of ART2 from its predecessor ART1 is the incorporation of several preprocessing layers (STM loops) in F1, which allows ART2 networks to stably categorise sequences of analog input patterns, by performing operations such as contrast enhancement, noise suppression and normalisation. An ART2 architecture is depicted in Figure 3.4(a).

Unlike the major architectural change from ART1 to ART2, ART3 is simply an extension of

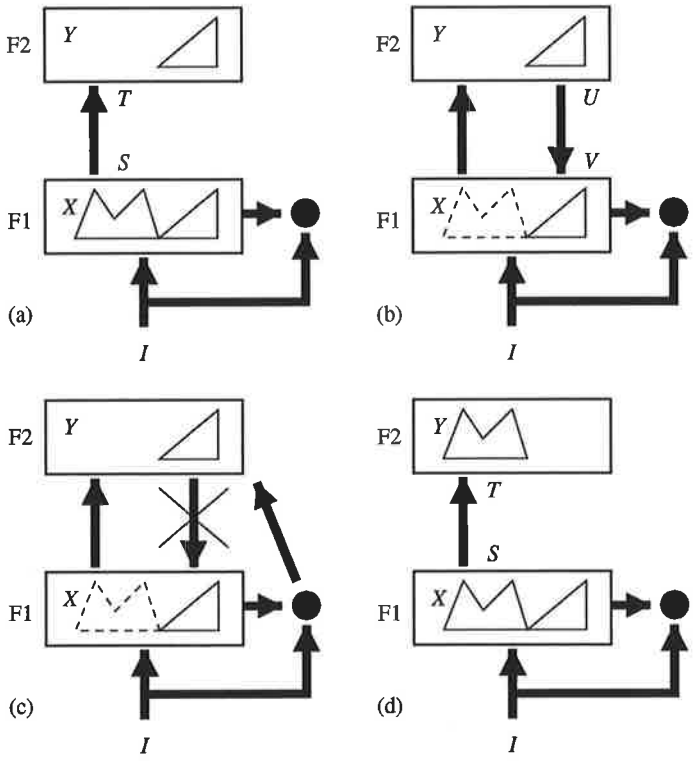


Figure 3.3: An ART search cycle. Adapted from [28].

ART2 with a more sophisticated parallel search mechanism. It enables ART to be embedded in network hierarchies. ART3 achieves this robust parallel search of a learned pattern recognition code by modelling computational properties of the chemical synapse in both bottom-up and top-down memory pathways. Among the properties modelled are intracellular operations such as transmitter accumulation, release, inactivation, and modulation. Another novel feature in ART3 is that it allows recognition codes in compressed or distributed representations.

An ART3 neural architecture is depicted in Figure 3.4(b). It is clear from the figure that STM loops are implemented in the same way for both ART2 and ART3. Since learning was not mentioned in the ART3 article [30], it is assumed that learning in ART3 is as given in ART2 [29]. Thus, it is sufficient to consider ART3 alone.

We begin reviewing ART3 from its search mechanism which is realised by a model of the chemical synapse. Figure 3.5 shows a model of the chemical synapse for the bottom-up pathways in Figure 3.4(b). The model is characterised by the dynamics of production and release of chemical transmitter, the inactivation of transmitter at postsynaptic sites, and the modulation of these transmitter processes. As a result, the postsynaptic cell is driven by the net excitatory or inhibitory signal arisen by the bound transmitter. According to Figure 3.5, we may write a set of differential equations in terms of u_{ij}^{bc}, v_{ij}^{bc}, z_{ij}^{bc}, S_i^{b3}, and x_j^{c1} to describe the dynamics of the

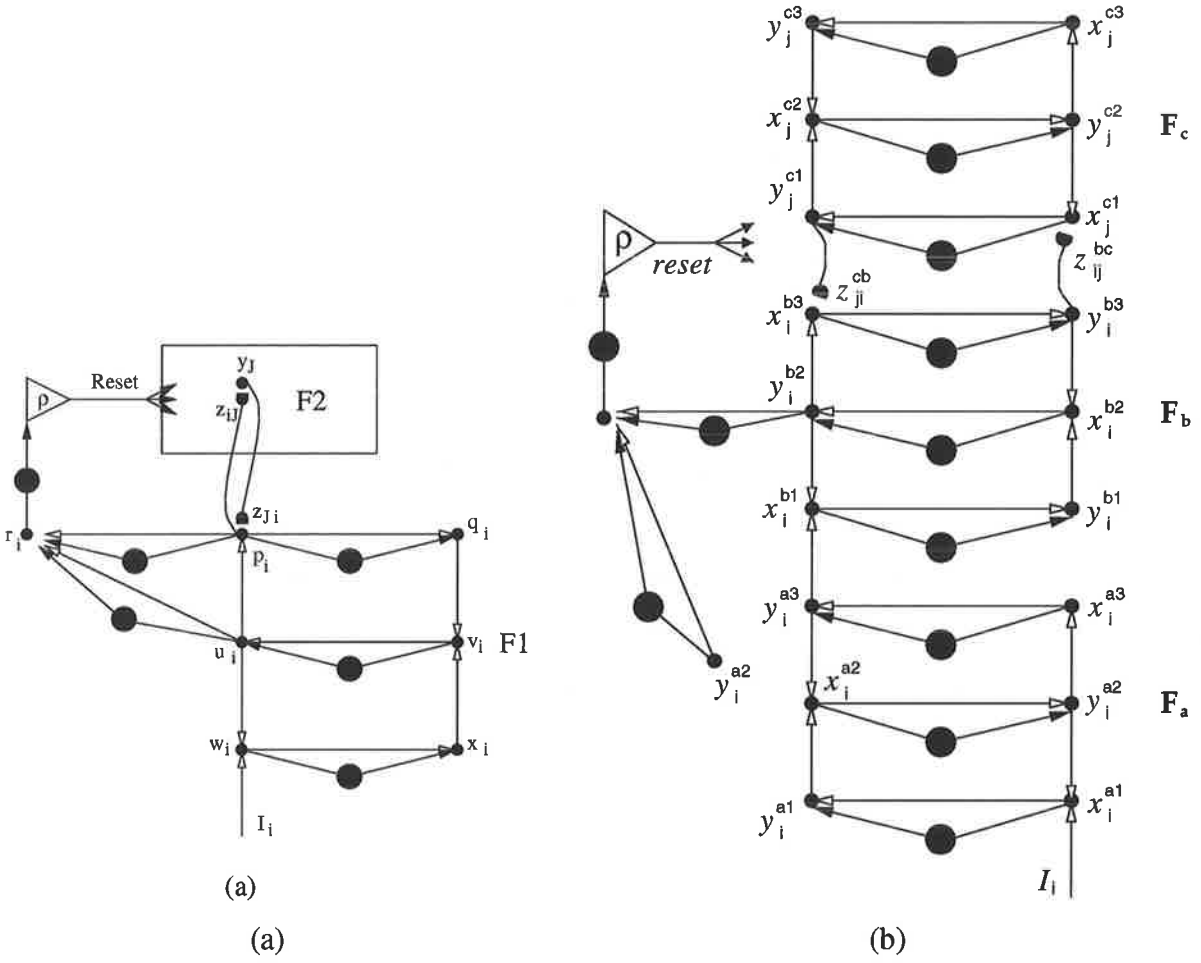


Figure 3.4: ART architectures: (a) ART2; (b) ART3. Large filled circles are gain control nuclei; small filled circles are nodes; and semi-circles are synapses. Adapted from [29, 30].

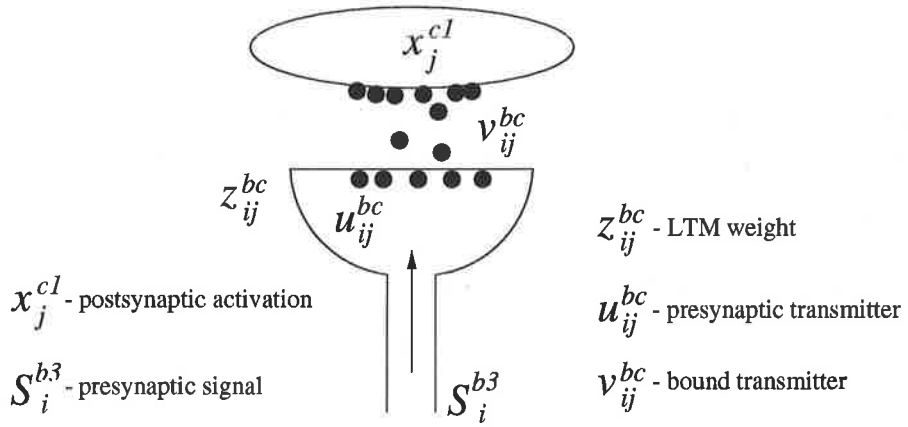


Figure 3.5: Model of the chemical synapse.

chemical synapse:

Presynaptic transmitter

$$\frac{du_{ij}^{bc}}{dt} = (z_{ij}^{bc} - u_{ij}^{bc}) - u_{ij}^{bc} p_5^c S_i^{b3} (x_j^{c1} + p_6^c). \tag{3.14}$$

Bound transmitter

$$\frac{dv_{ij}^{bc}}{dt} = -v_{ij}^{bc} + u_{ij}^{bc} p_5^c S_i^{b3} (x_j^{c1} + p_6^c). \tag{3.15}$$

Postsynaptic activation

$$\frac{dx_j^{c1}}{dt} = -x_j^{c1} + (A - x_j^{c1})[\sum_i v_{ij} + p_1^c S_j^{c2}] \tag{3.16}$$

where A is the upper limit of x_j^{c1} and $S_k^{nL} = g^n(y_k^{nL})$ is the presynaptic signal for field F_n, node k, and layer L.

Each field F_n in ART3 is completely described by the following set of equations:

$$x_i^{n1} = I_i^n + p_1^n g^n(y_i^{n2}) \tag{3.17}$$

$$y_i^{n1} = \frac{x_i^{n1}}{p_3^n + \|\mathbf{x}^{n1}\|} \tag{3.18}$$

$$x_i^{n2} = g^n(y_i^{n1}) + p_2^n g^n(y_i^{n3}) \tag{3.19}$$

$$y_i^{n2} = \frac{x_i^{n2}}{p_3^n + \|\mathbf{x}^{n2}\|} \tag{3.20}$$

$$x_i^{n3} = g^n(y_i^{n2}) \tag{3.21}$$

$$y_i^{n3} = \frac{x_i^{n3}}{p_3^n + \|\mathbf{x}^{n3}\|} \tag{3.22}$$

where g^n is the signal function for the nth field, p_k^n are nonzero constants, and $\|\mathbf{x}^{nk}\|$ is the L_2 norm. The signal function g^n can be selected to produce either distributed or choice codes. For choice networks with threshold θ, $g^n(y) = y$ if $y > \theta$, otherwise $g^n(y) = 0$

Matching is performed by finding the L_2 norm of the reset vector:

$$r_i^b = \frac{y_i^{a2} + y_i^{b2}}{p_3^a + \|\mathbf{y}^{a2}\| + \|\mathbf{y}^{b2}\|}, \tag{3.23}$$

so that matching is considered a success if $\|\mathbf{r}^b\| > \rho$, where ρ, called the vigilance parameter, is a predefined value that determines the degree of approximate match that needs to be satisfied prior to learning.

Simulations of Learning in ART

In this section, we present several simulations for learning in an ART neural network. The bottom-up long-term-memory (LTM) trace equation is given by

$$\frac{dz_{ij}^{bc}}{dt} = R\delta(S_i^{b3} - z_{ij}^{bc}) \tag{3.24}$$

where δ is the learning rate and R is a gating signal determined by the degree of match M and vigilance parameter ρ:

$$R = \begin{cases} 1 & \text{if } M > \rho \text{ and steady state} \\ 0 & \text{otherwise} \end{cases} \tag{3.25}$$

The following simulations are intended to illustrate several aspects of learning in ART: (i) uncommitted versus committed learning, (ii) match-reset tradeoff, and (iii) stability-plasticity tradeoff.

(i) Uncommitted learning simply refers to the creation of a new category for learning the current input pattern. In contrast, committed learning occurs in an existing category which can be seen as updating and reinforcing the category identity, or overwriting an existing category with a new one. Uncommitted learning is the result of mismatches with existing categories, while committed learning is performed in an approximate match phase under resonance.

On the left of Figure 3.6 are two very similar input patterns. Upon presenting the top input pattern to an ART system, the pattern is learned by an uncommitted node due to the nonexistence of memory. The learning process is displayed in the top row in four stages from left to right, showing the increase in weight size as learning progresses. When the bottom input pattern is presented, the same node is chosen due to their similarity, therefore the system is engaged in committed learning. In this case, it has the effect of updating an existing category.

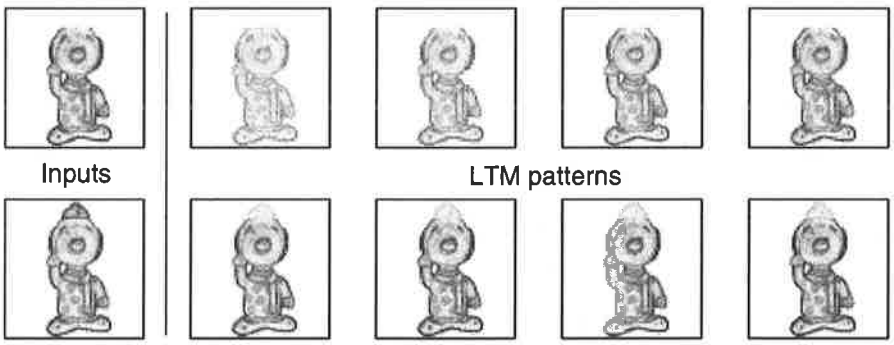


Figure 3.6: Committed versus uncommitted learning in ART.

(ii) Match-reset tradeoff: the criterion for a successful matching is determined by the level

of the vigilance parameter ρ. Therefore, under normal circumstances, ρ is a deciding factor in performing either committed or uncommitted learning. Under high vigilance, minor differences can cause matching to fail, whereas the same differences are tolerated for low vigilance levels, allowing committed learning to proceed.

Figure 3.7 shows five categories that are created when five input patterns of the same appearance are presented to an ART network under high vigilance.

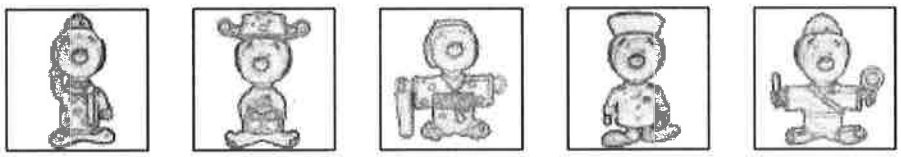


Figure 3.7: Learned categories under high vigilance.

We may observe the contrast enhancement property of ART in Figure 3.7 by increasing the threshold θ in the signal function g^n. The result of this contrast enhancement is shown in Figure 3.8.

Figure 3.8: Contrast enhancement in ART learning.

A very high vigilance level will ensure that each different input pattern will create a new category, however under a low vigilance only one category will be created and the resultant LTM pattern may take on one of several forms. This is discussed in (iii).

(iii) Stability-plasticity tradeoff: the system must be able to protect information stored at a given committed node, and yet maintain the ability to update the node with new or additional information, if needed. Once a committed node has been selected for learning, there are a number of factors that determine the outcome of the LTM pattern. Typically, we can identify LTM patterns produced by three types of system. The first is a highly stable system; the LTM pattern for the committed node is affected very little by the input pattern. The LTM trace tends to grow with learning, which is evident in Figure 3.9. Stable systems may be caused by slow learning rates, and large top-down feedback with weak input signal. It should be noted that each pattern in Figures 3.7 and 3.8 is a category, i.e., five categories in each figure, while in Figures 3.9-3.11 all five patterns represent the same category with respect to time (from left to

right). Each pattern corresponds to an instance when a new input pattern is presented, and the patterns are presented in the same order throughout.

Figure 3.9: Learning in a highly stable system.

On the other hand, limiting the top-down signal can increase the plasticity of the system. In the extreme case, the system becomes unstable and the LTM pattern for a committed node is overwritten each time the node is activated. Figure 3.10 shows such a case.

Figure 3.10: Learning in an unstable system.

Lastly, we can have a system that has the right balance between stability and plasticity such that the LTM pattern is a combination of the existing LTM pattern and the current input. Figure 3.11 shows the LTM pattern stored at a category is a blurry representation of all past input patterns combined together.

Figure 3.11: Learning in an intermediate system produces blurry LTM patterns.

3.3.2 Parameter Estimation - A Case Study of ART

As a further investigation of ART, a case study involving ART for solving a practical control problem is provided. The purpose of this study is to thoroughly examine the clustering nature of ART networks as a mapping function in the context of nonlinear dynamical systems. The

study will illustrate many ART related concepts such as input vector compression, clustering, and basins of attraction. We have chosen a control problem simply to demonstrate that ART is not restricted to visual applications.

Identification and parameter estimation problems have been tackled by artificial neural networks in recent years. Multilayer feedforward networks and radial basis function networks are particularly successful [5, 47, 135, 153] in solving these problems, due to their ability to represent arbitrary nonlinear mappings [67, 84]. An alternative approach based on ART, in particular ART2 and FuzzyART [29, 36] is introduced. It is hoped that through the use of a self-organising neural network, significant improvement could be gained in weight convergence time and also avoids difficult network topology design issues such as the number of hidden nodes and other related problems.

Problem Definition

Consider a simple inverted pendulum system as depicted in Figure 3.12. The dynamics of this system can be described by the following set of differential equations:

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ 9.81 \sin x_1 - 2x_2 + u \end{bmatrix} \tag{3.26}$$

where: $x_1 = \phi$ = angle between the pendulum arm and the zero degree position in clockwise.
$x_2 = \frac{d\phi}{dt}$ = angular speed of the pendulum measured in clockwise directions.
u = the external force applied to the pendulum by a motor.

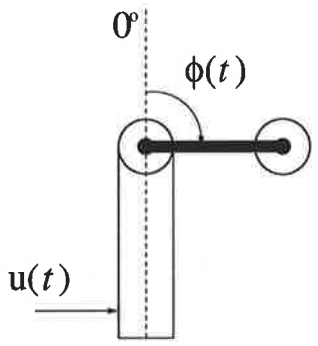


Figure 3.12: An inverted pendulum.

In the mathematical model of the system, the pendulum motion is characteristed by three factors: i) the force of the gravity on the pendulum represented by the $9.81 \sin x_1$ term, ii) the viscous friction acting against the motion by the $-2x_2$ term, and iii) the control signal u of the system.

Assuming that the model structure is unavailable but can be written in the form of:

$$\mathbf{X}_{t+1} = f(\mathbf{X}_t, u), \qquad \text{where} \quad \mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{3.27}$$

in discrete time. So given a particular state $\mathbf{X}_t$ and the control signal u, we can estimate the next state $\mathbf{X}_{t+1}$ by utilizing *a priori* information from the input-output data sets. In other words, we can use the current angular position and speed of the pendulum and the control signal to estimate the angular position and speed of the pendulum over one sampling period T based on experimental input-output observations of the system only.

This is a nonlinear mapping problem from $\Re^3 \rightarrow \Re^2$ and can be solved using any neural network that possesses the universal approximation property [83, 200]. Although ART networks are not universal approximators, they can form clusters of discrete states for mapping.

Methodology

The general neural approach to this kind of problems is to train a neural network to represent the unknown function f such that (3.27) becomes

$$\widehat{\mathbf{X}}_{t+1} = N(\mathbf{X}_t, u) \tag{3.28}$$

where $\hat{\mathbf{X}}_{t+1}$ is the next state estimate and N is a nonlinear function in the form of a neural network representing the nonlinear system.

The next state is given by

$$\mathbf{X}_{t+1} = \mathbf{X}_t + \Delta\mathbf{X}_t. \tag{3.29}$$

So instead of (3.28) we now have

$$\Delta\widehat{\mathbf{X}}_t = N(\mathbf{X}_t, u) \tag{3.30}$$

which is still a $\Re^3 \rightarrow \Re^2$ nonlinear mapping and is best illustrated by a block diagram in Figure 3.13.

The remaining issue is how to train a neural network to represent the nonlinear mapping in (3.30). With multilayer feedforward networks, a data set of angular position and speed and control signal combinations $(\mathbf{X}_t, u)^n$ covering the entire operating range together with the corresponding data set of state changes $(\Delta\mathbf{X})^n$ over one sampling period are generated. In practice where the system's mathematical model is unavailable or unknown the state change data set is obtained from the physical system by means of measurements.

Here the state change data set is actually obtained by solving the system model, (3.26), over one sampling period. The data set $(\mathbf{X}_t, u)^n$ is used as the network inputs while the set $(\Delta\mathbf{X})^n$

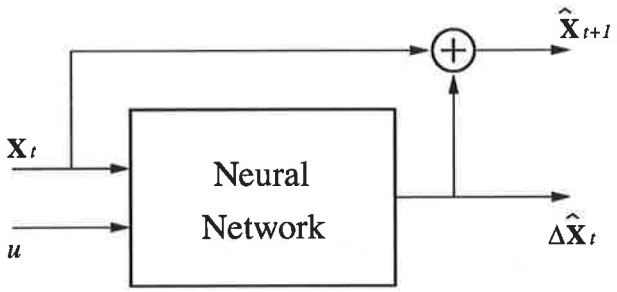


Figure 3.13: Parameter estimation block diagram.

provides the target vectors for which the network weights are being optimised:

$$|N(\mathbf{X}_t, u) - f(\mathbf{X}_t, u)| < \epsilon \tag{3.31}$$

thus enabling the network to achieve any desirable degree of accuracy.

In contrast ART networks do not perform optimisation of weights, instead input vectors containing sufficient similarities are grouped into "clusters", so that each cluster represents the centre of a region of attraction. In effect ART networks discretise the system input space and hence a compression of input vectors is achieved. This is illustrated in Figure 3.14, where the grey cubic region represents a continuous input space, the spheres are *basins of attraction*, and within each attraction region there is a cluster. Once learning is completed the clusters are linked to their corresponding state change vectors, the system becomes a content-addressable memory (CAM).

When an input vector is presented to the system, the cluster whose attraction region within which the vector falls is activated. The associated state change vector with the activated cluster is used to solve (3.29). It should be noted that this diagram represents the ideal case where clusters are evenly spaced, but in practice the clusters are usually scattered in space depending on the data distribution.

In the case of the inverted pendulum, when the first input vector $(\mathbf{X}_t, u)^1$ is presented to an ART network, it automatically becomes a cluster. If the next input vector $(\mathbf{X}_t, u)^2$ is within the predefined neighbourhood (unsupervised learning) of $(\mathbf{X}_t, u)^1$ then $(\mathbf{X}_t, u)^2$ becomes attracted to the cluster and is being grouped to that cluster. As a result of the grouping, the cluster or the centre of the attraction region is shifted to a position in between $(\mathbf{X}_t, u)^1$ and $(\mathbf{X}_t, u)^2$ (self-organising). However if $(\mathbf{X}_t, u)^2$ lies outside of the $(\mathbf{X}_t, u)^1$ neighbourhood then $(\mathbf{X}_t, u)^2$ becomes another cluster. This process continues until the entire continuous input space is represented by clusters, i.e., a continuous input space is being transformed into a discrete input space. These clusters are then pointed to the appropriate state change vectors. Subsequently an input vector will activate the cluster closest to itself and therefore acquiring the state change vector required.

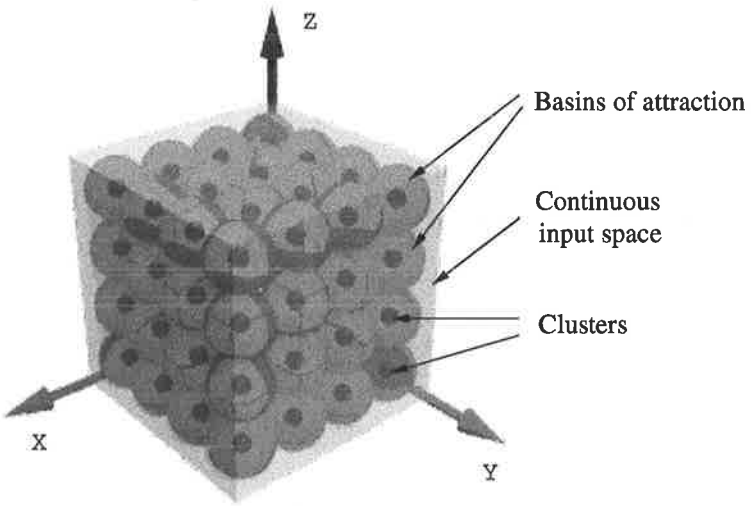


Figure 3.14: Cluster representation of a 3D continuous space.

This approach is often referred to as nearest neighbour search, and is similar to a lookup table. However with their massively parallel architecture, potential hardware implementations would allow this approach to be performed in a very efficient manner.

FuzzyART

A FuzzyART network has a similar structure to ART1. It also incorporates a similar pattern matching mechanism, in which the bottom-up input and top-down learned template are compared, leading to either the resonant state and prototype learning or to the mismatch state which triggers a search cycle. This search continues until an established category that satisfies the matching criterion is found or an untrained node is selected if no matching category can be found.

Although FuzzyART takes analog inputs as in ART2, its operations and characteristics are actually closer to those of ART1 due to the similarity of their structures. One major feature which highlights this is the comparison of input and template vectors, instead of comparing the vectors spatially, both ART1 and FuzzyART perform elementary comparison. For the binary case (ART1), elementary comparison is simply a bit-wise logical AND ($\cap$) operation between two vectors, whereas elementary comparison in the analog case (Fuzzy ART) is achieved by replacing the AND ($\cap$) operator with the MIN ($\wedge$) operator of fuzzy set theory. The MIN operator between two vectors $\mathbf{x}$ and $\mathbf{y}$ is also known as the fuzzy AND, and is defined by

$$(\mathbf{x} \wedge \mathbf{y})_i \equiv \min(x_i, y_i) \tag{3.32}$$

Simulation Results

In this section four sets of results comparing the relative performances of ART2 and FuzzyART for pendulum parameter estimation are presented, in each case both of the ART2 and FuzzyART networks are trained using the same data sets. These data sets are generated from samples of the pendulum input and output vectors within a predefined operating range so as to reduce the number of clusters required. So it is important to realise this method is only valid when applied to inputs within the trained region.

Simulation I

This simulation involves recalling some of the trained states. The purpose of this simulation is to examine the direct memory access property of ART networks. This is achieved by training the ART network with states that correspond to the actual behaviour of the pendulum system under a variety of initial conditions. That is, for a particular initial condition $(\phi_o, \dot{\phi}_o)_i$, we can obtain all subsequent states (assuming the free fall case, where $u = 0$) via the pendulum system. Upon presentation of the state $(\phi_o, \dot{\phi}_o)_i$ to an ART network trained with these states, it will recall all relevant states, thus the pendulum's dynamical response is generated.

In this simulation, the states used to train the ART network are actually generated using ten different initial conditions, each producing eighty one states with 0.05 seconds between the states, thus spanning a period of four seconds. The results for both ART2 and FuzzyART networks are shown in Figure 3.15(a) and (b), respectively.

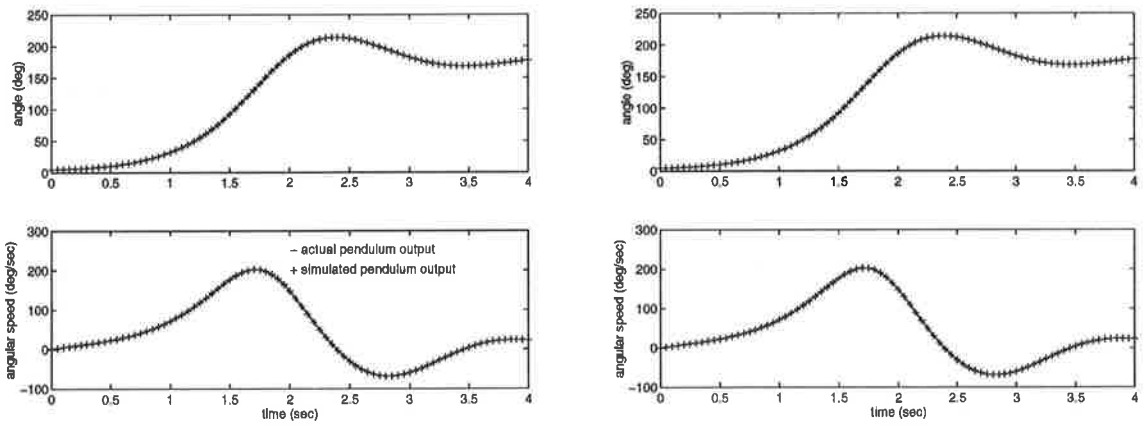


Figure 3.15: Complete recall of memory: (a) ART2; (b) FuzzyART.

Simulation II

Unlike the first simulation, this part requires the ART network to find the best match cluster

to the input state. The initial input state is chosen in such a way that none of the subsequent input vector has been used for training previously. This simulation can be used to verify the interpolation ability of ART networks. As in Simulation I, the ART networks are first trained with states corresponding to ten different initial conditions. An input state vector within the trained region but different from the training states is presented to the ART networks. Based on the clusters generated from the training states, the ART networks are able to estimate the pendulum behaviour. The resultant pendulum responses are shown in Figure 3.16.

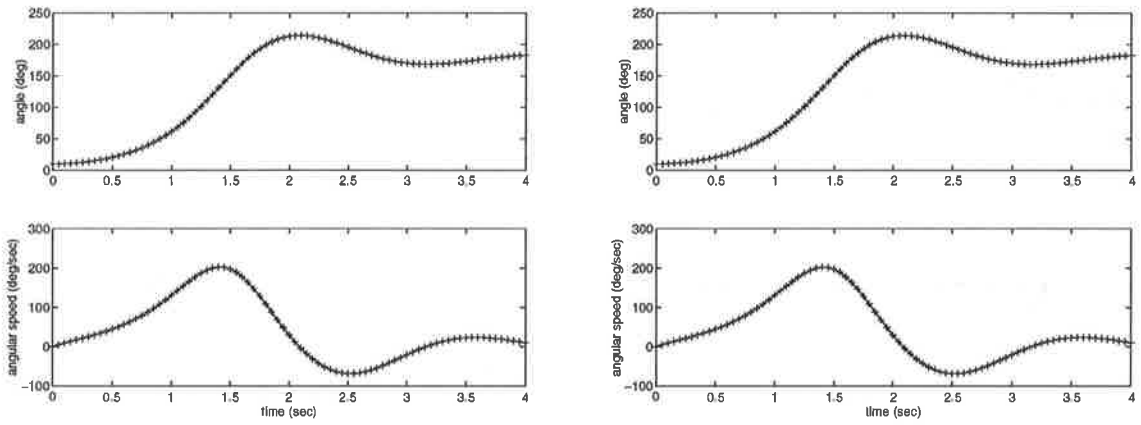


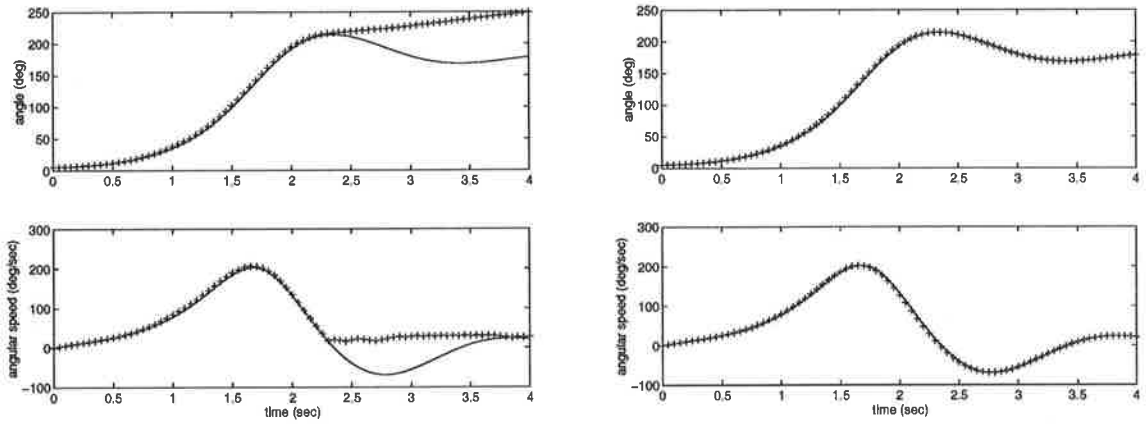
Figure 3.16: Untrained predictive estimation 1: (a) ART2; (b) FuzzyART.

Simulation III

This simulation is similar to Simulation II with the intention to demonstrate what could happen if the input was significantly different from the training data. The simulation also highlights that open-loop estimation reduces the effectiveness of the scheme, as errors accumulate over successive estimation iterations, finally inducing an incorrect estimation. Thereafter, due to the open-loop nature of the estimation scheme, the system becomes out of control. The result is shown in Figure 3.17.

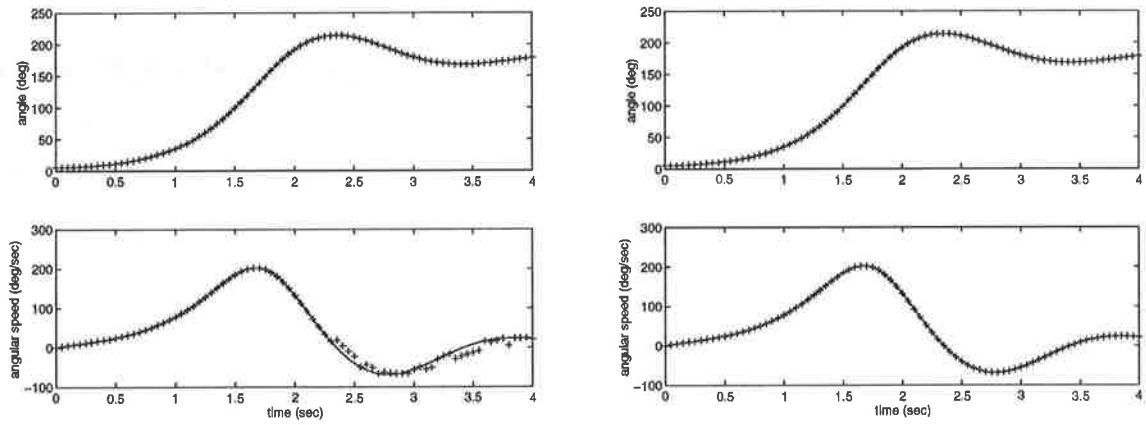
Simulation IV

This last simulation illustrates how the estimation scheme can be improved, especially for the ART2 network, by converting it into a closed-loop scheme, thus avoiding any accumulations of error. It can be seen from Figure 3.18(a) that an incorrectly estimated state will not lead to outright failure of the system. In practice, one step ahead prediction may not be sufficient, in which case the closed-loop scheme can be altered to allow for several states to be estimated in each cycle.



(a) (b)

Figure 3.17: Untrained predictive estimation 2: (a) ART2; (b) FuzzyART.



(a) (b)

Figure 3.18: Closed-loop parameter estimation: (a) ART2; (b) FuzzyART.

Discussions

The results in Figures 3.15-3.18 suggest that the FuzzyART network is more suitable for the problem described. Both networks achieve similar results in Simulations I and II, but the results from the ART2 network in Simulations III and IV clearly show the inadequacy of the ART2 network for parameter estimation problems. However this cannot be taken as FuzzyART networks are more effective than ART2 networks in general, the choice of network is dependent upon the nature of the problem concerned.

The superior performance of the FuzzyART network over ART2 can be attributed to the matching mechanism of the respective networks. In ART2, the degree of matching is defined by the angle between the input and cluster vectors, and matching is a multi-dimensional comparison process. In FuzzyART, however, matching is only one dimensional as it is performed on the vector elements.

In the parameter estimation problem, each input vector comprises three elements representing angle, angular velocity and force, so it is appropriate that matching is carried out between corresponding elements, not spatially between two vectors. This is the main reason for the better results generated from the FuzzyART network. It is also important to realise that in applications where the vector elements are not independent quantities but rather parts of an input, for example in an image recognition process where the vector elements represent image pixels, the matching mechanism in ART2 can still recognise a trained image that has been slightly shifted or deformed. This is one area where the vector matching is preferred over the matching in FuzzyART.

This study has shown that ART is applicable to problems other than visual pattern recognition. In particular, ART draws on its ability to perform vector quantisation and clustering to achieve nonlinear mapping.

3.4 Selective Attention Adaptive Resonance Theory

ART is a neural architecture with certain built-in attentional mechanisms, and is specifically developed for categorisation of recognition codes. Although it may be applied to 2D shape-based object recognition, it does not consider many visual conditions under which object recognition may occur. It has been shown [112, 113] that for ART to perform object recognition in the presence of background clutter and occlusion, image segmentation must be performed prior to the recognition phase, but such preprocessing greatly diminishes the role of the neural architecture in object recognition.

Recognising objects from a complex scene is difficult because object parts tend to merge with or be hidden by their surroundings. In human vision, our ability to select a portion of input stimuli for further processing is performed by the attentional system.

Experimental findings from neurophysiology suggest that visual attention has modulatory effects on neuronal signals [123, 131] and top-down mechanisms from memory may selectively favour desired bottom-up stimuli [50, 51]. These findings have led to the suggestion that feedback pathways from higher cortical areas affect the excitability of neurons coding the features of the attended or ignored stimuli. Ullman also suggests top-down feedback connections are directly involved in the direct activation of a lower area [191]. Several experimental studies [129, 172] have produced evidence supporting these theories. It prompted Lozo [113] to suggest that attentional phenomena may be modelled by using top-down feedback pathways to modulate bottom-up signals. That is, top-down feedback signals may be used to selectively process stimuli from a complex scene. Furthermore, Lozo proposed that the top-down pathways achieve attentional modulation by regulating the amount of chemical transmitter flow from synaptic terminals to postsynaptic cells, thereby controlling the net excitation or inhibition available to the postsynaptic cells. Figure 3.19 shows Lozo's proposed modification to ART that would enable it to deal with objects in cluttered images.

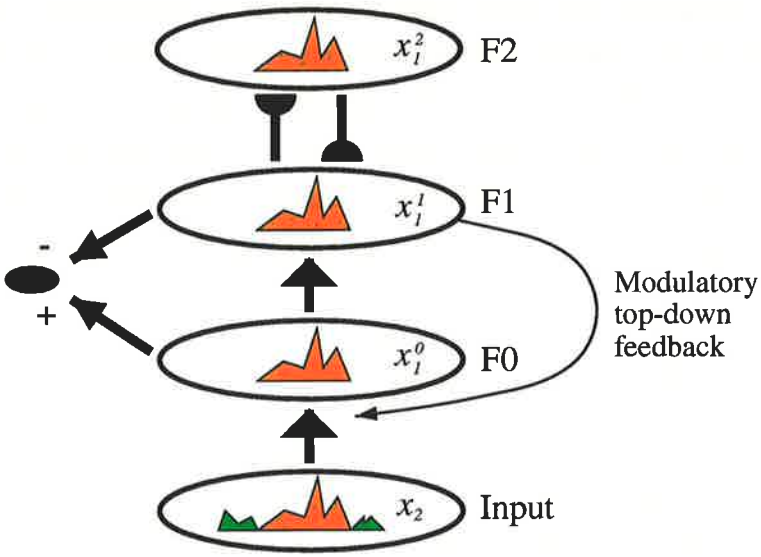


Figure 3.19: Selective attention adaptive resonance theory concept. The top-down feedback can be used to regulate bottom-up signals to achieve attentional modulation. Adapted from [113].

The model shown in Figure 3.19 is based on a generic ART structure derived from ART1 [28], therefore the fields displayed correspond directly to those in ART1. It simply emphasizes the

structural changes required to implement a top-down feedback pathway in an ART system to achieve attentional modulation.

Under a complex and cluttered environment as shown in Figure 3.19, an ART system would reset the choice node in $F2$ after failing to satisfy the matching criterion. The system has failed to recognise the input, even though a recognisable object is contained in the input. For the same input, the model by Lozo has a top-down modulatory feedback from the recalled memory that amplifies the corresponding bottom-up pattern into field $F0$. Furthermore, lateral competition in $F0$ will actively suppress the activity of all cells whose bottom-up signals are not amplified by the top-down memory pathways. This process is referred to as *top-down selective attention.* The interactions between the bottom-up and top-down pathways enable resonance to occur between the recalled memory and a familiar portion of the input. Thus, the new model has been named *selective attention adaptive resonance theory* (SAART).

3.4.1 The SAART Chemical Synapse and Neural Layers

The computational requirements of top-down selective attention in SAART are fulfilled by formal properties of chemical transmitters. Attentional modulation is achieved by regulating the amount of chemical transmitter flow from synaptic terminals to postsynaptic cells. In SAART, this regulation occurs in an idealised chemical synapse model which is an extension of the ART3 synapse introduced in Section 3.3.1. The transmitter dynamics of the SAART synapse are characterised by an external facilitatory signal which can alter the gain (defined as the the amount of mobilised transmitter) of a chemical synapse. A detailed implementation of the SAART synapse with the notation for transmitter dynamics is shown in Figure 3.20.

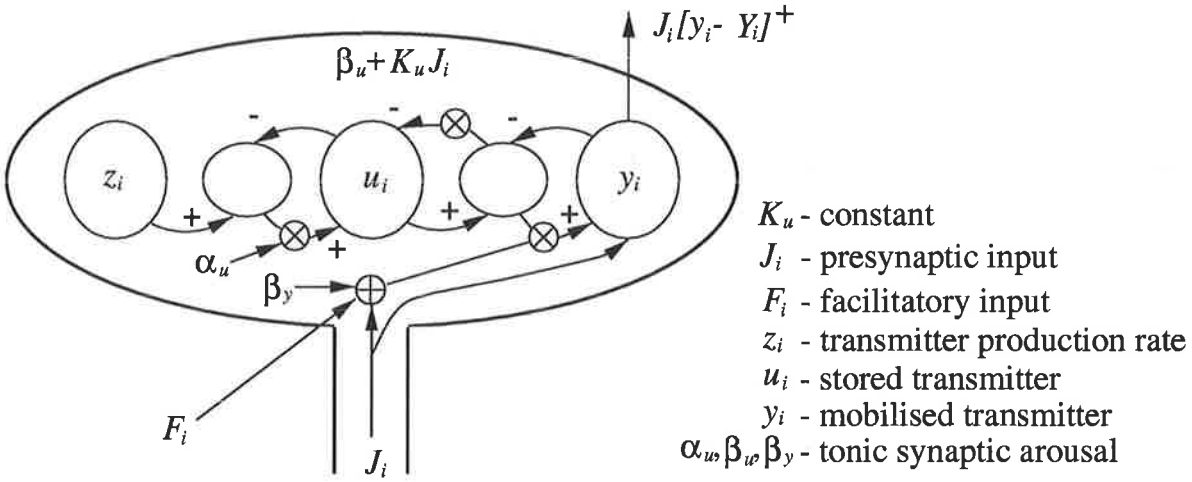


Figure 3.20: Model of the SAART chemical synapse. Adapted from [113].

According to Figure 3.20, there are two input signals to the SAART synapse: the presynaptic input J_i and the facilitatory input F_i, and one output $J_i[y_i - Y_i]^+$, where $[y_i - Y_i]^+ = \max(y_i - Y_i, 0)$ is a simple thresholding function. The internal dynamics are represented by u_i and y_i, which denote the amount of stored and mobilised transmitter, respectively. Mathematically, u_i and y_i can be expressed as two differential equations:

$$\frac{du_i}{dt} = \alpha_u(z_i - u_i) - (\beta_u + K_u J_i)(u_i - y_i) \tag{3.33}$$

and

$$\frac{dy_i}{dt} = (\beta_y + F_i)(u_i - y_i) - J_i\rho_y[y_i - Y]^+ - \gamma_y y_i \tag{3.34}$$

where α_u, β_y and β_u are the tonic activities of the synapse, z_i is the transmitter production rate, and K_u and γ_y are constants.

Equation (3.33) says the rate of change of the amount of chemical transmitter available in the synapse is controlled by its production rate z_i and the amount that has been mobilised already $(u_i - y_i)$. A close examination reveals that F_i in (3.34) acts to speed up the rate of transfer of available transmitter (u_i) to the mobilised state (y_i). Since F_i increases the presynaptic signal J_i, the process has been referred to as *presynaptic facilitation*. More specifically, top-down presynaptic facilitation if F_i is a feedback from top-down, i.e., from recalled memory as in Figure 3.19.

Lozo went on to propose a number of neural layers that are built upon the SAART chemical synapse using the shunting competitive neural architecture [70, 73]. These layers have been termed Presynaptically Modulated Competitive Neural Layers (PMCNLs). PMCNLs are modelled by a set of nonlinear differential equations that represent the dynamics between neurons organised in a shunting competitive fashion that are driven by SAART chemical synapses. Figure 3.21 shows the simplest implementation of a PMCNL.

The set of differential equations that completely describe the PMCNL in Figure 3.21 are given by the synapse equations (3.33) and (3.34), and three additional equations below:

- Postsynaptic cellular activity

$$\frac{dx_i}{dt} = -Ax_i + (B - x_i)Gv_i - (C + x_i)(\bar{G}\bar{v}_i + \Gamma) \tag{3.35}$$

  where A is the passive decay rate, B and C are the saturation limits for the upper and lower bounds respectively; both G and $\bar{G}$ are amplification factors, and Γ is the tonic level of inhibition. This equation represents shunted competition of a layer of neurons with the on-centre off-surround anatomy whose cellular activity is restricted to range $(-C, B)$.

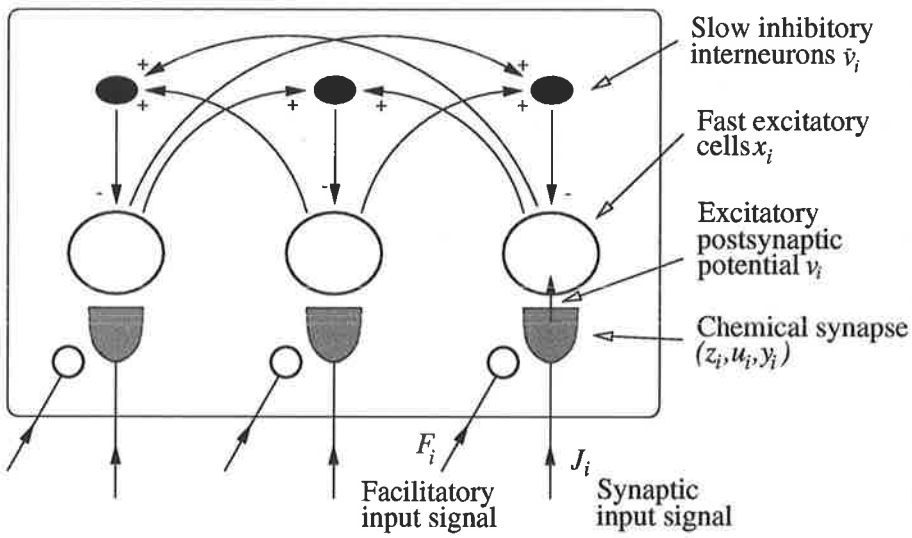


Figure 3.21: Simplest implementation of a presynaptically modulated shunting competitive neural layer. Adapted from [113].

- Excitatory postsynaptic potential

$$\frac{dv_i}{dt} = -Dv_i + J_i[y_i - Y]^+ \tag{3.36}$$

where D and ρ_v are constants, J_i is the input, Y is the threshold for transmitter release, and $[y_i - Y]^+ = \max(y_i - Y, 0)$ is the thresholding function. The excitatory postsynaptic potential acting on a cell is due to the bound transmitter on the postsynaptic cell.

- Lateral feedback inhibition

$$\frac{d\bar{v}_i}{dt} = -\bar{A}\bar{v}_i + \frac{1}{n}\bar{B}\sum_{j\neq i} f(x_j) \tag{3.37}$$

where $\bar{A}$ and $\bar{B}$ are positive constants, n is the number of neurons in a layer. The equation indicates that the postsynaptic cellular activity must be above the threshold ψ before the cell fires, and thus begins charging.

So how may we use SAART, in particular PMCNL, to improve the matching between bottom-up and top-down patterns? How can we use stored models in memory to selectively segment portions of the input scene? And, how do we achieve selective attention in object recognition?

Answers to the above questions lie in the postulate by Lozo [113] on top-down presynaptic facilitation, which states that "a facilitatory presynaptic feedback from the higher neural layer to the lower neural layer mediates the neural mechanism of selective attention". As a result, facilitated presynaptic signals are amplified while nonfacilitated ones are weakened, causing an

increase in the degree of match between bottom-up and top-down patterns. This is in agreement with Broadbent's theory [21] that unattended stimuli are degraded or attenuated as in a PMSCNL. Furthermore if the facilitatory signal is provided by the top-down memory, then we have used a stored model in memory to selectively focus on segments of the input scene. These segments, when grouped together, reminisce of or *resonate with* that stored model.

Figure 3.22 illustrates an application of presynaptic facilitation. The system is implemented using two competitive neural layers. The first layer, labelled as $F0$, holds a steady state pattern of the input. This pattern is gated at $F1$ by modulated chemical synapses. If the modulating signal is pattern specific as shown in Figure 3.22, then cells that correspond to the modulating pattern will be facilitated. Under mutual inhibition, spatial cells that are facilitated will act to suppress the non-facilitated cells, resulting in a filtering effect as shown in the steady state pattern in $F1$.

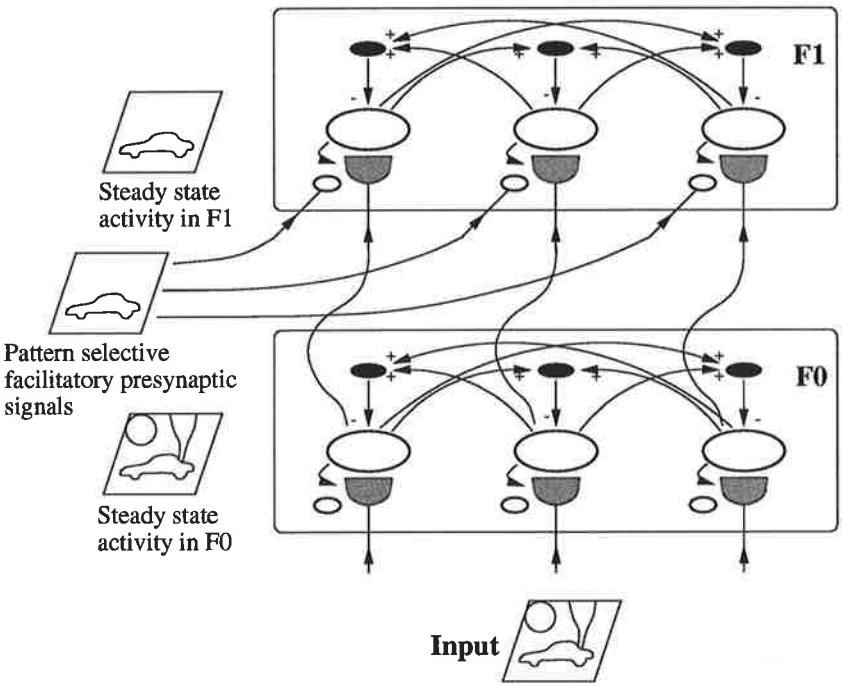


Figure 3.22: Pattern specific presynaptic facilitation of information transfer. A supplementary stage - in which attention is voluntary, applied only when required, akin to "looking harder" to locate something that is familiar but cannot be recognised instantly. Adapted from [113].

Neural Layer Parameter Selection

The PMSCNL consists of a considerable number of parameters. Some of the parameters are chosen according to the input attributes, and once chosen they form the basis for the selection

of the remaining parameters. Since there is not a set of parameters suitable for all conditions, inputs of very different nature may require adjusting the parameters. Tuning system parameters for improved performances is common for neural based systems [9, 111]. On most occasions, this can be achieved empirically from good intuition.

In (3.33) α_u controls the rate of transmitter storage, which must be slower than the dynamics of the postsynaptic potential v_i. So we need $\alpha_u < D$ in (3.36). β_u and K_u are constants for the depletion of the stored transmitter, which is triggered by the input J_i, therefore we have $K_u > \beta_u$. Similarly, in (3.34) we want $F_i > \beta_y$, so that the facilitatory signal has a greater influence over the flow of transmitters. Both ρ_y and γ_y counterbalance the flow, and should be set according to β_u.

The cellular activity x_i in (3.35) is bounded in the range (C, B); for convenience we can set this to $(0, 1)$. Since the dynamics of x_i are much faster than the synaptic dynamics, we have $A \gg \alpha_u, \beta_y$. If we let $A = 1$, then we can set $\alpha_u = 0.05$ and $\beta_u = \beta_y = 0.01$. The gain factor G should be set large enough to excite the layer beyond the threshold for small inputs. Whereas $\bar{G}$ is crucial for increasing competition among nodes. As the amount of competition required varies from case to case, $\bar{G}$ is usually determined empirically. D the decay rate of v_i should be greater than α_u but less than A, so we choose $D = 0.5$. The lateral feedback interneuron $\bar{v}$ in (3.37) is more responsive than the synapse but less than x_i, so we let $\bar{A}$ and $\bar{B}$ equal to one-tenth of A and B.

3.4.2 The SAART Architecture

A SAART neural network architecture is an implementation of the SAART concept in Figure 3.19 using SAART chemical synapses and neural layers introduced in the previous section. The SAART architecture is an extension of ART, therefore SAART is also a real-time and self-organising neural network. However, one significant property of SAART that is not shared by ART is its ability to learn and recognise patterns in a complex and noisy environment through the process of top-down presynaptic facilitation. Figure 3.23 shows one such implementation.

This particular implementation of SAART consists of five presynaptically modulated shunting competitive neural layers (PMSCNLs), which have been labelled as Fields $A1$, $B1$, $B2$, $B3$, and $C1$. These five fields are interconnected via dynamic synaptic pathways whose internal dynamics represent short-term-memory of the most recent neural activity pattern. Field $C1$ is a competitive winner-take-all neural layer, whose nodes represent stored categories. The synaptic connections between Fields $B1$, $B2$, and $C1$ are bottom-up and top-down long-term-memory pathways, respectively. Field $B3$ provides top-down presynaptic facilitation to Field $A1$. An

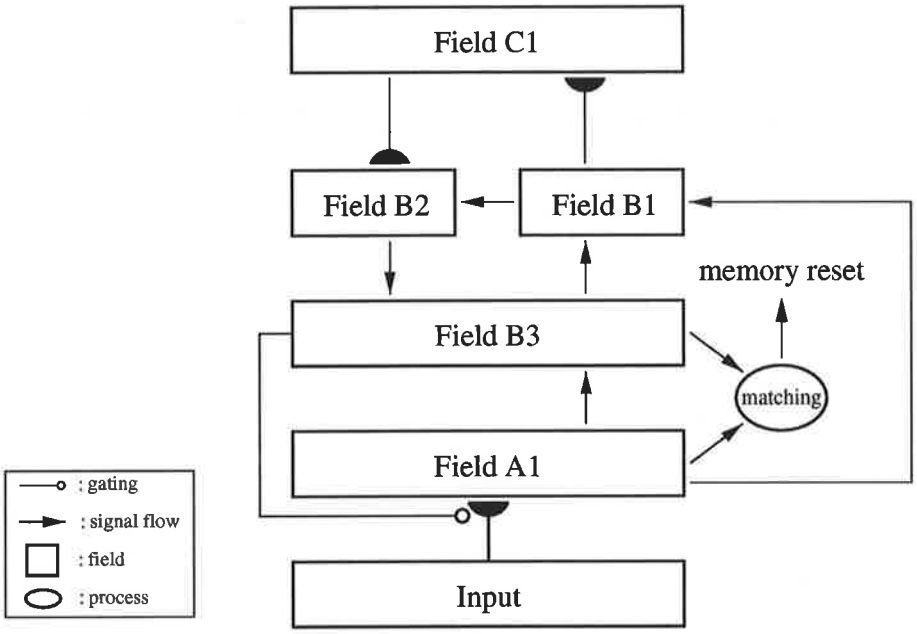


Figure 3.23: A SAART neural network architecture.

input pattern arriving at Field $A1$ may be presynaptically facilitated by top-down signals from Field $B3$, thereby amplifying desirable bottom-up signals from input to Field $A1$. Bottom-up signals that are not amplified by the top-down signals will be actively suppressed by lateral inhibition across Field $A1$. Resonance is established if the STM pattern reverberating between Fields $B1$, $B2$, and $B3$ matches the facilitated pattern in Field $A1$ and the network is in a stable state.

The SAART architecture in Figure 3.23 is one of the many possible implementations. Variations are only limited by the number of fields used and the type of neural layer employed, and their organisation. Fundamental to all SAART implementations is the essential process of top-down presynaptic facilitation.

3.5 Summary

In this chapter, we presented a literature review on the computational aspects of the higher visual function of object recognition. We have considered both the traditional machine vision approach and the biologically inspired approach based on neural networks. The computational power of artificial neural networks has been explored, in particular, a thorough and detailed study of Adaptive Resonance Theory neural networks has been provided, demonstrating many important properties of ART in learning and categorisation.

An extension of ART called Selective Attention Adaptive Resonance Theory, proposed by Lozo [113], was introduced in Section 3.4. This class of neural networks attempt to exploit the computational role of top-down feedback pathways and chemical synapses for selective attention. This theory suggests that top-down feedback signals may be used to selectively process stimuli from a complex scene, and attentional modulation is achieved by regulating the amount of chemical transmitter flow from synaptic terminals to postsynaptic cells, thereby controlling the net excitation or inhibition available to the postsynaptic cells. The most important proposition by Lozo is the SAART process of top-down presynaptic facilitation, which allows the use of top-down recalled memory as an external input to selectively facilitate individual synaptic signals by modulating the gain of the synaptic pathways.

SAART in its current form has limited applicability in terms of visual scene analysis. It can only deal with a single normalised input pattern, however it has provided an important theoretical and mathematical foundation for our research. Both top-down presynaptic facilitation and ART will be utilised in the modelling of our proposed visual scene analysis system in the next chapter.

Chapter 4

Models of Visual Object Recognition

4.1 Introduction and Overview

This chapter presents a neural architectural framework for visual object recognition and selective attention. The proposed model encompasses a two-stage theory of biological vision, namely the parallel preattentive stage and the serial attentive stage. These two stages enable the model to perform automatic attentional shifts and attentional gating, where information can be selected for further processing, or attenuated.

Architecturally, the framework consists of massively parallel feedback and feedforward connections. It is based on a bi-directional structure with both bottom-up and top-down pathways. Bottom-up signals from the input are converted to elementary features and used as visual cues for capturing attention. The captured region is further processed with the eventual goal of memory activation. Top-down signals from memory are used in memory-guided search and recognition. In particular, the recalled memory is used as a feedback to achieve attentional modulation of the bottom-up pathways.

Specifically, the model is capable of detecting, locating, and recognising any familiar objects automatically in a cluttered image. Moreover, it forms a translation, rotation and distortion invariant object representation of the recognised object in an object-based reference frame.

This chapter is structured using an incremental approach. We start from a very basic neural architecture for translation invariant pattern recognition, then additional visual function modules are gradually incorporated into the model to form an integrated neural framework for visual object recognition and selective attention. The additional functions incorporated are recognition in cluttered images and partial occlusion, automatic attentional capture and shift (model of preattentive stage), rotation invariance, and distortion invariance. A graphical illustration of the

integrated framework is given in Section 4.7.

4.2 Translation Invariance

Tolerance to shifts in position for visual object recognition is a widely acknowledged property of the biological visual system. An early study in electrophysiology by Hubel and Wiese [88] indicated that complex cells in the primary visual cortex exhibit approximate invariance to position within a limited range. There is further evidence from a more recent study that the inferotemporal (IT) cortex, an area of the brain that is highly sensitive to complex shapes, is invariant to certain stimulus transformations [90]. Studies in psychophysics [17, 52] have reinforced experimental findings that human visual recognition is translation invariant.

In many neural network based recognition systems [18], the translation invariant feature is not a part of the neural model, rather they rely on pre-processing stages such as feature extraction and segmentation to achieve position invariance. Such models fail to exploit fully the benefits of the massively parallel architecture of neural networks. Furthermore, they make no consideration for synaptic gain control mechanisms which are critical for visual attention [50].

In this section we propose a neural architecture for translation invariant recognition. The model consists of neurons whose synaptic connections can be selectively gated by attention, such that the attended visual information is transferred to an object-centred reference frame that is translation invariant. The model is derived from the translation invariant representation concept of Lozo[113].

4.2.1 Stages of Operation

Figure 4.1 depicts the basic schematic of the proposed neural architecture that is capable of performing translation invariant object recognition. The schematic depicts all the processing fields involved, and the directions of information flow between the fields.

The model consists of five neural representation fields. Starting from the bottom where the visual scene is first registered in the *input field*. A potentially known object is located by aligning partitions of the input field with all stored models in the memory field[1]. The best correlated pair, chosen by a winner-take-all (WTA) neural layer shown as the *spatial alignment WTA field*, are selected as the most likely match. Since every memory is correlated with the input field, there are as many spatial alignment WTA fields as stored memories. Based on the selected pair, the

[1]Models in memory are assumed to have been learned already.

memory field sends a read-out of that particular stored model in memory to the *top-down field*, while the *central representation* receives from the input field a bottom-up translation invariant object representation.

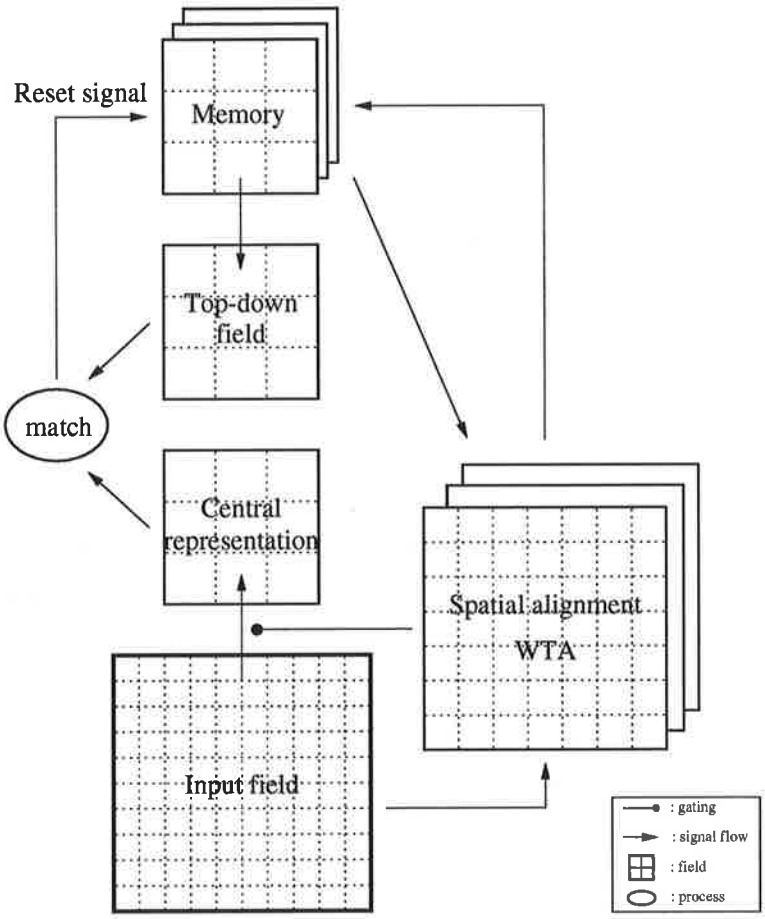


Figure 4.1: Neural architecture for translation invariant object recognition. The model is based on a bi-directional architecture, having both top-down and bottom-up pathways. Note: Long-term-memory (LTM) patterns are depicted as neural activity fields instead of synaptic weights to enable better visualisation.

The system is now ready to match the selected object with the activated memory. The object is deemed to be recognised if the matching is successful, otherwise a reset signal will be triggered, thereby suppressing the currently chosen memory. Therefore, another stored memory may be activated, and the matching process repeats until either the object is recognised, or all the stored memories are exhausted, in which case the object is declared unrecognisable.

Structurally, the fields are arranged in accordance to the bi-directional architecture of visual perception [190], consisting of both bottom-up (BU) and top-down (TD) pathways. These two streams eventually meet and take part in the matching process in an attempt to establish

resonance [28].

There are five main processes for translation invariant pattern recognition. These include:

1. partitioning of the input field;
2. bottom-up activation of a stored memory;
3. selective transfer of the bottom-up and top-down patterns;
4. matching of the bottom-up and top-down pair; and
5. mismatch reset if matching failed.

4.2.2 Partitioning of the Input Field

Consider an input pattern $\mathbf{P}$ stimulating a spatial pattern $\mathbf{X}$, shown in Figure 4.2. This is divided into overlapping sub-patterns $\mathbf{x}_{kl}$ whose individual activities can be described as $x_{ij,kl}$, where ij are local spatial indices for cell activity positions within each sub-pattern, and kl are spatial indices for the sub-pattern positions. These sub-patterns are sampled in parallel, emerging from each sub-region are thresholded synaptic signals $S_{ij,kl}$. Mathematically, the above can be expressed as follows:

$$\begin{aligned} & f : \mathbf{P} \rightarrow \mathbf{X} \\ & \mathbf{x}_{kl} \subset \mathbf{X} \\ & x_{ij,kl} \in \mathbf{x}_{kl}. \end{aligned} \tag{4.1}$$

Synaptic signals $S_{ij,kl}$ arisen from individual cells are given by a threshold function with threshold ψ, shown in (4.2). This is a piece-wise linear function with a discontinuity at ψ. Only cellular activities that are above the threshold may pass:

$$\begin{aligned} S_{ij,kl} &= f(x_{ij,kl}) \\ &= \max(x_{ij,kl} - \psi, 0). \end{aligned} \tag{4.2}$$

4.2.3 Bottom-Up Activation of Stored Memory

After partitioning the input field with overlapping regions, the model attempts to establish relationships between the stimuli in each partition and the stored models in memory. The purpose of this stage, illustrated in Figure 4.3, is to align a potentially familiar object with its possible counterpart in memory. In order to perform this alignment, the region must undergo a translation

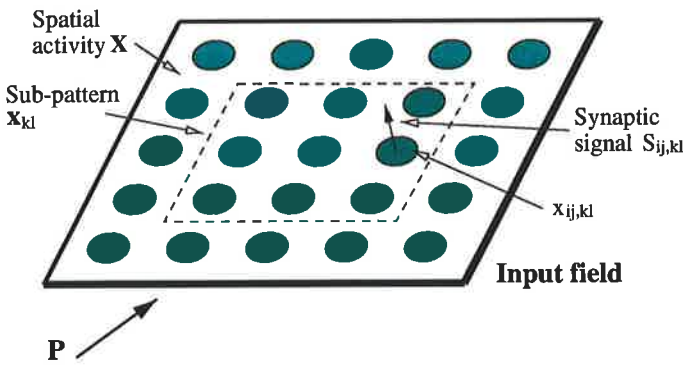


Figure 4.2: Partitioning of the input field. The input field is the neural representation of the visual scene which is processed in parallel in a number of overlapping sub-patterns.

invariant transformation. It determines the best correlated stored model with the transformed region, thereby generating an invariant object representation. The mechanism that determines the best correlated pair is a multi-dimensional WTA, termed as the spatial alignment WTA field, whose function is to localise the most active pair of stored memory and bottom-up pattern.

The synaptic signals are gated by long-term-memory (LTM) weight vectors, $M_{ij,r}$, resulting in postsynaptic signals:

$$G_{kl,r} = \sum_i \sum_j M_{ij,r} S_{ij,kl} \tag{4.3}$$

where the subscript, r, denotes the rth stored model in memory. From Figure 4.3, $G_{kl,r}$ become the input signals to the spatial alignment WTA field.

Equation (4.3) shows that $G_{kl,r}$ is a measure of correlation between patterns $M_{ij,r}$ and $S_{ij,kl}$. The strength of $G_{kl,r}$ is an indication of how well the sub-pattern from location (k, l) in the input field is matched to the rth LTM vector. The signals $G_{kl,r}$ converge at the spatial alignment WTA field and compete against each other to produce a unique winner. The winning node indicates that the region (k, l) from the input field and the rth stored model in memory are a likely match.

In its simplest form, a WTA network is equivalent to a maxima-finding operator, and is given by

$$y_{kl} = \begin{cases} 0 & \text{if } y_{kl} < \max\limits_{pq}\{y_{pq}\} \\ 1 & \text{if } y_{kl} = \max\limits_{pq}\{y_{pq}\}. \end{cases} \tag{4.4}$$

A parallel implementation of the WTA network can be achieved using the shunting competitive neural layer [31], however alternative parallel implementations also exist [58, 98].

In general, a shunting competitive layer with cellular activity x_i, fluctuating within the finite interval $[-C_i, B_i]$, stimulated by excitatory and inhibitory inputs I_i and J_i, and nonlinear feedback

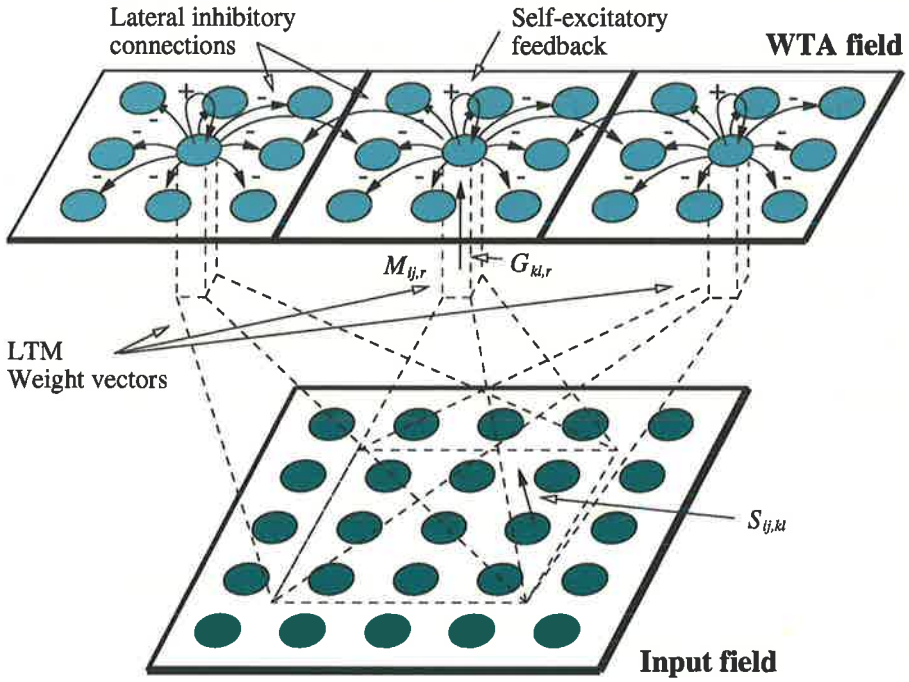


Figure 4.3: Bottom-up activation of stored memory. Each sub-pattern from the input field is gated by all LTM weights, and the resultant synaptic signals then engage in a WTA competition to identify a likely bottom-up and top-down match, with the winning cell representing the spatial location of the bottom-up pattern and the location of the top-down pattern from memory.

signals $f_i(x_i)$ and $h_j(x_j)$, can be expressed as

$$\frac{dx_i}{dt} = -A_i x_i + (B_i - x_i)[I_i + f_i(x_i)] - (C_i + x_i)[J_i + \sum_{j=1}^{n} D_{ij} h_j(x_j)]. \tag{4.5}$$

where A_i is the passive decay rate.

A close examination of the equation reveals that the main function provided by the shunting competitive layer is contrast enhancement - improving the stronger signals while suppressing the weaker ones. WTA can be regarded as an extreme case of contrast enhancement, where only the strongest survives. This is achieved through strong competition by having very large lateral inhibitions, so that under fierce competition only the cell with the largest activity remains.

It follows that competition in the spatial alignment WTA field is given by

$$\frac{dg_{kl,r}}{dt} = -Ag_{kl,r} + (B - g_{kl,r})[G_{kl,r} + f(g_{kl,r})] - (C + g_{kl,r})D \sum_{uv,p \neq kl,r} f(g_{uv,p}). \tag{4.6}$$

Equation (4.6) can be regarded as a three dimensional competition, with two dimensions, k and l, in spatial competition, and the remaining dimension, r, for determining the best correlated LTM vector. This is graphically illustrated in Figure 4.3.

4.2.4 Selective Transfer of Bottom-Up and Top-Down Patterns

After the multi-dimensional competition, the winning stored model in memory is sent to the top-down field, and its corresponding winning region from the input field is transferred to the central representation - an object-centred invariant reference frame. These two transfers are summarised in Figures 4.4 and 4.5.

The purpose of the bottom-up transfer is to place the attended object into its canonical form in accordance to [143], in which it was suggested the invariant object representation (usually refers to invariance in position, size, and orientation) of a fixated object may be routed to higher cortical areas in a bottom-up manner. Similarly the process in Figure 4.5 is for the read-out of a top-down expectation, triggered by a bottom-up input.

Figure 4.4 shows that individual cellular activities C_{ij} in the central representation are given by

$$C_{ij} = \sum_k \sum_l (S_{ij,kl} \sum_r g_{kl,r}). \tag{4.7}$$

Similarly, activities T_{ij} in the top-down field are

$$T_{ij} = \sum_r (M_{ij,r} \sum_k \sum_l g_{kl,r}). \tag{4.8}$$

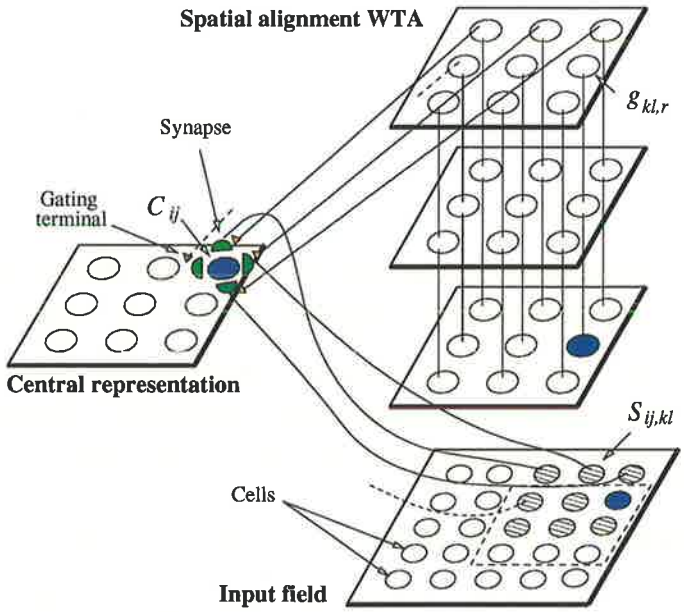


Figure 4.4: Selective transfer of bottom-up pattern from input field. Each cell in the central representation is connected to a neighbourhood of cells in the input field. In addition, each synaptic pathway is connected to a cell in the spatial alignment WTA field via a gating terminal. The gating terminals act as switches, allowing signals through only when the gating signals are on. Under WTA only one cell remains active, and its corresponding region is transferred to the central representation.

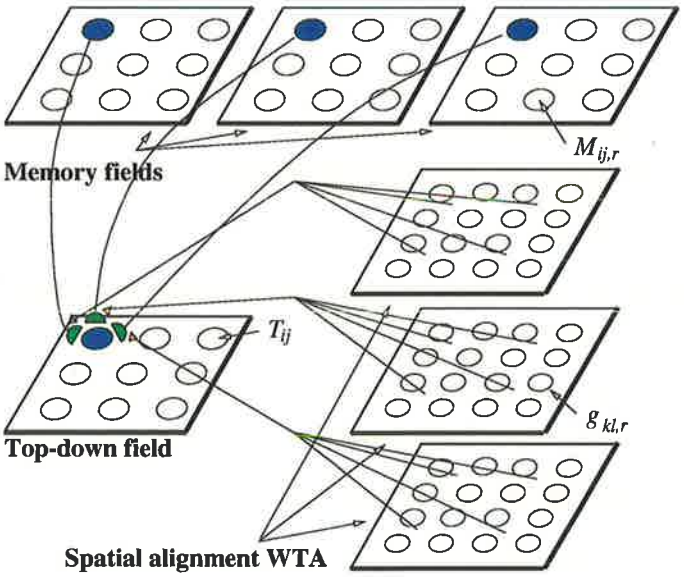


Figure 4.5: Selective transfer of top-down pattern from memory. A complementary process to the bottom-up transfer. Here the selected model in memory is transferred to the top-down field for matching with the bottom-up pattern. For better visualisation, LTM patterns are represented as neural activity fields instead of synaptic weights. This, however, has no bearing on the mathematical model.

4.2.5 Matching of Bottom-Up and Top-Down Patterns

The mechanism employed for matching is similar to that used in ART based neural networks [29, 30]. The degree of match between the bottom-up and top-down patterns is determined by the vector $\mathbf{r}$, such that each individual element

$$r_{ij} = \frac{C_{ij} + T_{ij}}{\|\mathbf{C}\| + \|\mathbf{T}\| + \epsilon}, \quad \epsilon > 0 \tag{4.9}$$

where $\|V\|$ is the L_2-norm of a vector V and ϵ is a non-zero constant, installed in order to avoid zero division if both $\|\mathbf{C}\|$ and $\|\mathbf{T}\|$ become zeros.

4.2.6 Mismatch Reset

The matching is considered a success if the degree of match is greater than a preset value, called the vigilance parameter ρ:

$$\frac{\rho}{\epsilon + \|r\|} < 1. \tag{4.10}$$

However, if the selected memory fails to match the bottom-up input, a reset signal is triggered causing a temporary suppression of the selected memory location in a process similar to the mismatch reset in ART3 [30]. This reset allows another memory to be selected and compared with the input.

4.3 Recognition in Cluttered Images

Single, well defined objects can be recognised by a variety of methods and approaches. This area of object recognition has been well researched and a wealth of literature exists [8, 39, 28, 29, 122, 191]. However, real-world images are usually rich in context and high in detail. Objects contained within real-world images are rarely seen in isolation, they are usually seen against some background, with other objects next to, or in partial occlusion to, them. Recognising objects in cluttered scenes is a very challenging problem.

The traditional approach deals with complex scenes by performing a process called *segmentation* or *figure-ground separation* [6, 77, 191], whereby a well defined region that contains an object to be recognised is separated from other objects and background clutter. The isolated region can then be recognised using any desired method.

Segmentation is arguably the most difficult and crucial stage in a traditional object recognition system. Generally, this stage determines the success of the entire system. Failures in segmen-

tation are usually caused by vaguely defined object contours. There are many instances where ill-defined contours may occur, the most commonly encountered ones are low image resolution, noise corruption, complex background, occlusion, bad illumination and low contrast scenes.

This section presents an extension to the translation invariant model proposed in the previous section. The new model incorporates an additional capability for recognising familiar objects in cluttered environment. The model is based on Selective Attention Adaptive Resonance Theory, or SAART [113, 116], introduced in Section 3.4. SAART is a self-organised real-time learning, memory-guided search neural network for recognition in cluttered images. SAART is incorporated into the model in the form of a neural layer modelling the effect of selective attention.

The underlying principle and mechanisms which enable SAART to perform recognition in cluttered images is the use of top-down feedback connections as a control signal to bottom-up synaptic signals for attentional modulation or gating. In this way, memory or prior knowledge can be used to selectively retune the signal transmission gains and the filtering characteristics of the lower neural layers. This allows desired synaptic signals that resemble a familiar object to be strengthened, while the remaining bottom-up synaptic signals from the background and other objects are weakened, or in the extreme case, totally suppressed under shunting mutual competition [31].

The extended model for translation invariant object recognition in cluttered images is shown in Figure 4.6. In comparison with Figure 4.1, an additional processing layer called the *selective attention field* is incorporated for its ability to selectively process stimuli that correspond to a familiar object. The layer is constructed upon chemical synapse models [113], consisting of neurotransmitters, excitatory and inhibitory neural connections, and facilitatory connections, modelled entirely using differential equations. In Figure 4.6, synaptic signals from the central representation are gated by a feedback pathway from the memory field to perform *presynaptic facilitation* (a major interaction in SAART). The synaptic strength of each neural pathway is dynamically governed by a gain control mechanism, so that the neural signal from each spatial location can be amplified or attenuated as desired (discussed in detail in Section 3.4). This has the effect of attending to familiar object stimuli while ignoring any other background stimuli.

Besides the extra neural layer and top-down feedback pathway, the main difference between this model and the one in Section 4.2 is a more sophisticated processing unit at the heart of the model. We have employed a structure similar to ART2 [29], for short-term-memory (STM) reverberation, normalisation and pattern matching. In order to make minimal modifications to ART2, where appropriate we have kept the same notations and STM equations as given in [29].

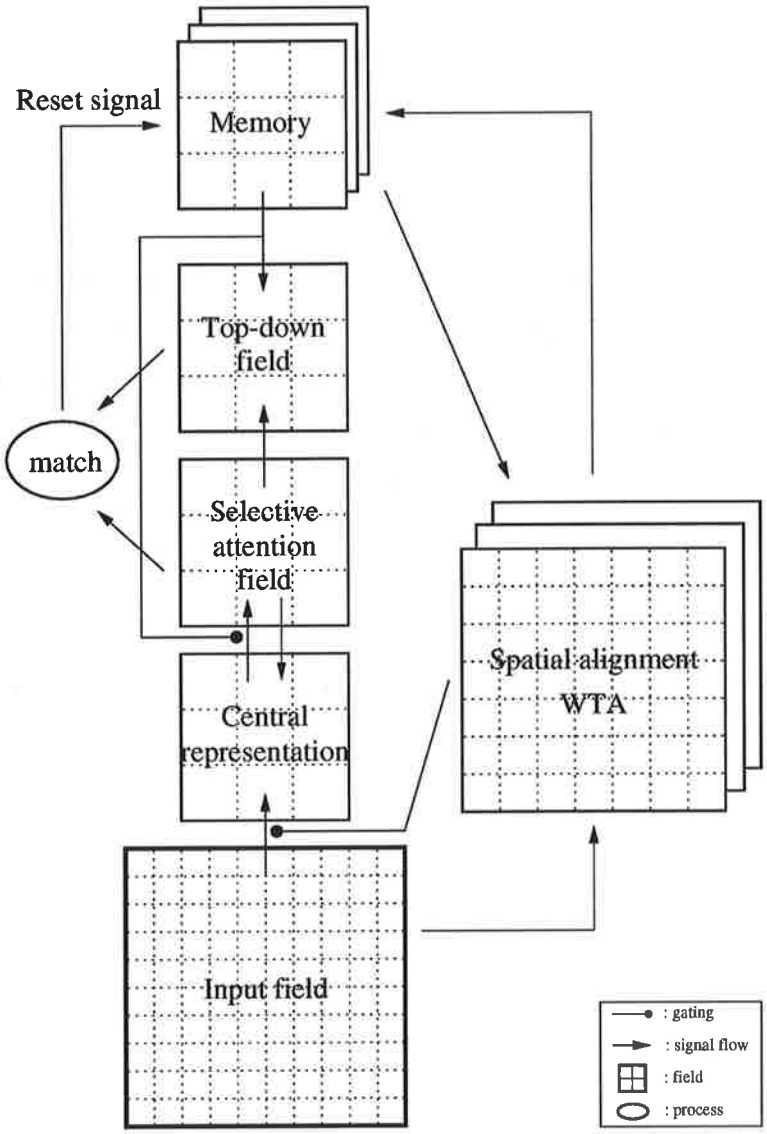


Figure 4.6: Neural architecture for translation invariant object recognition in cluttered images. Small filled circles are gating terminals.

4.3.1 Stages of Operation

The framework in Figure 4.6 operates in the same way as the one in Figure 4.1 up to the selective transfer of the bottom-up and top-down patterns. After that, the top-down and bottom-up patterns are processed in a number of STM loops, spanning over several processing fields, which has the effects of normalisation, contrast enhancement, noise suppression and self-stabilisation. After several iterations, the STM patterns are stabilised and ready for matching. Two courses of action may be taken after a matching failure. The first is when the degree of match is unacceptably low, then a mismatch reset process is initiated. The second is when the matching fails marginally, i.e., the degree of match is sufficiently high, which is often due to the object being occluded by other objects or is seen against some background. To handle such problems, the aligned object region is further processed in the selective attention field which utilises learned 2D object patterns in memory as a feedback signal to achieve top-down segmentation. For selective attention to be initiated, we propose there is a *secondary vigilance parameter* that controls the activation of the presynaptic facilitation process. The original vigilance parameter may be called the primary vigilance parameter. Whenever the degree of match is above the secondary vigilance but below the primary, we can interpret this as the input object strongly resembles the activated top-down memory pattern but inconclusive.

In addition to the processes discussed in Section 4.2, two more processes must be performed in order to achieve translation invariant object recognition in cluttered images. The overall operation can be summarised as follows:

1. partitioning of the input field;
2. bottom-up activation of a stored memory;
3. selective transfer of the bottom-up and top-down patterns;
4. short-term-memory iteration;
5. matching of the bottom-up and top-down pair; and
6. choice after matching:
    - mismatch reset or
    - top-down presynaptic facilitation, and back to 4.

In the next two sections, we will discuss the details of incorporating ART2 and presynaptic facilitation into our model.

4.3.2 Incorporation of Adaptive Resonance Theory

An ART2-like neural structure, shown in Figure 4.7, is incorporated into the framework for STM reverberation, normalisation and pattern matching. To begin with, we ignore the details of the selective attention field and top-down feedback pathway, and simply incorporate ART2 into the translation invariant model. The layer named "selective attention field" in Figure 4.7 is then modified using differential equations in the next section.

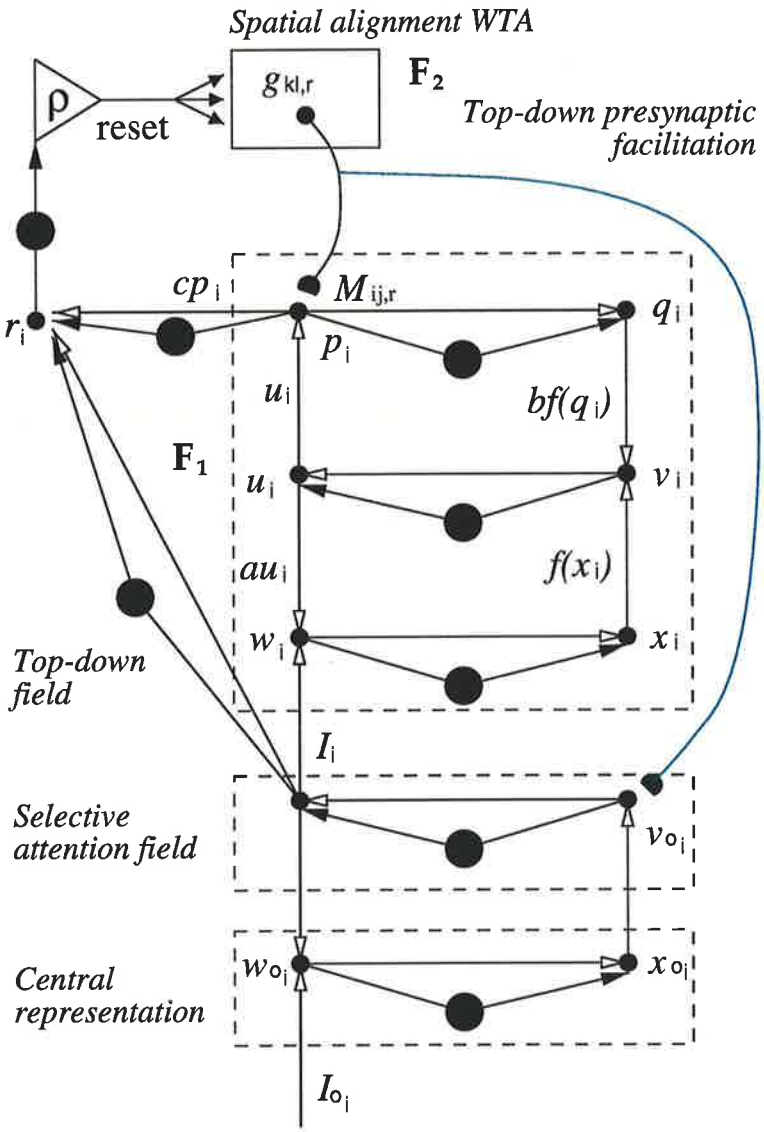


Figure 4.7: STM and matching of top-down memory and bottom-up patterns. An ART2 architecture is incorporated in the heart of the framework to provide STM patterns, as well as the extension for top-down presynaptic facilitation.

For the incorporation of ART2, we have treated the top three STM layers or field F_1 in ART2 as the top-down field; the next STM layer as the selective attention field; the bottom as the

central representation, as indicated by the dash-lined boxes in Figure 4.7. The spatial alignment WTA field may also be regarded as an extended version of the category representation field F_2 in ART2; both are responsible for making a choice for a top-down LTM pattern, however in the former case it has the added function of spatial alignment.

Since the top-down field and central representation are the top and bottom STM layers, T_{ij} and C_{ij} are equivalent to p_i and w_{o_i}:

$$f_{2\to1}(T_{ij}) = p_i \text{ and } f_{2\to1}(C_{ij}) = w_{o_i}, \tag{4.11}$$

where $f_{2\to1}$ is a transformation, mapping a 2D vector to a 1D vector.

This gives (4.12) and (4.13):

$$p_i = f_{2\to1}(\sum_r (M_{ij,r} \sum_k \sum_l g_{kl,r})) + u_i, \tag{4.12}$$

$$w_{o_i} = f_{2\to1}(\sum_k \sum_l (S_{ij,kl} \sum_r g_{kl,r})) + I_{o_i}. \tag{4.13}$$

Besides (4.12) and (4.13), the rest of the STM equations are in the same form as in ART2 [29], and are given below:

$$q_i = \frac{p_i}{\epsilon + ||p||} \tag{4.14}$$

$$u_i = \frac{v_i}{\epsilon + ||v||} \tag{4.15}$$

$$v_i = f(x_i) + bf(q_i) \tag{4.16}$$

$$w_i = I_i + au_i \tag{4.17}$$

$$x_i = \frac{w_i}{\epsilon + ||w||} \tag{4.18}$$

where $||V||$ is the L_2-norm of a vector V, ϵ is a non-zero constant, and f is a nonlinear function with a threshold θ and is given by

$$f(x) = \begin{cases} 0 & \text{if } 0 \leq x < \theta \\ x & \text{if } x \geq \theta. \end{cases} \tag{4.19}$$

Equations for the STM variables with a subscript o are similar to their counterpart in field F_1, except for v_{o_i} which is without its second term in accordance to Figure 4.7, i.e., $v_{o_i} = f(x_{o_i})$.

Once stable STM patterns are established across all layers, the system proceeds to compare or match the patterns from top-down and bottom-up, representing internal memory and external input, respectively. The degree of match is determined by the reset vector $\mathbf{r}$:

$$r_i = \frac{I_i + cp_i}{||\mathbf{I}|| + ||c\mathbf{p}|| + \epsilon}, \quad \epsilon > 0. \tag{4.20}$$

The matching process is considered a success if the degree of match is greater than the vigilance parameter ρ, thus satisfying the constraint (4.10).

If the constraint cannot be met, either a mismatch reset is initiated or a top-down presynaptic facilitation process is required. The former results in the currently active choice node being temporarily suppressed, thus allowing another category (stored model in memory) to be chosen. The latter assumes that a familiar object is present but successful matching is only hindered by cluttered background or partial occlusion. For this to happen, some ART2 equations may be replaced by a set of dynamic differential equations that constitute a selective attention neural layer. Details of this process will be discussed in the next section.

4.3.3 Implementation of Selective Attention

An essential process of visual attention is attentional gating whereby some input stimuli are selected for further processing, while the rest are ignored or attenuated [2, 131]. The effect of attentional gating can be either spatial or object-based [94]. The former requires spatial cues for priming purposes, whereas the latter is responsible for perceptual grouping.

We showed in Section 3.4 that perceptual grouping could be modelled using top-down presynaptic facilitation in a Presynaptically Modulated Shunting Competitive Neural Layer (PMSCNL) [113]. To incorporate the process of presynaptic facilitation, we simply convert node v_{o_i} in Figure 4.7 into a PMSCNL by modelling the five equations given in (3.33)-(3.37). Therefore, x_{o_i} becomes the synaptic input signal J_i in PMSCNL, the top-down signal T_{ij} (without the STM term) in (4.8) as the facilitatory signal F_i, and the resultant cellular activity x_i in PMSCNL, after normalisation, as I_i in Figure 4.7. The whole process continues on as usual in iterating STM loops, except I_i is a presynaptically facilitated pattern of x_{o_i}. If after presynaptic facilitation the patterns still fail to match, a mismatch reset is triggered, causing the currently selected LTM pattern and STM equations to be reset for a new category. The entire process is repeated until either a match is found or all stored models in LTM are exhausted, in which case the pattern is declared unrecognisable. For easy reference, the five equations are listed here:

Postsynaptic cellular activity

$$\frac{dx_i}{dt} = -Ax_i + (B - x_i)Gv_i - (C + x_i)(\bar{G}\bar{v}_i + \Gamma) \tag{4.21}$$

where A is the passive decay rate, B and C are the saturation limits for the upper and lower bounds respectively; both G and $\bar{G}$ are amplification factors, and Γ is the tonic level of inhibition. This equation represents shunted competition of a layer of neurons with the on-centre off-surround anatomy whose cellular activity is restricted to range $(-C, B)$.

Excitatory postsynaptic potential

$$\frac{dv_i}{dt} = -Dv_i + J_i[y_i - Y]^+ \tag{4.22}$$

where D is a constant, J_i is the input, Y is the threshold for transmitter release, and $[y_i - Y]^+ = \max(y_i - Y, 0)$ is the threshold function. The excitatory postsynaptic potential acting on a cell is due to the bound transmitter on the postsynaptic cell.

Lateral feedback inhibition

$$\frac{d\bar{v}_i}{dt} = -\bar{A}\bar{v}_i + \frac{1}{n}\bar{B}\sum_{j\neq i} f(x_j) \tag{4.23}$$

where $\bar{A}$ and $\bar{B}$ are positive constants, n is the number of neurons in a layer. This equation indicates that the postsynaptic cellular activity must be above the threshold ψ before the cell fires, and thus begins charging.

Stored transmitter

$$\frac{du_i}{dt} = \alpha_u(z_i - u_i) - (\beta_u + K_u J_i)(u_i - y_i) \tag{4.24}$$

where α_u and β_u are tonic adaptation constants, z_i is the transmitter production rate, and K_u is a constant.

Mobilized transmitter

$$\frac{dy_i}{dt} = (\beta_y + F_i)(u_i - y_i) - J_i\rho_y[y_i - Y]^+ - \gamma_y y_i \tag{4.25}$$

where β_y is the tonic activity of the synapse, F_i is the facilitatory signal, and ρ_y and γ_y are constants.

4.4 Preattentive Processing: Deploying Automatic Attention

Attention is necessary in visual perception due to the limited information processing capacity of the brain [22, 128, 139]. Visual attention must maximise our speed of response, yet maintain an acceptable level of analysis. Somehow, it must find a compromise between these two contrasting and competing requirements. For this reason that a number of psychophysical studies have suggested a two-stage theory of human visual perception [139, 187]. The first stage is the *parallel-preattentive* mode, in which the visual scene is rapidly processed to identify important portions to be processed thoroughly by the second stage, the *serial-attentive* mode. The existence of two distinguishable modes during attentional processes are supported experimentally [118, 121, 198].

Two influential theories of visual attention [186, 97], that are based on the two-stage model, have suggested that the preattentive mode locates regions of interest from the visual scene by processing simple features such as colour, direction of motion, orientation of edges, and luminance contrast rapidly in parallel. These regions are then processed by a spatially limited serial stage. Experimental results [187] on visual search tasks involving single feature and multi-feature targets have supported the theory. It was found that response times for single feature searches were independent of the set size, and in contrast, response times for conjunction searches increased with the set size.

Another important issue in preattentive processing is whether these elementary features capture attention in a bottom-up (stimulus-driven) manner. This has been referred to as *feature singleton attentional capture* [54]. Bacon and Egeth [7] propose that singleton attentional capture may occur under *singleton detection mode*, in which attention is directed to the location exhibiting the largest local feature contrast.

In the previous section, we have implemented the serial attentive mode. Here we model the effect of preattentive processing. The model uses luminance contrast as an elementary feature for attentional capture. As a result, a region of interest is located based on the strength of its luminance contrast. The region is then transferred via *selective mapping* [98] to a higher cortical area for a thorough analysis, as suggested in the serial-attentive mode.

With this latest incorporation, the model is enhanced in many ways. Firstly, it will have a more theoretically sound parallel-serial architecture. Secondly, it improves the efficiency of processing with a parallel front end. Lastly, the model becomes more biologically plausible due to the fact that translation invariance in the visual system is very limited in range [52]. To recognise objects away from the fixation point, eye movements are necessary. These are carried in the superior colliculus, driven by some attentional mechanism. By modelling the preattentive mode, we have shifted the translation invariant transformation to a higher cortical layer which is limited in size. Also the signals resulting from singleton attentional capture may be used to drive the attentional mechanism for eye movements. Both of these agree with experimental observations [52, 54].

4.4.1 Stages of Operation

Two noticeable changes present in the framework for modelling the effects of the preattentive mode, shown in Figure 4.8, are two additional neural processing fields: the *high activity WTA field* and *high activity field*. More subtle changes are also made in neural connections, for example, a new pathway is directed towards the input field; the central representation receives

its input from another processing field, etc.

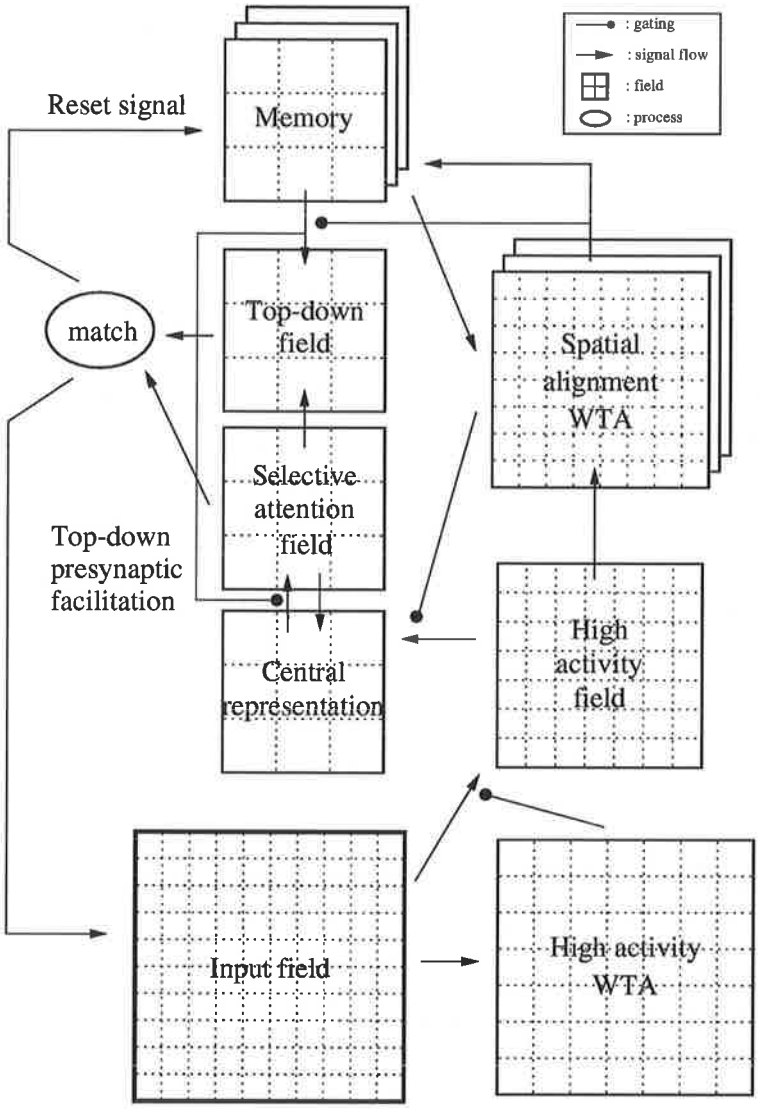


Figure 4.8: Neural architecture for object recognition with visual selective attention. Small filled circles are gating terminals.

As their names suggest, the new processing fields are involved in detecting, locating and selecting a region of interest based on the level of activity of some feature contained within that region. The input field is partitioned into a number of overlapping and equal size regions. These regions are sampled in parallel by a 2D Gaussian receptive field for spatial averaging - finding the average spatial activity across a well defined region using a weighted sum approach, so that each average value represents the overall activity level of a region. Each region is represented by a cell in the high activity WTA field, such that each cell receives the average spatial activity value of its corresponding region as input. Under mutual WTA competition, a winner cell is

chosen corresponding to the region with the maximum level of spatial activity. Thus, a region of interest is located based on a local feature contrast. Since our model is designed for object recognition in 2D gray scale images, we have chosen luminance contrast as the elementary feature measured.

The selected region is projected to the high activity field via synaptic gating, in turn the high activity field holds the selected region which plays a similar role to the input field in the previous model. That is, one may regard the model in Figure 4.8 the same as the one in Figure 4.6 if the bottom two processing fields in Figure 4.8 were removed. The reason is that the model in Figure 4.6 assumes the translation invariance property is applicable throughout, but it turns out that translation invariance is limited in range and is only processed in the second stage of attention. Whereas the model proposed in this section takes into account the effects of attentional capture.

It should be noted that attentional capture can only occur if the level of neural activity stimulated by the visual scene is above some activation level which we simply refer to as the activation threshold.

Another function performed by the preattentive mode is attentional shift. Once a region of interest has been analysed another region may capture attention, thereby shifting the focus. In our model, attentional shifts are based on the degree of conspicuity, measured in terms of an elementary feature contrast, so that shifts are carried out in descending order from the most conspicuous region to the least one. Koch and Ullman [98] propose that there are two rules governing shifts in selective visual attention: *proximity* and *similarity* with the presently selected location. These are discussed in Section 4.4.4.

In order for attentional shift to function properly, there are time lapses in which previously selected locations cannot capture attention again until the lapses are over. In our model we simply assume that the model operates within these time lapses, so that after a region has been selected, it is precluded from selection. This is carried out by means of an inhibitory Gaussian receptive field which acts to suppress any selected locations, thereby preventing them from winning the WTA competition in the high activity WTA field.

From the above, several more stages are now involved in the model for the inclusion of preattentive processing. The overall processing can be summarised as follows:

1. attentional capture - high activity region selection based on elementary features and WTA competition for locating the most active cell, provided the minimum activity level is above the activation threshold;
2. selective transfer of high activity region;

3. bottom-up memory activation and translation invariant transformation;
4. selective transfer of bottom-up and top-down patterns;
5. short-term-memory iteration, matching of the bottom-up and top-down pair; and
6. choice after matching:
   a) automatic attentional shift, details of this stage are discussed in Section 4.4.5, and back to stage 1,
   b) mismatch reset and back to stage 3, or
   c) top-down selective attention and back to stage 5.

In the following sections, stages 1, 2 and 6a) will be discussed in details, while the remaining stages are the same as presented in previous sections.

4.4.2 Attentional Capture: High Activity Region Selection

Upon encountering a visual scene, one is inevitably attracted to the most "eye-catching" region, by which we mean a region having the highest contrast to the rest of the scene. This contrast may be formed by simple elementary features such as colour, direction of motion, orientation of edges, and luminance. One reason for singling out this region is for the visual system to allocate its resources to the "most" important part of the visual scene first. Thereafter one's attention shifts to the next most conspicuous region. The purpose of this processing stage is to locate this "eye-catching" region.

Consider Figure 4.9, where a partitioning operation is required, and therefore equations (4.1) and (4.2) apply. Converging at the high activity WTA layer are postsynaptic activities Y_{kl}, given in (4.26). This has the effect of space averaging a region, with respect to its centroid:

$$Y_{kl} = \sum_i \sum_j S_{ij,kl} W_{ij}. \tag{4.26}$$

The Gaussian receptive W_{ij} is given by

$$W_{ij} = W_o \exp\left(-\frac{(i_o - i)^2 + (j_o - j)^2}{\sigma^2}\right) \tag{4.27}$$

where W_o is a constant, (i_o, j_o) is the centroid of the sub-pattern, and σ is the standard deviation.

In the high activity WTA field, signals y_{kl} compete against each other to produce a unique winner, which indicates that its corresponding region has the highest level of activity based on some elementary feature.

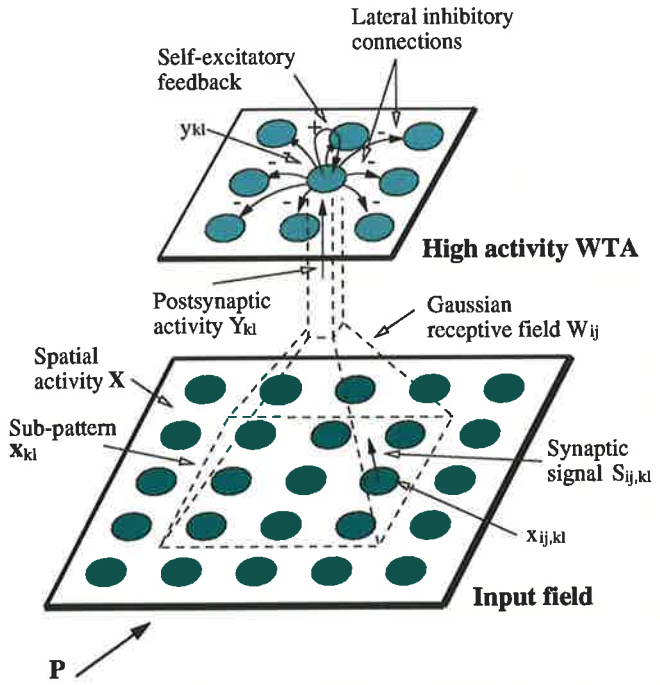


Figure 4.9: High activity region selection. Overlapping regions are gated by a Gaussian receptive field. The resultant cellular activities participate in a WTA competition in a shunting competitive neural layer to locate the region with the highest level of contrast.

Thus, a shunting competitive neural layer implementation of the high activity WTA field is given by

$$\frac{dy_{kl}}{dt} = -Ay_{kl} + (B - y_{kl})[Y_{kl} + f(y_{kl})] - (C + y_{kl})D \sum_{i,j \neq k,l} f(y_{ij}) \tag{4.28}$$

where $f(y_{kl}) = \max(y_{kl} - \psi, 0)$ is a threshold function with threshold ψ, and A, B, C and D are constants.

Before competition individual activities are bounded, such that $y_{kl} \in [0, 1]$. After competition the winner cell is quantised to one, so that (4.4) is satisfied.

Figure 4.9 is very similar to the competitive learning module of Grossberg [31], which is constructed upon two separate model architectures: the instar and the shunting competitive networks. The competitive learning module can be expressed as shown in Figure 4.10, where an input pattern $\mathbf{p}$ is gated by a weight vector $(\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_n})$. A unique winner is then produced from the gated signals using shunting competition.

The processing in Figure 4.9 is summarised in Figure 4.11. Here instead of having one single input pattern and many different synaptic weights, we have a set of patterns $(\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_n})$ and a fixed weight $\mathbf{w}$. In a similar way, postsynaptic activities are generated, and from which a winner is chosen.

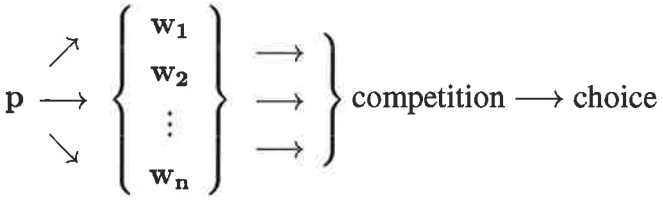


Figure 4.10: The competitive learning module.

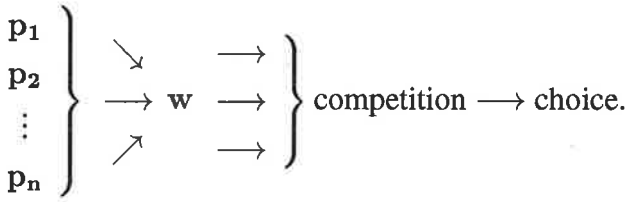


Figure 4.11: Competition in the high activity WTA field.

4.4.3 Selective Transfer of High Activity Region

In this stage, the winner region from the previous stage is transferred from the input field to the high activity field, using a massively parallel set of neural connections, as shown in Figure 4.12.

For each cell in the high activity field, its activity H_{ij} is given by

$$H_{ij} = \sum_k \sum_l y_{kl} S_{ij,kl} \tag{4.29}$$

which describes a selective transfer process of information in parallel.

However, since only one unit in the high activity WTA field remains active after competition, it may be interpreted as a control signal to a digital multiplexer, thus the transferred pattern is

$$H_{ij} = S_{ij,(\max)}. \tag{4.30}$$

The equation boils down to a direct transfer of information in parallel. The selected region is the one with the maximum average spatial activity, and in this case it is the region of highest luminance contrast. This is a gating stage where no signal is allowed through initially, i.e., with all control signals zero indicating no passing or switched off, until a choice is made to selectively pass the chosen region.

After the transfer to the high activity field, the framework operates as in Figure 4.6 until the current region of interest is analysed. At this point, the system must shift its focus to another region of interest.

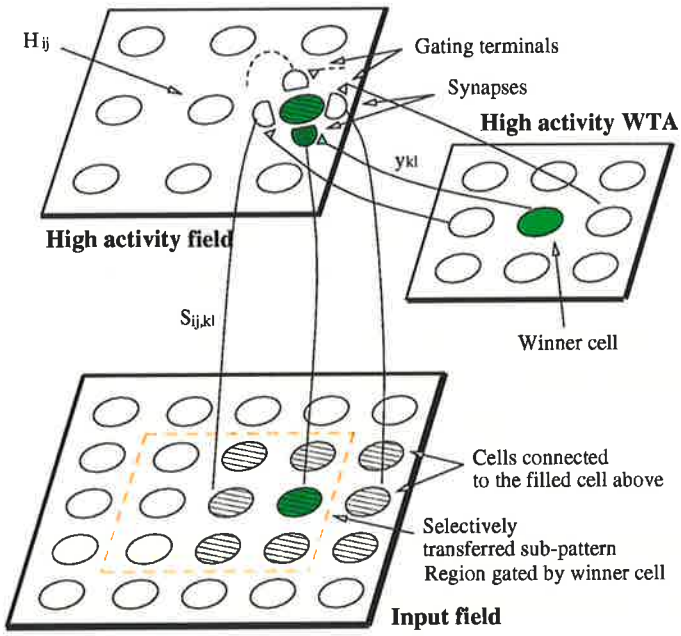


Figure 4.12: Selective transfer of high activity pattern.

4.4.4 Attentional Shift Considerations

Studies in psychophysics have come up with conflicting suggestions regarding how attentional shifts are carried out. At the centre of the debate is whether attentional shifts occur in an analog continuous or discrete spontaneous onset fashion. That is, whether shifts in attention are distance-dependent. Shulman *et al.* [171] and Tsal [188] have provided evidence that it takes some measurable time to shift the focus of attention from one location to another. Furthermore, this time increases with distance between these locations. Contrary to those claims, more recent studies by Kwak *et al.* [103], and Sperling and Weichselgartner [175] have reported that longer movements of attention do not require more time.

Regardless of the exact nature of movement of attention, these findings only serve to provide some inspirations and insight into the implementation of attentional shift across the visual scene. In both cases, attentional movements are performed in descending order of conspicuity, i.e., from the most salient location to the least. For that to happen the currently selected region must be prevented from selection, so that another region of interest may be chosen. Koch and Ullman [98] proposed that the cell representing the currently selected location be allowed to decay, even if constant stimuli are present. They further proposed that this decay may occur locally within the cell or centrally from an external source which inhibits the cell as its conspicuity fades. As a result, the WTA network responds to the changes by shifting to the next most salient location. Based on findings from Shulman *et al.* [171] and Tsal [188], they claimed that the WTA network convergence time is depended on the distance between selected locations.

After selecting this new location, the visual stimuli representing this location are selectively transferred to the central representation as in Section 4.4.3.

According to Koch and Ullman [98], there are two rules for shifting the processing focus, in particular, they concern the relationship between the currently selected location and the next location to be selected. The first rule is called *proximity preference*, in which local salient locations are preferred over global locations of similar level of saliency. The second is called *similarity preference*, where locations that share some common or similar features with the current location are selected ahead of those unrelated locations. For example, if the current processing focus is green in colour, then other green locations will be facilitated, thus enhancing their saliency.

As mentioned above, the distance-dependence theory of attentional shift is at odds with more recent findings, and the conclusiveness of earlier findings by Shulman *et al.* [171] and Tsal [188] have been challenged by Eriksen and Murphy [56], and Yantis [206]. Emerging results are certainly indicating a distance-independent nature of movement of attentional focus. This has led us to suggest that there are two separate issues at hand regarding attentional shift: where to shift, i.e., the location of the next processing focus, and how to shift, i.e., the mechanics of carrying out the shift itself. The former concerns the selection process and has no bearing on how the shift is performed. Selection can be classified as either initial or on-going. The first type is unaffected by internal or external factors, is based purely on local feature contrast, and is thus unbiased. Whereas the on-going type is affected by both bottom-up and top-down factors, which are automatic stimulus driven and voluntary goal-directed, respectively. However, it is not always possible to clearly distinguish top-down or bottom-up attentional capture. It seems that they complement each other and operate in a pair, often attention capture is a result of both top-down and bottom-up factors, with the balance and mixture depending on the situation. We can thus interpret the two rules by Koch and Ullman [98] as external factors which influence the selection process, in particular on-going selection. So preferences are given to neighbourhood regions or features associated with the current location by enhancing their saliency. Selection aside, the shift itself appears to be invariant and discrete in nature, which agrees with the parallel architecture in the early stages of the visual system. Furthermore, Koch and Ullman [98] adopted the distance-dependence theory of attentional shift for supporting their particular implementation of a WTA network, while our implementation using a shunting competitive neural layer is distance-independent. Indeed, if there was any delays found in attentional shift experiments, they would most likely be caused by the selection process, rather than the shift itself.

4.4.5 Attentional Shift Implementation

The theories and findings on the movement of attention, considered in the previous section, have enabled us to make a number of propositions regarding the implementation of attentional shift in our model:

- Attentional capture is only triggered if the minimum level of activity present in the visual scene exceeds some activation threshold.
- Initial selection is based purely on elementary feature contrast.
- After attentional processing, the currently selected cell is allowed to decay centrally by means of a temporary inhibitory signal. This signal may take the form of many individual inhibitory signals to each synaptic pathway or a receptive field covering the appropriate feature spaces.
- Ongoing shift selection is primarily based on elementary feature contrast of the remaining regions; bottom-up or top-down factors may influence the outcome of the selection process by providing biases, such as spatial and similarity biases.
- The shift process is distance-independent and discrete, thus the processing focus is relocated to the newly selected location abruptly.
- Finally, we assume that the model operates within the inhibition period, thus any selected locations may not be revisited.

Based on the above propositions, we have derived the complete attentional shift process as follows:

1. attentional capture in initial selection mode as given in Section 4.4.2 - elementary feature contrast detection, provided the activation threshold is exceeded;
2. WTA competition to select the most conspicuous region, referred to as the high activity region;
3. selective transfer to high activity region;
4. attentional focus processing;
5. decay of the currently selected cell by an inhibitory signal, lasting for the whole duration of a scene analysis;

6. attentional capture in ongoing selection mode, i.e., selection is based on biased elementary features, and then back to Stage 2. The shift between the old and new locations is distance-independent and discrete, thus the processing focus is relocated to the newly selected location abruptly; and

7. the process continues until the level of activity falls below the threshold.

The shift process above is a more detailed version of the operation described in Section 4.4.1, and is graphically illustrated as a flowchart in Figure 4.13.

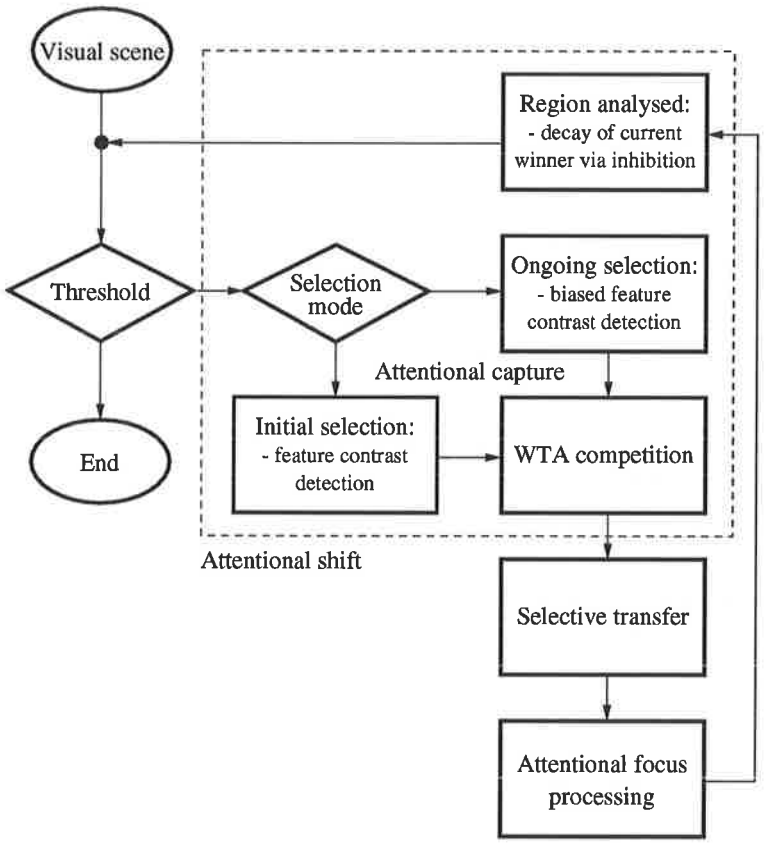


Figure 4.13: Flowchart for attentional capture and shift. For highlighting the processes involved in attentional capture and shift, other processes are grouped as one in the unit labelled "attentional focus processing".

The activation threshold may be placed immediately after the visual scene stimulates a spatial neural pattern of activity, so that the threshold gates the overall or global contrast level as in Figure 4.13. Alternatively, it may be placed locally, in which contrast levels for individual partitions of the visual scene are computed in parallel, and the resultant activity levels are thresholded, allowing through only highly contrasted regions for WTA competition.

Decay of the selected cell. Since inputs to the high activity WTA field are from overlapping regions, neighbourhoods that are close to the current winner must also be inhibited based on their proximity to it. In other words, a region that is close to the winner region may contain partially stimuli that are already processed; how much depends on the amount of overlap. For this reason, we have applied an inhibitory Gaussian filter to cells in the high activity WTA field, centered at the centroid of the current winner, so that the amount of inhibition falls away with distance:

$$\bar{G}_{kl} = 1 - \bar{G}_o \exp\left(-\frac{(k-k_o)^2 + (l-l_o)^2}{\varrho^2}\right) \tag{4.31}$$

where $\bar{G}_o$ is a constant and (k_o, l_o) is the centroid of the region to be inhibited. The spread (standard deviation) of the receptive field is specified by ϱ; the size of which is determined by the amount of overlap between neighbouring partitions.

Proximity and similarity biases. There are two obvious ways in which the proximity bias may be implemented. The first is by a receptive field that is similar to the one we have used above for the decay of the selected cell, but instead of inhibitory in nature, this receptive field acts to enhance all neighbouring cells in the high activity WTA field. Also, the scope of this receptive field is applied globally covering the entire WTA field, while the one above is only applicable locally. The other way is by presynaptic facilitation, as in Section 4.3.3, such that synaptic signals to the high activity WTA field are facilitated according to their proximity to the centroid of the currently selected region.

Implementation of the similarity bias requires separating elementary features into feature maps, so that features that are similar to the currently selected location will be facilitated. This facilitatory signal may be a feedback from the high activity WTA identifying the major feature associated with the current location.

4.5 Rotation Invariance

Another important invariant property for object recognition is rotation invariance. Under the scope of this thesis, we restrict our discussion to in-plane rotation for two-dimensional objects. In this context, rotation invariance refers to the ability of an object recognition system to recognise a learned object despite changes in its planar orientation.

There are at least three basic approaches to achieving rotation invariance in neural recognition systems, in fact invariance in general, including translation and scale as well [10, 18]. However, these techniques are more suitable to the supervised multilayer feedforward type of neural net-

works than the proposed visual scene analysis system, thus we shall restrict our discussion to some of the more relevant ideas.

The first of these three approaches is *invariance by structure*. Under this scheme, the neural architecture is arranged in such a way, through weight sharing and a higher-order structure, so that rotated patterns of the same input produce the same output classification. This is a particularly popular implementation of invariance in feedforward neural networks, examples of which are [49, 147, 176]. However, as Barnard and Casasent [10] pointed out, such networks are constrained by the need to duplicate weights for locations of equal distance. Because of this the number of connections required is unrealistically large, making it almost infeasible to implement.

The second, *invariance by training*, is somewhat a brute-forced approach to rotation invariant recognition. It involves the training of a neural network with a large number of examples of the same object in various orientations. An obvious deficiency is the need for a substantial data set that is able to cover the entire operating range, otherwise the invariance ability of the network is reduced. Furthermore, the time required to train such a large set of data is impractically long. It should however be pointed out that the human visual system seems to have multiple representations in specific orientations for highly overlearned objects. For example, one could learn to read upside down if enough practice is performed.

The last one is *invariance by preprocessing* or *feature spaces*, in which features that are invariant to transformations are extracted by a suitable choice of preprocessing. Some of the preprocessing methods that have been employed are moments [147], log-polar transform [162, 163], Fourier transform [69], and direct rotation [60, 178]. It has been pointed out [10] that the use of moments is plagued by noise. Furthermore, this approach is computational intensive, requiring new computation for each new input image.

It is not unusual to employ a combination of the abovementioned approaches. One example is by Perantonis and Lisboa [147], in which a higher-order structure is utilised together with moment classifiers. The main benefit of this hybrid approach is a reduction in the number of weights required. Another example is by Rumelhart, Hinton, and Williams [158]; their neural model has been regarded as a form of invariance by training [10] as well as by structure [18].

Having reviewed, in mostly an engineering perspective of rotation invariance, let us consider the psychological view of the issue. It has been well documented in the findings by Shepard and his co-workers [167, 168] that humans take more time to recognise a shape that is oriented away from its upright position. In fact, the recognition time increases with increasing rotation angles up to 180^o, independent of whether the rotation is clockwise or anticlockwise. This finding has led to the interpretation that a representation of the shape is rotated mentally so as to align the

represented object with a perceptual frame of reference. The resultant mental transformation is referred to as *mental rotation*. The level of consideration for mental rotation in the current approaches to rotation invariance is very minimal. Particularly, the first and second approaches above do not perform explicit rotation at any stage of processing.

However, Tarr and Pinker [183] have suggested it is not necessary to perform mental rotation each time a rotated object is recognised. According to [183], rotated objects can be recognised readily via some short-cuts:

- *unique visual cues* – orientation-independent features or parts that are unique;
- *multiple stored models* – as mentioned in the second approach that highly familiar objects may be stored in multiple orientation-specific representations in memory, which allows direct matching against the input; and
- *dimensionality* – different objects consists of the same parts may be distinguished if the objects differ only in how the parts are arranged along a single dimension.

There are other factors that may affect the execution of mental rotation, and thus the identification of disoriented objects. For examples orientation congruency effects are studied in [95]; effects of stimulus complexity and familiarity are investigated in [15]; Jolicoeur and Cavanagh [96] considered the relationship between mental rotation and physical motion, as well as the effects of surface medium on mental rotation; the effects of dimensionality and type of task on mental rotation by Shepard and Metzler [169]. Some of these factors along with the visual short-cuts mentioned above go beyond the scope of the current study, and thus are not implemented in our proposed model. They do, however, serve as important factors that help shape the proposed model and its implementation.

An interesting attempt for a rotation-invariant neural pattern recognition system to take into account the phenomenon of mental rotation has been proposed by Fukumi, Omatu, and Nishikawa [59]. The emphasis of the paper is the application of the *theory of information types* developed by Takano [182] to explain mental rotation. The theory seeks to classify various information types into two categories for recognising disoriented patterns: orientation bound and orientation free. The first category represents disoriented objects that must be mentally rotated in order for recognition to occur, while mental rotation is not required in the latter. This has important implications on when and how mental rotation is triggered, and the processing of any proposed recognition system.

In this section, we consider how rotation invariance can be implemented and incorporated into the model we have proposed so far. In particular, we are interested in implementing some of the

ideas inspired by mental rotation. We begin by analysing and summarising these findings and theories, so that we may propose how this incorporation may take place.

4.5.1 Rotation Invariant Model Propositions and Assumptions

After considering the issues involved in rotation invariance and mental rotation, we are able to make a number of assumptions and propositions regarding the incorporation and implementation of rotation invariance into our model.

Assumptions:

1. Objects to be recognised are complex in nature, consisting of parts arranged in more than one dimension.

2. There are no other unique visual cues available since the proposed model detects only one feature.

3. Objects to be recognised are low in familiarity, thus multiple views of the same object are not present in memory.

4. The variation in reaction times, for recognising a rotated object in varying orientations, has no direct effect on the operation of the overall model, and is thus not modelled.

The first three assumptions are necessary for the implementation of mental rotation so that only orientation bound objects are considered. There are several reasons for ignoring orientation free cases. For an initial implementation of mental rotation, it is logical to derive a relatively simple model as a basis for future refinements. Assumption 1 requires the analysis of an object and its parts to determine the dimensionality of the object, but the proposed model treats each object as a whole and parts analysis is not supported. Furthermore, the current understanding of the underlying mechanisms that perform this function is minimal and unclear, therefore we assume that all incoming objects consist of parts arranged in more than one dimension. It is obvious, as stated in assumption 2, that with our implementation there is only one feature detected, so it is not possible to provide another elementary feature as a visual cue for orientation free rotation invariant recognition. Training our neural system with multiple views of each object in varying orientations fails to illustrate the phenomenon of mental rotation, thus defeating the need to implement mental rotation. For computational reasons and the lack of complete understanding of the underlying neural mechanisms, we have chosen to ignore the variation in reaction time as stated in assumption 4.

Propositions:

1. Disoriented objects are processed by the visual system using mental rotation.
2. Mental rotation is to be carried in a bottom-up manner.
3. There exists a number of parallel frames of reference, each representing a transformation of the input via mental rotation, and are competing for activation.
4. The activated parallel frame of reference is aligned with another perceptual frame of reference - the central representation.

Proposition 1 is based on the above-stated assumptions, which have enabled us to propose that mental rotation is performed each time a disoriented object is recognised, i.e., rotation invariance is achieved through rotating the input object explicitly. While propositions 2, 3 and 4 are based on the concept of parallel frames of reference by Hinton and Parsons [81], which suggests that the brain possesses multiple and competing frames of reference in the bottom-up visual pathways through which the visual input can be transformed and analysed. The proposed model is also in agreement with the interpretation of mental rotation [96, 167] that a potentially recognisable object pattern is rotated internally so that it is aligned with a perceptual frame of reference which corresponds to the central representation.

4.5.2 Stages of Operation

For the implementation of mental rotation three major modifications are needed. Figure 4.14 reveals, in comparison with Figure 4.8, that the three modifications are located in the top right-hand quarter of the depicted framework. They are *invariant transformation*, *rotational templates* and *rotational and spatial alignment WTA fields*. We shall begin our discussion from the high activity field, since stages prior to this field have been described already in Section 4.4.1.

Overlapping regions from the high activity field are projected in parallel through an invariant transformation unit. This unit serves to transform an incoming pattern into several patterns consisting of the same elements in various orientations. These patterns are mapped directly onto a number of parallel frames of reference which we call the rotational templates. The templates can be regarded as a set of topographical, cortical maps encoding a subset of the visual scene in various orientations. Since each overlapping region is carrying its own spatial information and each rotational template has an associated orientation, the rotational template resulting from a particular spatial region has both translational and angular information. When these templates are gated by top-down memory patterns, synaptic signals corresponding to the

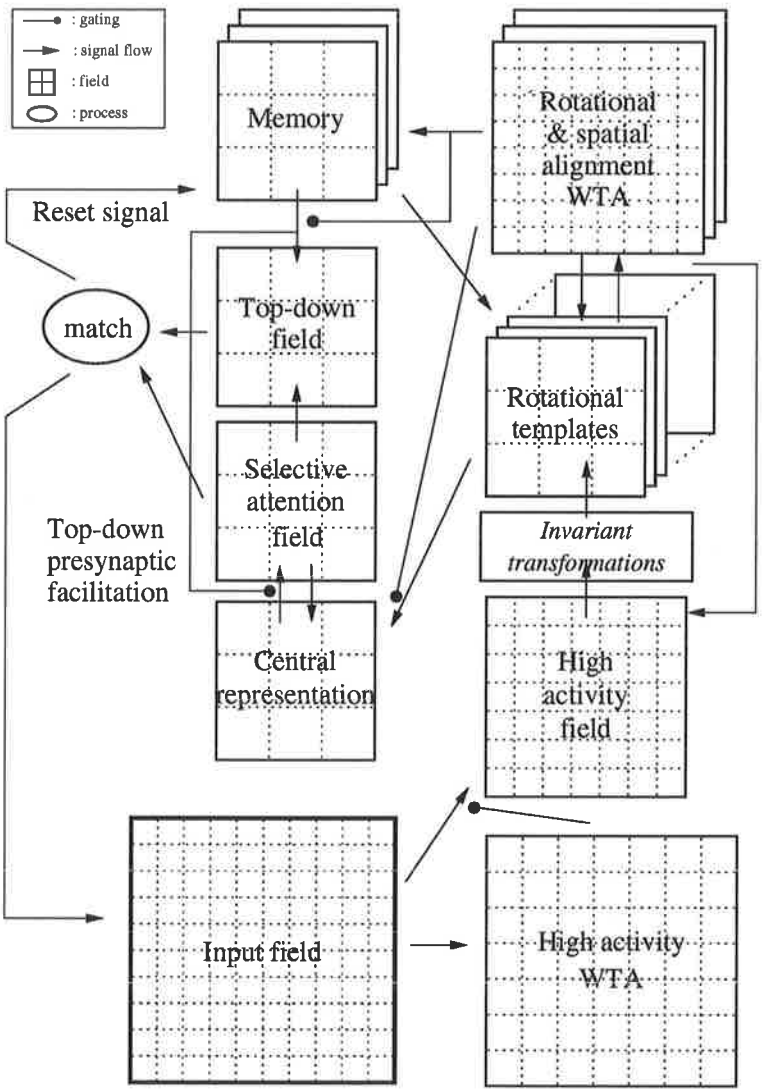


Figure 4.14: Visual object recognition and selective attention framework with rotation invariance. Small filled circles are gating terminals.

correlation between input and memory patterns are generated. The signals then participate in a multi-dimensional WTA competition in the rotational and spatial alignment WTA fields. We can regard this as a four-dimensional competition with two dimensions in spatial alignment, one in locating the canonical orientation, and the last one for the most likely match in memory. Therefore the rotational and spatial alignment WTA fields are a choice representation field as well as a spatial location field. So the winning cell provides information for four quantities on the identities of the bottom-up and top-down pair: the location of the input pattern in the input field along the x-axis, the location of the input pattern in the input field along the y-axis, the orientation of the input pattern, and the corresponding top-down pattern in memory. As a result, the activated rotational template is selectively transferred to the central representation, and its counterpart in memory is transferred to the top-down field. Thereafter, the processing continues on as described in Section 4.4.1.

By extending the operational process with the three modifications, we have now the stages of operation as follows:

1. attentional capture in initial selection mode for high activity region selection: elementary feature contrast detection, provided the activation threshold is exceeded;
2. WTA competition to select the most conspicuous region, referred to as the high activity region;
3. selective transfer of high activity region;
4. translation and rotation invariant transformations;
5. mapping of transformed patterns to parallel frames of reference;
6. multi-dimensional bottom-up memory activation;
7. selective transfer of bottom-up and top-down patterns;
8. short-term-memory iteration, matching of the bottom-up and top-down pair;
9. choice after matching:
   a) automatic attentional shift, go to stage 10,
   b) mismatch reset and back to stage 6, or
   c) top-down selective attention and back to stage 8;
10. decay of the currently selected cell by an inhibitory signal, lasting for the whole duration of a scene analysis;

11. attentional capture in ongoing selection mode, i.e., selection is based on biased elementary features, and then back to Stage 2. The shift between the old and new locations is distance-independent and discrete, thus the processing focus is relocated to the newly selected location abruptly; and

12. the process continues until the level of activity falls below the threshold.

Our treatment of rotation invariance is consistent with the three-level approach employed for the overall model, consisting of a high abstraction level using psychological experimental evidence, a mid-level using models and theories of neuromechanisms from physiological experimental results, and a low-level using simple models of the neuron to implement the model. In this case, we examine the findings in psychology regarding rotation invariance, in particular the phenomenon of mental rotation. For the mid-level, we employ a massively parallel interconnected neural structure; synaptic gain mechanisms for gating and selective transfer; an on-centre off-surround shunting competitive feedback neural network for WTA; and rotational templates as parallel frames of reference. In practice, rotational transformations are usually associated with some loss in information due to the discrete nature of image pixels, resulting from the lack of one-to-one correspondence between the mapping fields. Besides loss of information, a pixel location may be mapped to more than once causing a mapped value to be overwritten (this is graphically illustrated in Figure 4.15), so that a suitable computational algorithm must be derived to deal with multiply-mapped pixels, as well as information loss to best realise the invariance property.

4.5.3 Implementation of Mental Rotation

Our implementation begins after locating and transferring a region of interest for analysis in the previous stages, as the system attempts to establish a relationship between the stimuli in that region and a stored model in memory. The neuro-representation of this stage is illustrated in Figure 4.16, for aligning a potentially familiar object with its possible counterpart in memory. In order to perform this alignment, the region must undergo an invariant transformation, and determine the best correlated stored model with the transformed region. Thereby, generating an invariant object representation which is aligned with that particular stored model, both spatially and orientationally. The mechanism that determines the best correlated pair is a multi-dimensional WTA network, the rotational and spatial alignment WTA fields introduced in the previous section, whose function is to localise the most active pair of stored memory and bottom-up pattern.

The high activity pattern H, from stage 3 of the operational process, is partitioned into overlap-

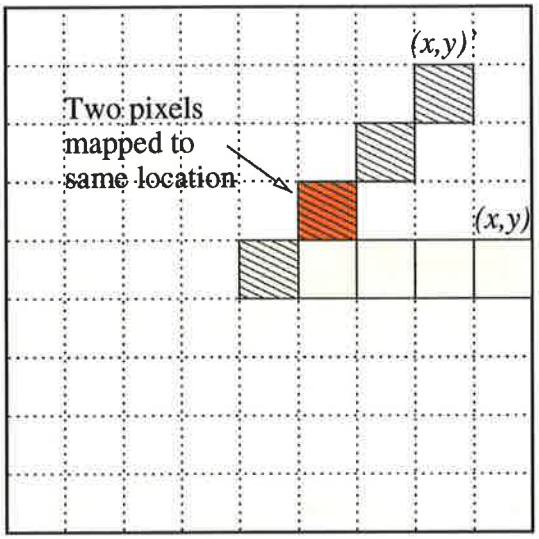


Figure 4.15: Discrete nature of digital images. It shows a horizontal bar of five pixel elements is reduced to four pixel in length after a 45^o rotation, and the highlighted element is mapped to twice.

ping regions $\mathbf{h}_{uv}$, whose individual activities may be described as $h_{\alpha\beta,uv}$:

$$\begin{aligned} \mathbf{h}_{uv} &\subset \mathbf{H} \\ h_{\alpha\beta,uv} &\in \mathbf{h}_{uv}. \end{aligned} \tag{4.32}$$

A set of parallel frames of reference, called the rotational templates, are generated for each sub-pattern $\mathbf{h}_{uv}$. The templates generated are of the same pattern but different in orientation, with a fixed angular distance apart. The transformation is denoted $\mathcal{R}$:

$$\mathcal{R} : \mathbf{h}_{uv} \rightarrow \mathbf{h}_{uv,\phi} \tag{4.33}$$

where $\phi \in [0, 2\pi]$.

Within each rotational template, the transformed indices α' and β' are given by

$$\begin{aligned} \alpha' &= \cos\phi(\alpha - \frac{N+1}{2}) - \sin\phi(\beta - \frac{N+1}{2}) + \frac{N+1}{2} \\ \beta' &= \sin\phi(\alpha - \frac{N+1}{2}) + \cos\phi(\beta - \frac{N+1}{2}) + \frac{N+1}{2}, \end{aligned} \tag{4.34}$$

where N is the dimension of the rotational template, thus $(N+1)/2$ is the centroid.

The above equation is derived from the rotation matrix $\mathbf{R}$ with the $(N+1)/2$ term representing the offset from the origin:

$$\mathbf{R} = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}. \tag{4.35}$$

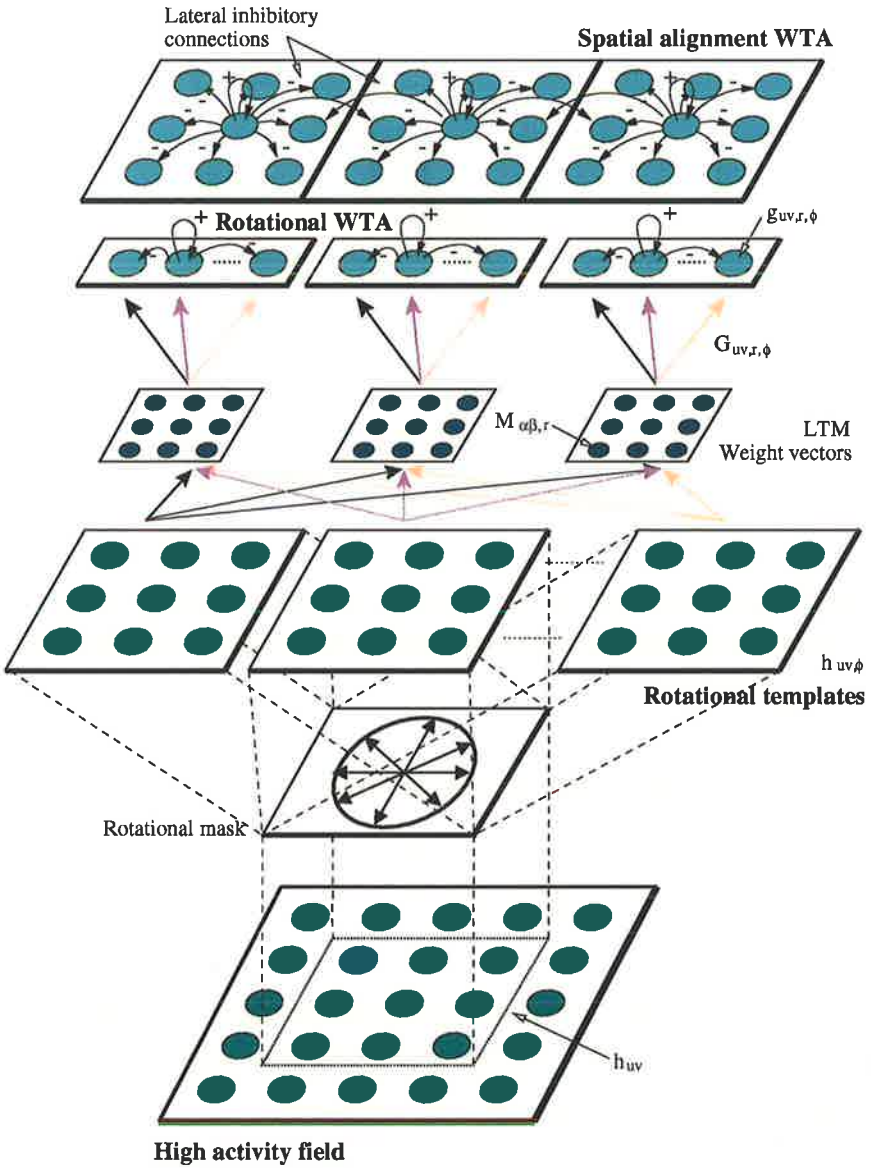


Figure 4.16: Bottom-up memory activation and invariant transformation. Consists of three sub-stages: i) invariant transformations; ii) mapping transformed patterns to rotational templates; and ii) multi-dimensional bottom-up memory activation. This structure serves to produce an object invariant representation of the attended object through transformations, and activates a stored memory location that correlates well with the fixated object. Overlapping regions from the high activity field are passed through a rotational mask, forming a set of parallel frames of reference for various orientations. These reference frames are then gated by LTM weight vectors, resulting in a multi-dimensional WTA competition, with the winner cell indicating the attended object's centroid, canonical orientation, and its potential match in memory.

Activities in a rotational template can be denoted as either $h_{\alpha'\beta',uv}$ or $h_{\alpha\beta,uv,\phi}$. A sub-pattern from location (u, v) in the high activity field, with spatial indices α and β, which are rotated through an angle of ϕ.

The rotational templates are gated by LTM traces, $M_{\alpha\beta,r}$, (stored models in memory), resulting in postsynaptic signals $G_{uv,r,\phi}$ as

$$G_{uv,r,\phi} = \sum_{\alpha} \sum_{\beta} M_{\alpha\beta,r} h_{\alpha\beta,uv,\phi} \tag{4.36}$$

where the subscript, r, denotes the rth stored model in memory. From Figure 4.16, $G_{uv,r,\phi}$ become the input signals to the rotational and spatial alignment WTA field.

Equation (4.36) shows that $G_{uv,r,\phi}$ is a measure of correlation between signals $M_{\alpha\beta,r}$ and $h_{\alpha\beta,uv,\phi}$. The strength of $G_{uv,r,\phi}$ is an indication of how well the sub-pattern from location (u, v) in the high activity field, rotated through an angle of ϕ, is matched to the rth LTM trace.

A multi-dimensional WTA network is employed to select a winner as shown below:

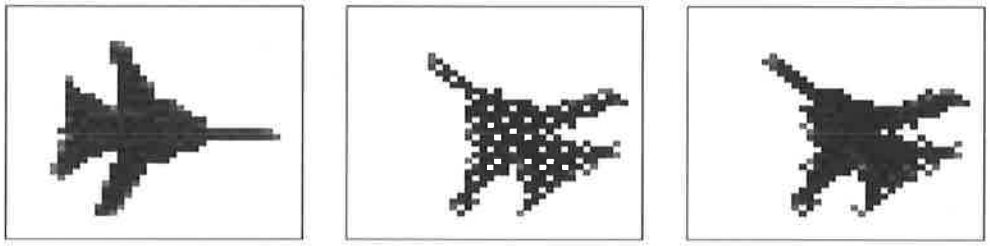
$$\frac{dg_{uv,r,\phi}}{dt} = -Ag_{uv,r,\phi} + (B - g_{uv,r,\phi})[G_{uv,r,\phi} + f(g_{uv,r,\phi})] - (C + g_{uv,r,\phi})D \sum_{ij,k,\theta \neq uv,r,\phi} f(g_{ij,k,\theta}) \tag{4.37}$$

where A, B, C and D are constants, and f is a threshold function.

Equation (4.37) is a four dimensional competition as discussed earlier, with two dimensions, u and v, in spatial competition, one dimension, ϕ, in rotational competition, and the remaining dimension, r, for determining the best correlated LTM trace. This is graphically illustrated in Figure 4.16.

We have already demonstrated in Figure 4.15 that on a square lattice of a visual scene, exact transformations in rotation by arbitrary angles are not realisable due to the lack of one-to-one correspondence on matrix arrays. Even though real world scenes are analog in nature, for modelling and simulations, digital images must be used. Therefore, we need to somehow deal with this quantisation effect [38], so that invariant properties can best be realised.

The algorithm we employed is to double map a point to its ceiling (rounding up) and floor (rounding down), so that information loss due to quantisation is being compensated to a certain extent. Also, if a location is mapped to more than once, the average value is sought as the pixel value. It is obvious that this process is noninvertible.



(a) (b) (c)

Figure 4.17: Quantisation effect of rotational transformation. (a) The original pattern; (b) after rotating 135^o about the origin without compensation; and (c) the same rotation with compensation.

4.6 Distortion Invariance

In order to obtain a perfect match between an input pattern and its counterpart in the stored memory, there must be a perfect elementary alignment between the two. If, however, due to various reasons, some part of the input pattern is out of place or deformed, depending on its severity, it could lead to nonrecognition of a known object. A robust mechanism for dealing with distortion is required for object recognition in images that contain minor distortion.

Of all the invariant properties considered in this thesis, distortion invariance is by far the most difficult and least researched. This can be attributed to the fact that distortion may take on many different forms, and thus is not easily modelled. Distortion of object figures in an image can occur in signal transmission, noisy environment, or objects being physically deformed prior to the capturing of the scene.

With its generalisation ability, a multilayer feedforward neural network can generally tolerate distortion or deformation to a small extent. This is also true for several other neural network models, in particular the Neocognitron by Fukushima [65, 64] has a tolerance mechanism that allows for small positional errors and distortions for pattern recognition. Convolutional network models that have a shared weight structure and involve subsampling have been found to be tolerant to minor distortion [105].

However, these models often have limited applicability in that they do not include other invariant properties or fail to take into account the presence of occlusion and clutter. For example, a distortion invariant object recognition system is proposed in [104], but has only been applied to images containing a single well-defined pattern. Another recognition system by Würtz [204] is not rotation invariant. A useful approach to distortion invariance is via the use of deformable shape templates; examples of this approach include [92, 108]. One drawback of this approach is the need for *a priori* shape knowledge and a set of probabilistic deformation transformations,

i.e., any deformed shape is deterministic, and thus distortions other than those generated by the probabilistic deformation transformations are not accounted for.

In this section, we aim to equip our proposed system with a robust mechanism that can deal with arbitrary distortions. In order to achieve this goal we need to define two new concepts in *band transformation* and *shape attraction* that are required for its implementation.

An arbitrarily distorted object is likely to have its outline deviated away from its normal topographical arrangements. Such deviations, however, are often confined by its neighbourhood relations, i.e., nearby locations in the canonical form of an object are still nearby locations in its distorted form, albeit not in exactly the same arrangement. This means local elements of a distorted pattern are within a close proximity to their former locations, such that a controlled tolerance at each spatial location can account for minor elementary positional shifts. Such tolerance may be in the form of a spatially aligned 2D Gaussian receptive field. When applied to every spatial location of a canonical object, a region, in the form of a band covering the object is created, hence the name "band transformation". The band is the result of combining circular expansions based on elementary locations. A graphical illustration of this operation is shown in Figure 4.18. Part (a) of the figure shows a normal, undistorted shape. This is sampled by Gaussian receptive fields in Part (b). The result in Part (c) is a band transformed shape of the original shape. Finally, Part (d) shows how a distorted shape can be tolerated under band transformation.

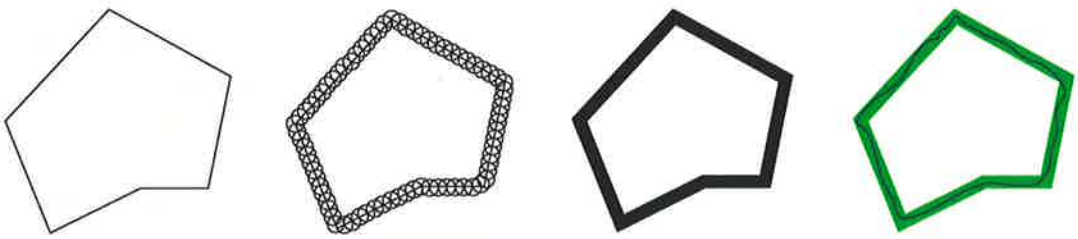


Figure 4.18: Band transformation: (a) canonical shape; (b) applying receptive field; (c) band transformed shape; and (d) distorted shape.

A complementary process of band transformation is shape attraction. The concept was first proposed by Lozo [113] to solve minor elementary shifts in the selective attention adaptive resonance theory (SAART) network, reviewed in Section 3.4, for shape recognition. This model recognises shape patterns that are positioned in the centre of the input scene such that it can only recognise a pattern whose centroid is positionally aligned with the centroid of its stored model counterpart. With the introduction of shape attraction, SAART is able to deal with minor elementary shifts, so that SAART is robust against minor distortions in the input pattern. However SAART has yet to address how distortion can be handled in conjunction with translation

and rotation in a large input scene. The concept of shape attraction is graphically illustrated in Figure 4.19. Part (a) shows conceptually the effect of applying shape attraction to a SAART neural network model, in which a distorted shape is sampled by 2D Gaussian receptive fields whose centroids correspond to spatial locations of another neural representation field, called the shape attractor. Those regions covered by the 2D Gaussian receptive fields can be regarded as regions of attraction generated by the shape attractor, such that elements within a region of attraction converge to a single location through the shape attractor on a projected layer as shown. An enlarged version of this process is shown in Figure 4.19(b), and this is analogous to a funnel or a vortex, where randomly located elements are drawn in through an opening, thus redirected to another spatial location. As a result, the distorted shape is transformed into a clean copy of the shape attractor pattern, provided the distorted shape consists of elements within the regions of attraction generated by the shape attractor. Obviously, the shape attractor must be a pattern from one of the stored models in memory.

It should be noted that there are various ways in which the concept of shape attraction may be implemented. A successful implementation of shape attraction (without translation, rotation and automatic attention) in SAART has been reported in [199], with its operation similar to that shown in Figure 4.19(a). This is a bottom-up approach to distortion invariance where the input pattern is transformed to match the top-down pattern; the other way is also possible where the top-down pattern may be transformed to match the incoming distorted pattern. Alternatively, elastic matching [25, 209] is a potentially suitable method for distortion invariance.

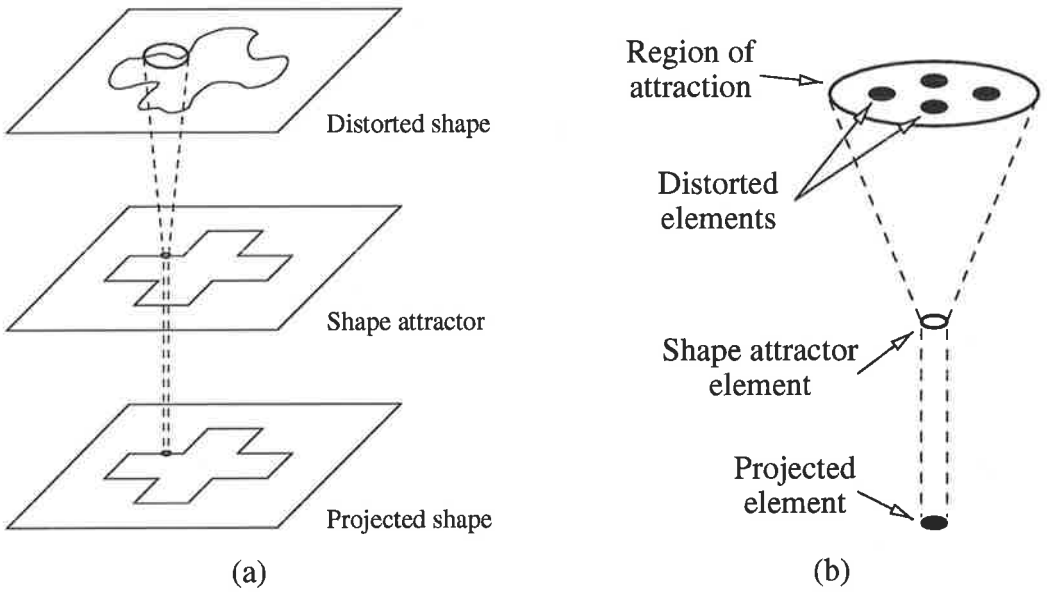


Figure 4.19: Shape attraction. (a) A distorted pattern is shape attracted into a clean copy of its attractor. (b) A zoom-in view of shape attraction in operation, in which the region of attraction acts as a vortex, through which distorted elements are channelled to a desired spatial location.

4.6.1 Stages of Operation

Several issues must be resolved for the enhancement of distortion invariance to our proposed visual scene analysis system. At the forefront of these issues is the question of when one should try to locate and recognise a distorted object. Assume that we have all the necessary mechanisms to deal with a distorted figure but what are the suitable conditions under which such processes are to be performed?

Once we have decided that there is a potentially known object in the visual scene but is deformed in shape, we need to determine, at the very least, coarsely its spatial location and orientation. That aside, we need to decide how we can compare the focussed pattern with the stored models in memory, and what constraints can be used to positively identify a distorted object, or through what transformation can such an identification be performed.

Our approach to these issues is based on two separate steps. First, there must be no familiar object recognised in the high activity field. Then we apply band transformation to the LTM patterns, so that when gated with the rotational templates, the extra tolerance provided can locate a distorted pattern, thus giving its spatial location and orientation. If the resultant match between the distorted pattern and a band transformed top-down pattern is reasonably good, then the second step can be taken to attempt to recognise the distorted pattern. This step involves the use of the activated top-down pattern, in its original form, i.e., non-band transformed, as a shape attractor to reshape the distorted pattern into a clean copy of the pattern. Figure 4.20 shows that these two steps have been incorporated into the framework as two separate units, denoted as *band tolerance transformation* and *shape attraction field*, respectively.

Extending upon the existing stages of operation to include band transformation and shape attraction, the overall stages of operation are given as follows:

1. attentional capture in initial selection mode for high activity region selection: elementary feature contrast detection, provided the activation threshold is exceeded;
2. WTA competition to select the most conspicuous region, referred to as the high activity region;
3. selective transfer of high activity region;
4. translation and rotation invariant transformations;
5. mapping transformed patterns to parallel frames of reference;
6. multi-dimensional bottom-up memory activation;

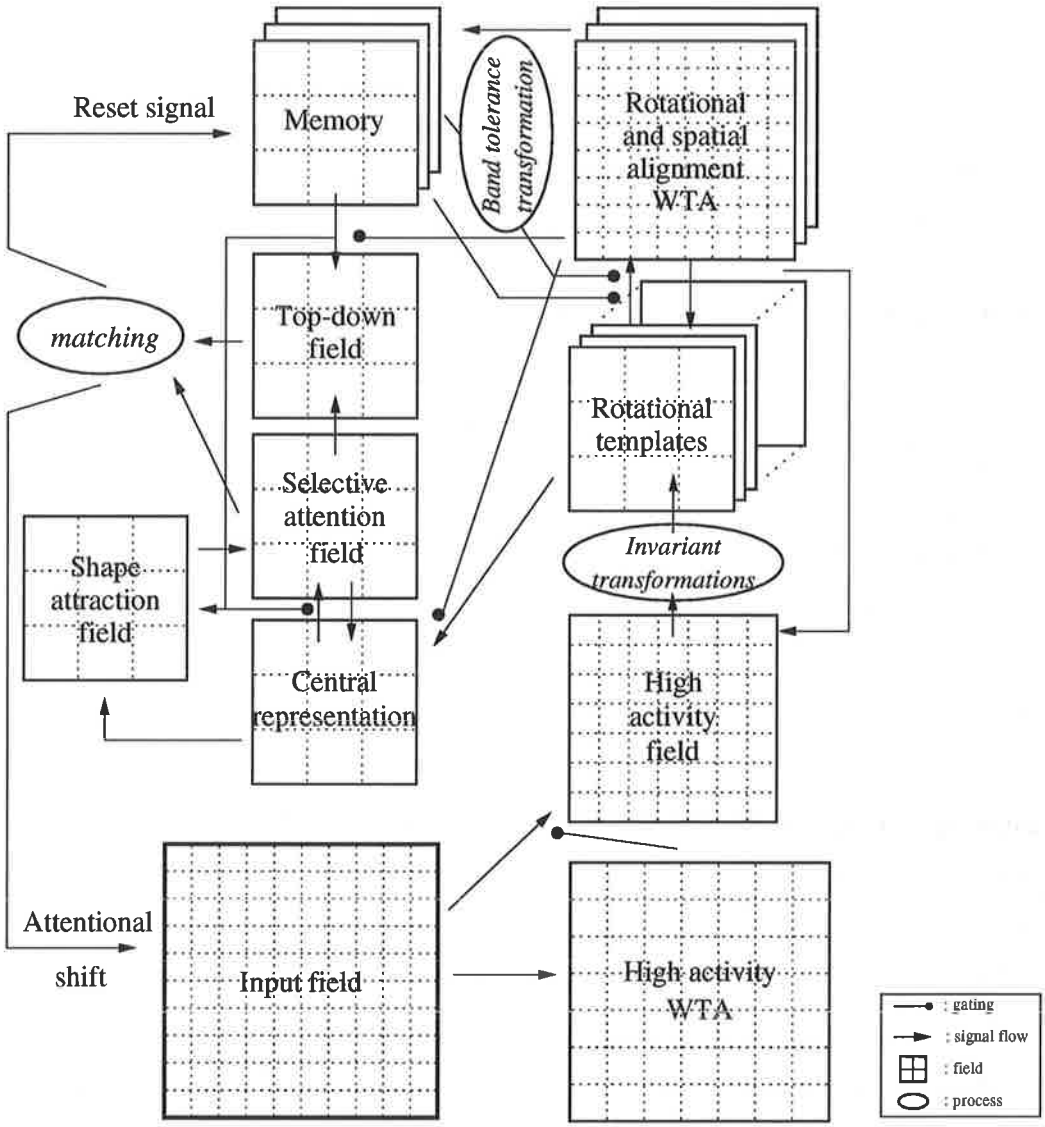


Figure 4.20: The complete visual object recognition and selective attention framework.

7. selective transfer of bottom-up and top-down patterns;

8. short-term-memory iteration, matching of the bottom-up and top-down pair;

9. choice after matching:
   a) automatic attentional shift, go to stage 10,
   b) mismatch reset and back to stage 6,
   c) top-down selective attention and back to stage 8,
   d) band tolerance transformation and back to stage 6, or
   e) shape attraction and back to stage 8;

10. decay of the currently selected cell by an inhibitory signal, lasting for the whole duration of a scene analysis;

11. attentional capture in ongoing selection mode, i.e., selection is based on biased elementary features, and then back to Stage 2. The shift between the old and new locations is distance-independent and discrete, thus the processing focus is relocated to the newly selected location abruptly; and

12. the process continues until the level of activity falls below the threshold.

As can be seen above, the overall process is highly complex and the use of a flow chart, in Figure 4.21, can be of a great aid to understanding the operations involved. For simplicity we have grouped a number processes that are performed prior to distortion invariant operations into a single block, details of which can be seen in Figure 4.13 or the complete flowchart is given in Figure 4.24, so that we can concentrate on operations that are required for distortion invariance.

It can be seen that band transformation and shape attraction are complementary operations with the former representing expansion and the latter contraction. Furthermore, band transformation serves to detect and locate distorted patterns, while shape attraction is required for recognition. These dual operations are analogous to the two fundamental morphological operators: *erosion* and *dilation* [174].

While the method proposed here is capable of recognising familiar patterns with minor distortions, it is not realistic to expect the system to be able to handle highly distorted shape patterns, since the shape or contour of an object is the very feature that we use for matching. By increasing the size of the region of attraction and the band of tolerance, we increase the system's ability to deal with distorted patterns, but we also increase the probability of incorrect recognition.

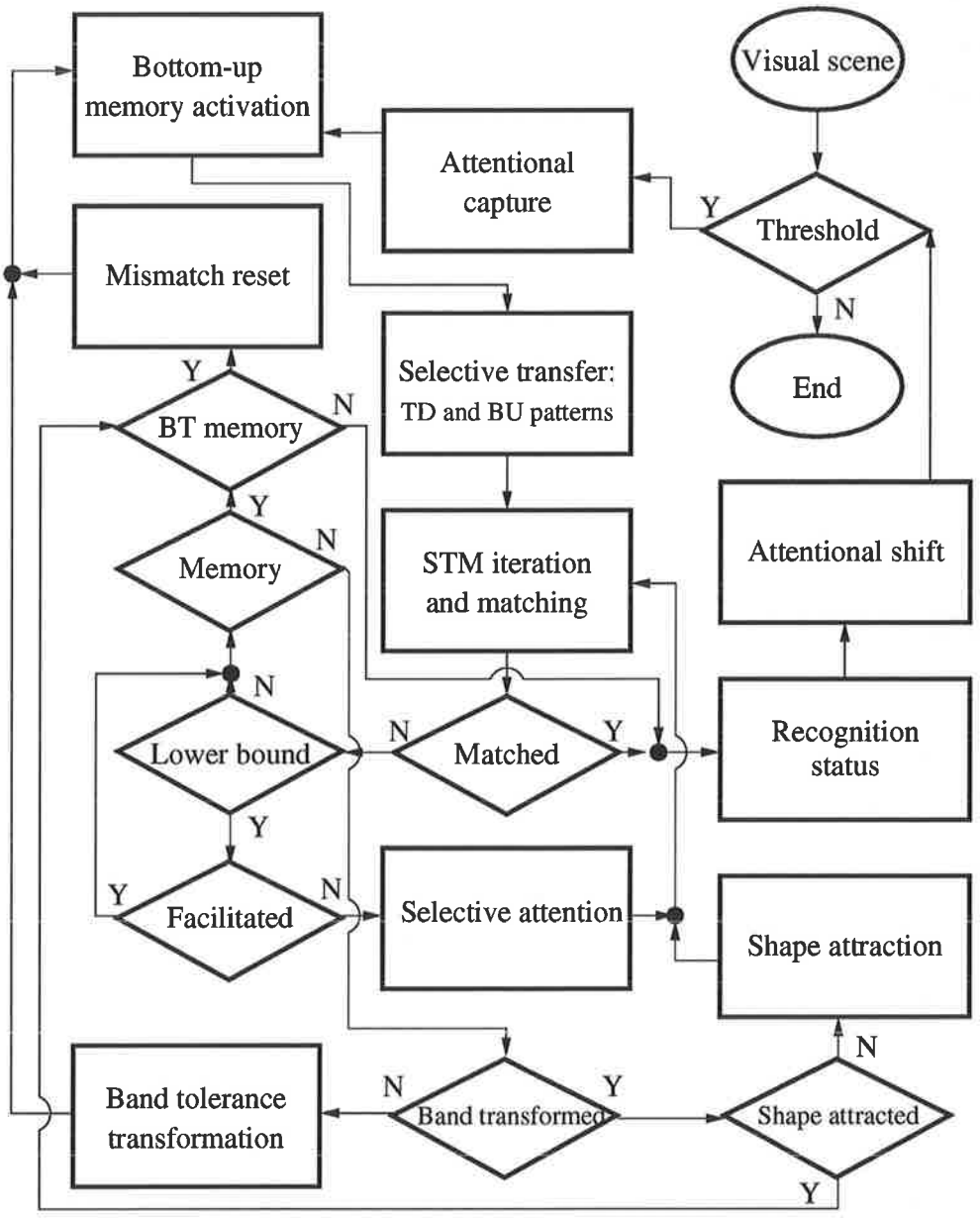


Figure 4.21: Flowchart for distortion invariance.

4.6.2 Band Transformation

Figure 4.22 shows a neural implementation of the concept of band transformation. A single LTM pattern element spawns a number of elements, labelled as band transformed LTM elements, which are confined within the predefined receptive field region as shown. The activities of individual band transformed elements are the activity of that single LTM pattern gated by distributed 2D Gaussian weights.

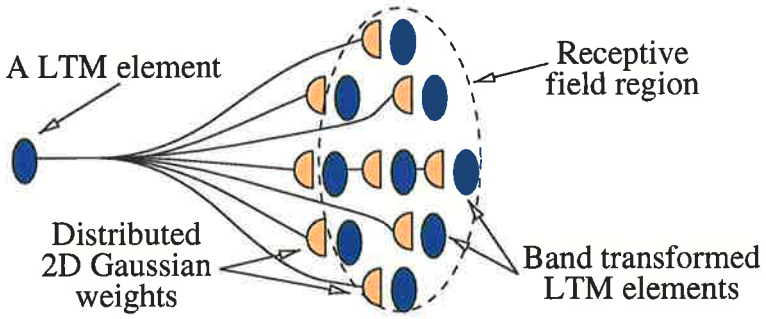


Figure 4.22: Neural diagram for band transformation.

Suppose $M_{\alpha\beta,r}$ is a stored model in memory and $\mathcal{M}_{\alpha\beta,r}$ is its band transformed model, then band transformation $\mathcal{B}$ is a mapping process such that $\mathcal{B} : M_{\alpha\beta,r} \rightarrow \mathcal{M}_{\alpha\beta,r}$.

For a 2D Gaussian receptive field of size $N_g \times N_g$ with standard deviation δ the transformation can be computed by using

$$\mathcal{M}_{\alpha\beta,r} = \max\left\{ M_{\alpha+i,\beta+j,r} \exp\left(-\frac{i^2+j^2}{\delta^2}\right)\right\}, \tag{4.38}$$

for $i, j = -\frac{N_g-1}{2}, \ldots, \frac{N_g-1}{2}$.

4.6.3 Shape Attraction

Since shape attraction is the dual process of band transformation, its neural implementation, shown in Figure 4.23, is thus a reverse of the one for band transformation in Figure 4.22.

We can see from Figure 4.23 that elements of the distorted pattern are modelled as neurons or cells, which are synaptically connected to a single cell, whose spatial location corresponds with the origin of the shape attraction region, in the shape attraction field. The weights connecting may be modelled using a distributed 2D Gaussian receptive field, such that greater contribution is extracted from the central part of the shape attraction region. The shape attractor element depicted acts as a reference for forming the desired top-down pattern by modulating the synaptic

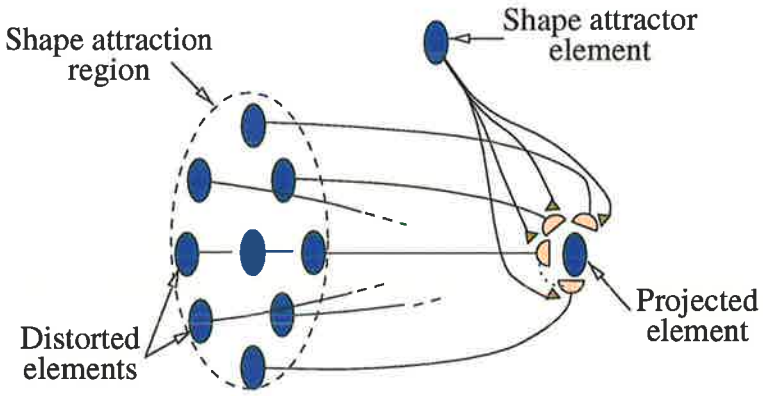


Figure 4.23: Neural diagram for shape attraction. The shape attractor element acts as a reference to the distorted elements.

gains. Depending on the requirement, the facilitatory signal provided by the shape attractor can be made to filter out distorted elements that are outside regions of attraction, so that a clean copy of the top-down pattern is obtained, or these elements may be suppressed but not eliminated if the facilitation is inhibitory for non-shape attraction regions.

Alternatively, we can choose the maximum element from the attraction region as the one passed to the shape attraction field, which is more direct and efficient for practical implementations.

Let $\mathcal{D}_{\alpha\beta}$ be a distorted object and $M_{\alpha\beta,r}$ be the reference for shape attraction, i.e., $\mathcal{D}_{\alpha\beta}$ is located by $\mathcal{M}_{\alpha\beta,r}$, then the shape attracted object, $\mathcal{S}_{\alpha\beta}$, generated by a shape attraction region of size $N_a \times N_a$ is given by

$$\mathcal{S}_{\alpha\beta} = \max\{\mathcal{D}_{\alpha+i,\beta+j}\} \quad \text{if} \quad M_{\alpha\beta,r} > 0, \tag{4.39}$$

for $i, j = -\frac{N_a-1}{2}, \ldots, \frac{N_a-1}{2}$.

4.7 Integrated Model of Architectural Framework

This section illustrates the operation of the overall framework, combining all of the visual function models proposed in this chapter, except distortion because the need to simplify the illustrations. However, distortion invariance of the proposed model will be demonstrated in simulations in Chapter 5.

Since the stages of operation and the framework diagram for the integrated model are as given in Section 4.6.1, they are not repeated here. By combining the flowcharts in Figures 4.13 and 4.21, we have the complete flowchart as shown in Figure 4.24. The flowchart is best explained using graphical illustrations depicted in Figures 4.25 and 4.26; they show exactly the expected outcomes at each processing stage.

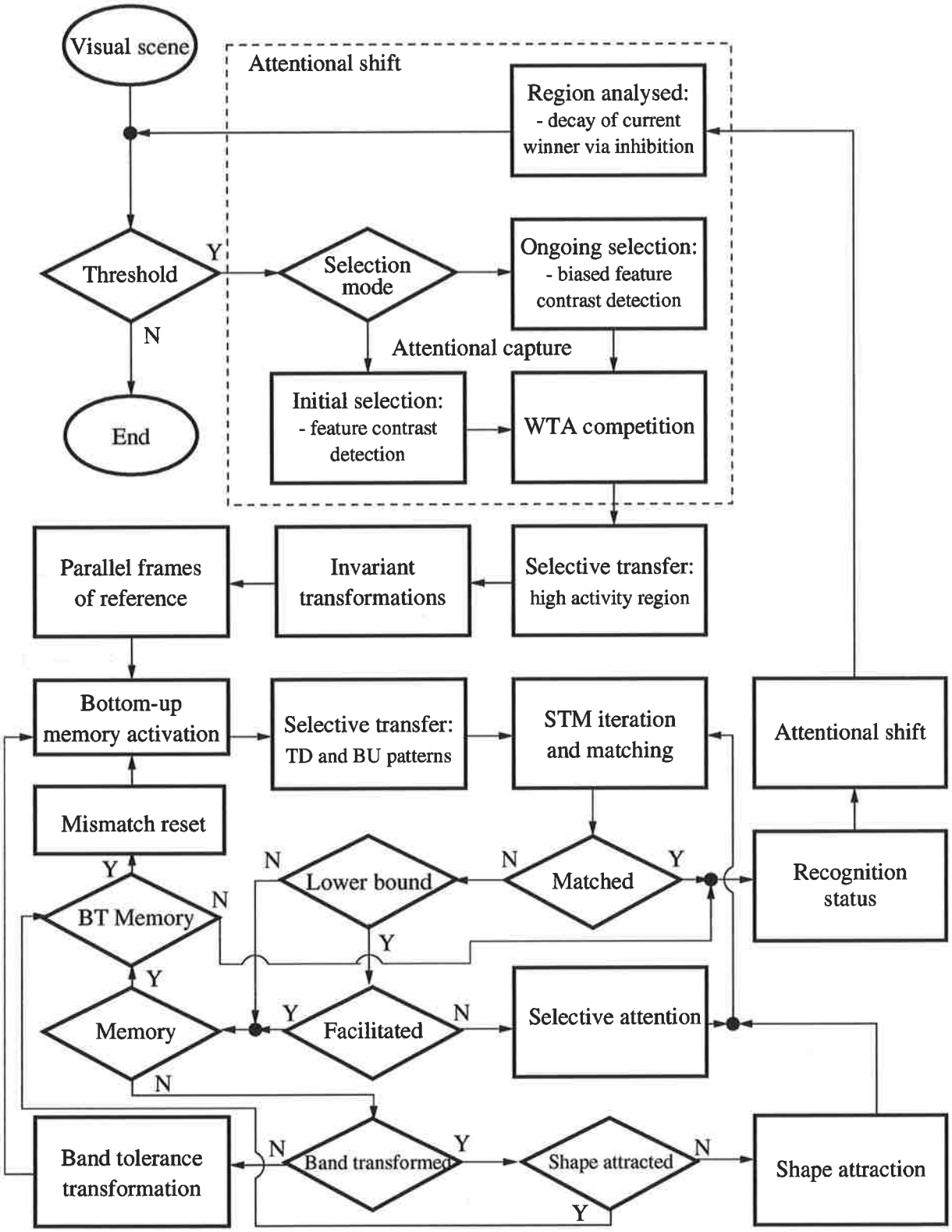


Figure 4.24: Processing flowchart for the complete framework.

Starting with Figure 4.25, where a visual scene consisting of a dog, horse and tree is first registered by the model in the input field. The scene is checked for activation before scene analysis can begin. Upon activating the attentional mechanism, it proceeds to capture a region of interest for further analysis. Based on elementary featural contrasts, a region of interest is located, indicated by the dotted region in the input field, as well as in the high activity WTA field. The disk in the high activity WTA field represents the winning cell which contains the spatial information of the region that has captured attention. This region of interest is transferred to the high activity field from which parallel processings for invariant transformations are performed, resulting in a large number of parallel frames of reference. Due to space limitation, we are only able to show the parallel frames that correspond to the chosen spatial location, indicated by the disk in the rotational and spatial alignment WTA field. These parallel frames are gated by top-down memory patterns, as a result the best correlated top-down and bottom-up pair are chosen as a potential match. This is indicated by the highlighted rotational template, memory, and rotational and spatial alignment WTA field. The chosen top-down and bottom-up patterns are transferred to the top-down field (TD) and central representation (CR).

If the outcome of the matching process satisfies the preset criterion, then the selected input pattern is deemed to be recognised and the attentional shift mechanism is set in motion. If the match is not satisfactory, then a mismatch reset occurs for unacceptably low degrees of matching, else top-down selective attention is triggered for reasonably good degrees of matching. In Figure 4.25, the pattern in CR failed to match the pattern in TD. Subsequently, pattern CR is top-down presynaptically facilitated by pattern TD to produce the pattern in the selective attention field (SA). The effect of top-down presynaptic facilitation is the suppression or in extreme cases removal of irrelevant information from the bottom-up pattern, as shown in SA. As a result of top-down presynaptic facilitation the matching criterion is satisfied, allowing the attentional system to be activated to shift the focus of attention from the current region of interest to the next by inhibiting the current winning cell in the high activity field.

The processing of the next region of interest can be seen in Figure 4.26. As before the region of interest is indicated by a filled circle in the high activity WTA field and the dotted region in the input field, where there is a dog with some background information. Repeating the same processing steps as above, we see that the rotational template with a 45^o anticlockwise rotation is selected as a potential match to the highlighted memory. With top-down presynaptic facilitation the bottom-up pattern matches the top-down pattern, so the object is recognised. Since there are no more regions of interest, the operation is complete until a new scene is registered.

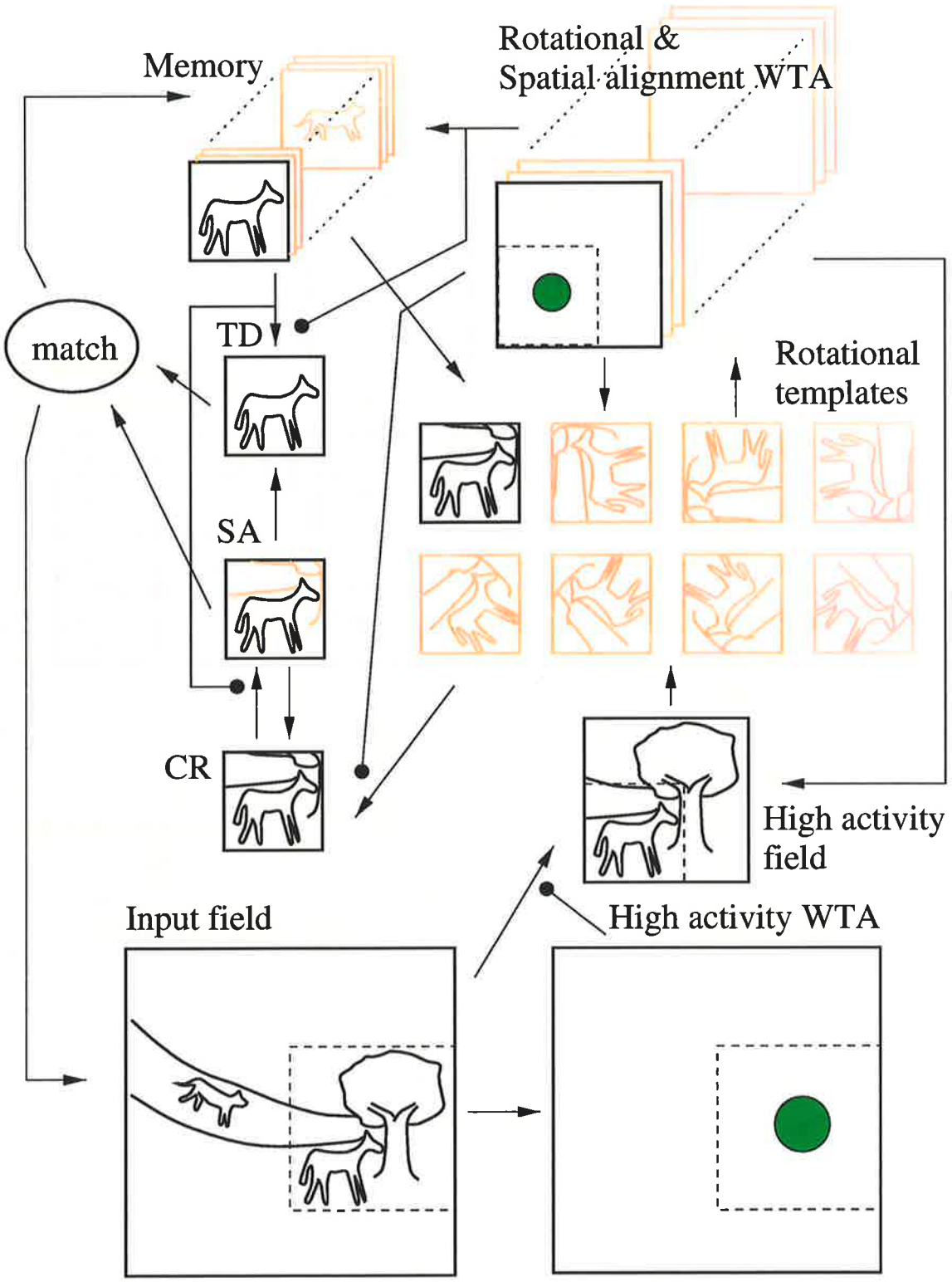


Figure 4.25: Graphical illustration of framework in operation part 1. Darker outlines represent the focussed region and activated fields.

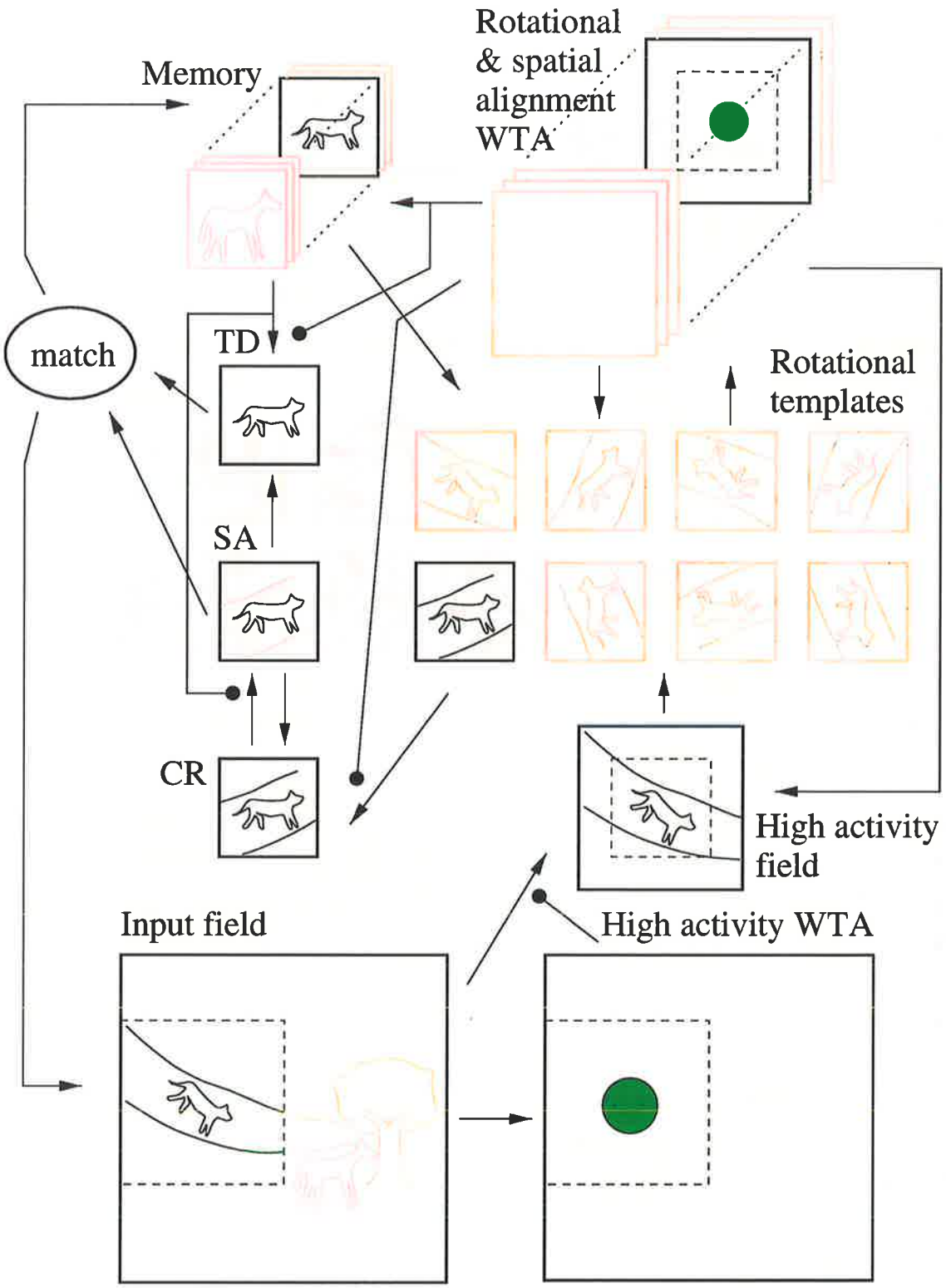


Figure 4.26: Graphical illustration of framework in operation part 2.

4.8 Conclusions

This chapter is the main theoretical part of this thesis. Theories and models reviewed in Chapters 2 and 3 are brought together to develop a number of neural models for visual object recognition and attention. The visual functions modelled include translation invariance, cluttered background with minor occlusion, rotation invariance, automatic attentional shift and capture, and distortion invariance. Together these models form an architectural framework that is based on a two-stage theory of biological vision and attention. The major contribution of this chapter is our attempt to merge selective attention with object recognition to form a framework for visual scene analysis. The chapter also highlights the research methodology adopted in Chapter 1.

Our models of visual object recognition and attention are based on a three level-approach from the highest abstraction to the lowest: psychological, neurophysiological, and implementational. First we attempt to develop a framework for our proposed model from psychological theories and models based on experimental findings and observations using a "blackbox" strategy, where the details are hidden, and with the emphasis entirely on the input and output relationship of the system concerned. Models and concepts on the findings and results of neurophysiological studies are used to devise neural representations, as well as establishing computational relationships and connections between various neural substructures within the framework. Finally, the framework is implemented using building blocks such as chemical synapses, and neural layers.

In the psychological level, the framework is able to perform higher-level visual functions by employing a two-stage parallel-serial architecture with bi-directional pathways in top-down and bottom-up connections. Attended bottom-up stimuli are stored in an object-based invariant frame of reference, while disoriented patterns are processed in competing parallel frames of reference generated by mental rotation. In the neurophysiological level, the framework consists of massively parallel feedforward and feedback connections. Some feedback pathways are engaged in attentional modulation (presynaptic facilitation) and gating. Also, many equations are based on the on-centre off-surround receptive field organisation. It accepts luminance contrast as visual stimuli and it has both short-term-memory and long-term-memory. In the implementation level, the neural networks are modelled using synapses, excitatory and inhibitory connections, and cells. Models of the neurotransmitter are used in synaptic facilitation, and existing artificial neural network models such as Adaptive Resonance Theory (ART) and Selective Attention Adaptive Resonance Theory (SAART) are used in implementation.

The next chapter aims to verify and analyse the proposed visual scene analysis system using computer simulations. Further extensions to the framework will be presented in Chapters 6 and 7.

Chapter 5

Model Simulations and Analysis

5.1 Introduction

This chapter provides a comprehensive simulation study of the proposed visual scene analysis system. The study aims to demonstrate the system's capability and effectiveness in various visual conditions, and analyse the system behaviour. The chapter also includes discussions on the selection of system parameters, and system limitations.

The chapter is arranged in an incremental fashion as in Chapter 4. An ART2 based neural network model for learning input patterns is presented in Section 5.2. Simulations on the translation invariant object recognition model are presented in Section 5.3. Sections 5.4 - 5.7 provide simulation results for additional visual functions in the following order: recognition in cluttered backgrounds and occlusion, automatic attentional shift and capture, rotation invariance, and distortion invariance. A discussion on the choice and design of system parameters is provided in Section 5.8, and a real-world application of the proposed visual scene analysis system is presented in Section 5.9. Subsequently, any limitations identified during the simulations are discussed in Section 5.10.

For each visual function, there are at least two simulations provided using synthetic images. Additional simulations are included for more complex visual conditions such as recognition in the presence of clutter and occlusion. In order to illustrate the operations involved and the model's capability, the two cases considered are usually one simple and one complex. We have chosen to use synthetic images in some of the simulations as they provide greater flexibility for object manipulation, a reduction in visual scene complexity which simplifies the problem under consideration, and a large database that can readily be generated. Furthermore, it allows fewer parameters to be chosen, and enables the model's inadequacy and limitations to be located with

ease. Another advantage of employing synthetic imagery is that it allows a systematic approach to implementing and developing a pilot simulation model. However, in the absence of common benchmark databases, a credible way to demonstrate the effectiveness of the proposed system is to apply it to real-world problems. So we provide simulations on real-world imagery in addition to those obtained using synthetic imagery.

Some of the simulation results are presented in the same format as the framework so as to provide greater ease for interpretation and understanding. However, for input scenes that contain several recognisable objects, only the first recognised object is presented in the framework form. The subsequently recognised objects are presented in a more compact form with LTM patterns and their input scene omitted, as these do not provide any further useful information and yet take up space needlessly. All simulations provided were achieved with software programs written in the C programming language.

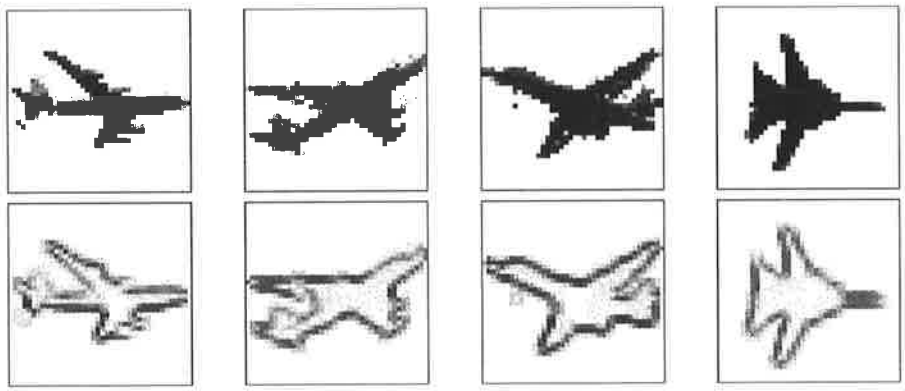
5.2 Learning

An important stage of a neural network recognition system is the learning phase. It is this ability to learn and adapt to its environment that has attracted most interest from researchers and scientists worldwide. A neural network learns about its environment through an iterative process of adjustments applied to its synaptic weights and thresholds.

In Chapter 1, we adopted a shape-based approach to object recognition. A common image processing technique to achieve this is edge detection. Unlike intensity, edge maps are less sensitive to noise and changes in the illumination. On the other hand, intensity maps can provide a better representation visually. It has also been found that the proposed system works equally well using intensity maps in most visual conditions, but intensity maps are prone to have a relatively high false alarm rate for visually cluttered environment.

Figure 5.1 shows four input objects in both intensity and edge forms. The depicted objects are images of four different types of aircraft captured digitally. Several preprocessing steps were performed to convert the images to the depicted form. The original images containing the aircrafts were resized and cropped to a size of 40×40 pixels. They were then converted to grayscale and segmented from their backgrounds using computer graphics tools. Finally, the images were edge detected using a 3×3 Sobel edge operator with both the horizontal and vertical kernels.

The learning algorithm chosen for the proposed system is the ART2 neural architecture [29], which has been reviewed in Chapter 3. Hence we shall not repeat the steps involved. However,



(a) (b) (c) (d)

Figure 5.1: Aircraft simulation input objects and their edge maps. (a) Aircraft I. (b) Aircraft II. (c) Aircraft III. (d) Aircraft IV.

results and the parameters used are presented in this section. The learned LTM weights will be used in all subsequent simulations, unless specified.

Figure 5.2 shows the degree of match between the top-down and bottom-up patterns for the duration of the simulation. Instances where new input objects were applied are highlighted by vertical dotted lines; the vigilance parameter is represented as a dashed horizontal line.

With no learned object prior to the start, Aircraft I activates an uncommitted node, which will always have a perfect match with the bottom-up pattern due to a zero top-down pattern. The adaptation of synaptic weights associated with the activated node is performed iteratively until the degree of match falls below the vigilance parameter (learning in ART is performed in the approximate match phase, instead of mismatch [73]). As expected, this occurs at the switching point from Aircraft I to Aircraft II, indicated by a single dip at the $25th$ iteration mark. As a result, the previously chosen node is reset, allowing another node to be selected, which is duly adapted with the bottom-up pattern.

As mentioned in Chapter 3 bottom-up activation is based on the strength of the dot product (or linear combination) between the bottom-up pattern and LTM weights, therefore upon switching to Aircraft III, the node that has been learning Aircraft II is reset, and the node that learned Aircraft I is selected ahead of an uncommitted node due to its LTM weight size. Since Aircraft I is very different from Aircraft III, this node is also reset, allowing an uncommitted node to be selected and adapted with Aircraft III. For Aircraft IV, there are now three learned nodes, so three resets occurred before it was learned. Details of the simulation are provided in Table 5.1.

Figure 5.3 is a very good illustration of the iterative nature of learning in neural networks. It depicts, for each input, the synaptic weights at various stages of the learning process. In conjunction with Figure 5.2, we can see that during the simulation the neural network is stimulated

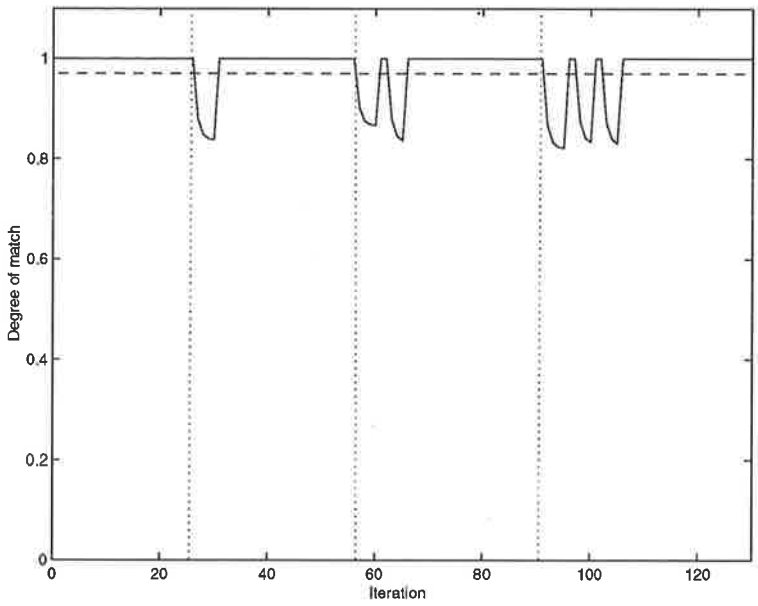


Figure 5.2: Graph of degree of match during learning. There are four input objects indicated by the vertical dotted lines. The horizontal dashed line is the vigilance parameter. The dips are caused by mismatch resets.

Table 5.1: Simulation details for learning

Input object size $N_p \times N_q$	40×40 pixels
Number of input objects	4
Vigilance parameter ρ	0.97
Number of STM iterations	5 per time step
Number of learning iterations	25 per input
Simulation time step Δt	0.5
Simulation method	Euler's method
STM equation parameters	$a = 10; b = 10; c = 1; d = 0.9; e = 10^{-6}; \theta = 10^{-3}$
Initial values	$p_{ij} = q_{ij} = u_{ij} = v_{ij} = w_{ij} = x_{ij} = 0;$ $Z_{ijk}^{bu} = \frac{0.001}{(1-d)\sqrt{(N_p+N_q)/2}}; Z_{ijk}^{td} = 0$

Refer to Section 3.3 for relevant equations and symbols.

by its input, as a result it undergoes changes to its LTM weights. The network becomes more knowledgeable about its stimuli after each iteration of the learning process. This is evident in the relative strengths in the weight patterns. As learning progresses, the LTM weights get stronger and stronger.

Due to the changes to its LTM weights, the network behaves in a new way to its stimuli over time. For example, at the end of the learning process, the committed (learned) node will respond maximally to bottom-up patterns that are familiar. So the network can begin to recognise more and more objects if the network continues to be exposed to novel stimuli.

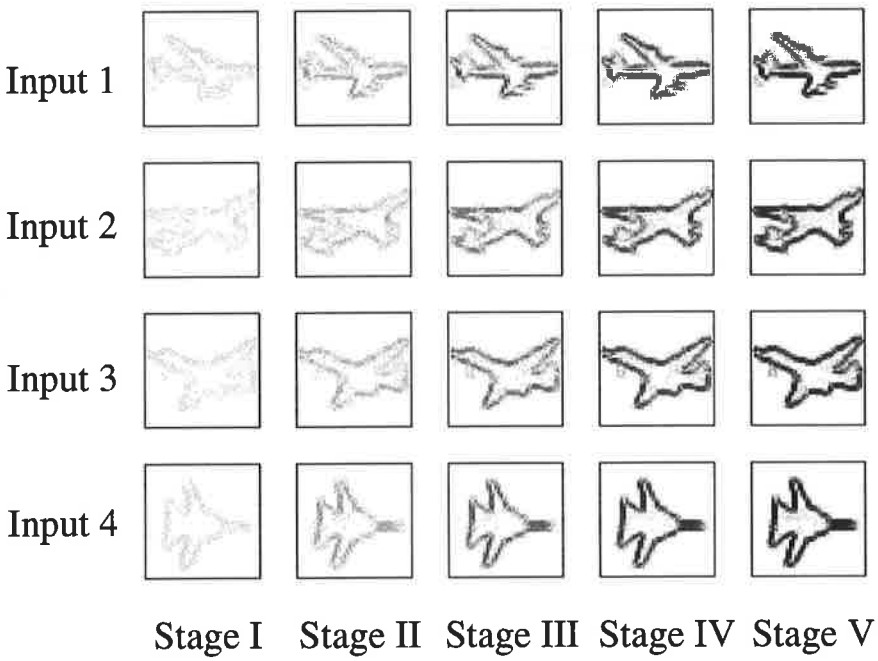


Figure 5.3: Bottom-up LTM patterns at various stages of learning.

5.3 Translation Invariance

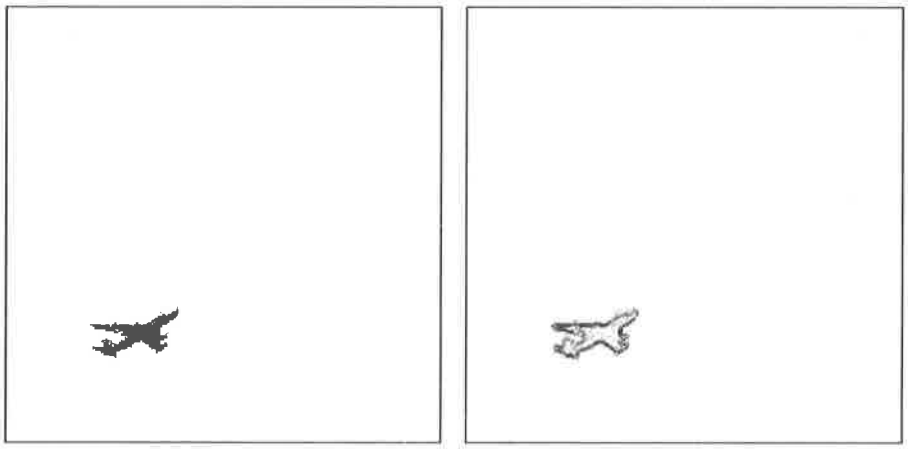
This section demonstrates the translation invariant recognition ability of the proposed system. The model simulated in this section is as described in Section 4.2, thus it does not possess other visual functions such as rotation invariance, and automatic attention. These will all be dealt with in the remaining sections in this chapter.

Typically, a scene requiring translation invariant object recognition can contain one or several objects that may appear anywhere within the visual scene. Hence, we consider two scenarios here: i) a single known object visual scene; and ii) a multiple known object visual scene. The

first is to highlight how a familiar object may be detected, located and recognised. The second emphasizes the bottom-up competition among objects within the visual scene. In both cases, we would expect a familiar object to be located by the WTA field, from which the located region is selectively transferred to the central representation to be compared with the bottom-up activated top-down memory.

5.3.1 Simulation I

Figure 5.4 shows the input scene, in both intensity and edge forms, for the single object simulation. It depicts an aircraft located randomly near the bottom left hand corner of a clear background scene. The input is simulated using the model in Section 4.2 with simulation details as given in Table 5.2. The choice of system parameters will be discussed in Section 5.8.



(a) (b)

Figure 5.4: Translation invariance: Simulation I input scene. A simple scene requiring translation invariant object recognition. (a) Intensity map; (b) edge map.

The simulation results are arranged in the same structural format as in Figure 4.1, and are shown in Figure 5.5. The region highlighted by a gray square in the WTA field is selected after a WTA competition based on convolution outputs between the input scene and the LTM weight patterns (acting as convolution kernels). The selected region is transferred to the central representation via presynaptic gating as shown in Figure 4.5. At the same time a memory pattern is activated by that selected region, causing the memory pattern to be transferred to the top-down field. The bottom-up and top-down patterns are then engaged in a matching process to determine whether the pair are an acceptable match.

To assist in visualising how the correct spatial location was selected, a 3D neural activity profile of the WTA field is shown in Figure 5.6. Under WTA competition, only the strongest cell

Table 5.2: Translation invariance simulation I details

Input scene size $N_i \times N_j$	200 × 200 pixels
Central representation size	40 × 40 pixels
Top-down field size $N_p \times N_q$	40 × 40 pixels
LTM weight pattern size	40 × 40 pixels
Available LTM patterns	4
Threshold ψ	0.01
Threshold θ	0.002
Vigilance parameter ρ	0.97
WTA field competition	$A = 1; B = 1; C = 0; D = 100$
Simulation time step Δt	0.005
Simulation method	Euler's method

Refer to Section 4.2 for relevant equations and symbols.

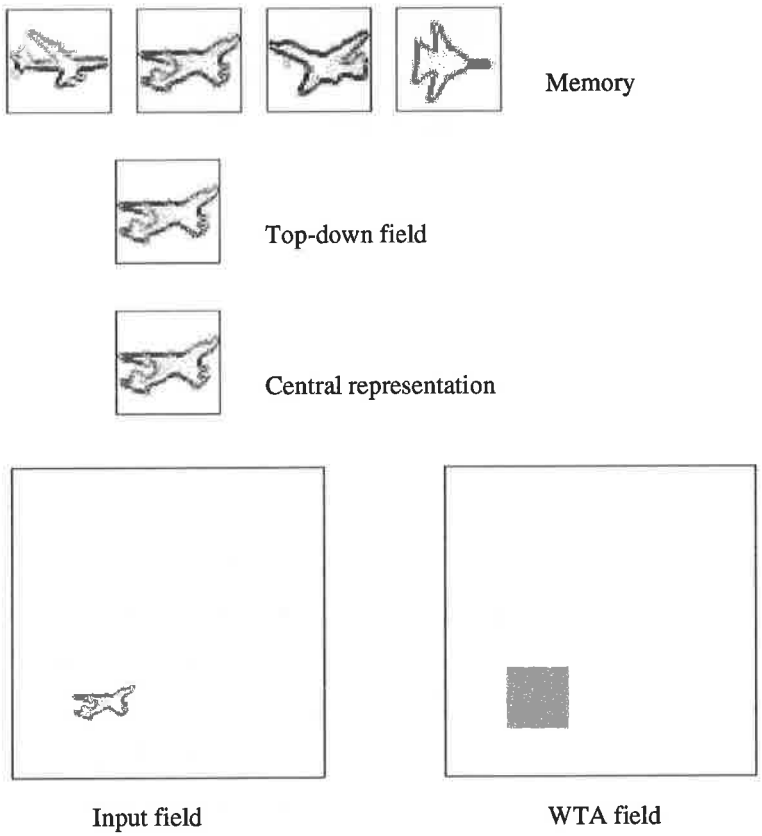


Figure 5.5: Translation invariance: Simulation I results. The results are set out in the same format as the framework model introduced in Section 4.2.

survives. So when applied to the synaptic signals, only the region that has the most common features remains, and this is evident in the shape peak depicted in Figure 5.6. Note that the spatial alignment WTA field is in fact multi-dimensional in nature (see Figure 4.1), so there would be four layers (four stored models in memory) thus giving four different 3D profiles. We have only shown one to illustrate the competitive nature between spatial locations for aligning the origins of the bottom-up and top-down patterns. While the other spatial alignment WTA layers concern the selection of the bottom-up activated memory pattern, which will be further demonstrated in the next simulation.

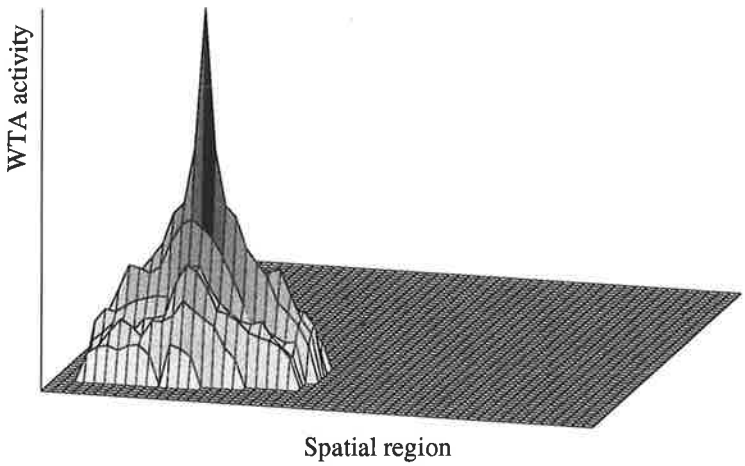


Figure 5.6: Translation invariance: Simulation I - 3D WTA field profile.

5.3.2 Simulation II

Let us examine what happens if an input scene contains multiple objects. Consider the scenario in the input field of Figure 5.7, there are four aircrafts randomly placed in the input scene. So how does the system know which aircraft to recognise or would the presence of other familiar objects cause any problem to the system? It is obvious from Figure 5.7 that other familiar objects pose no problem to the system. But why is aircraft III chosen ahead of the other three in the simulation results? The answer lays in the multi-dimensional nature of the spatial alignment WTA field. Besides competing for spatial alignment as shown in Simulation I, it also competes to activate a LTM pattern.

To illustrate the competition in the spatial alignment WTA field, all four layers of the WTA

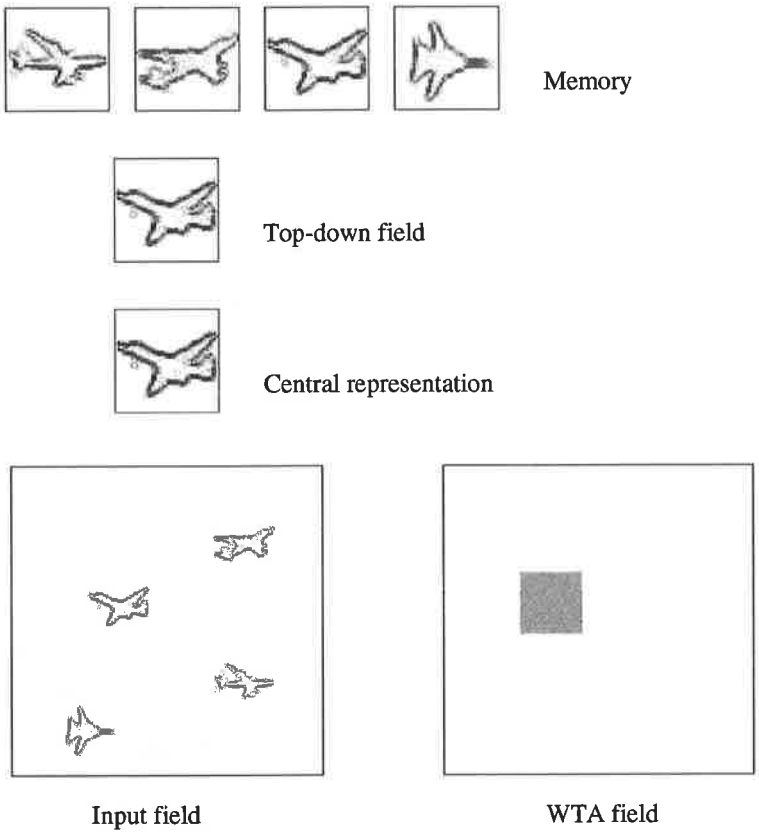
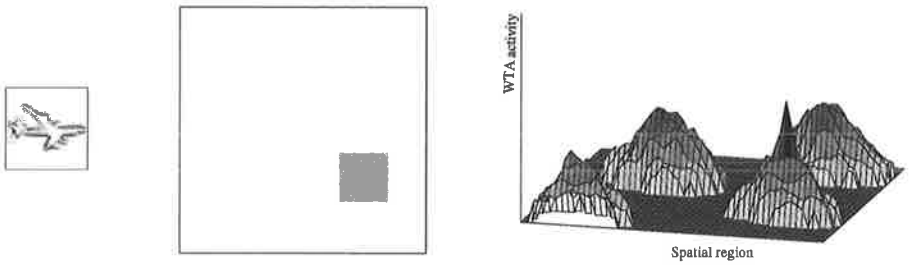
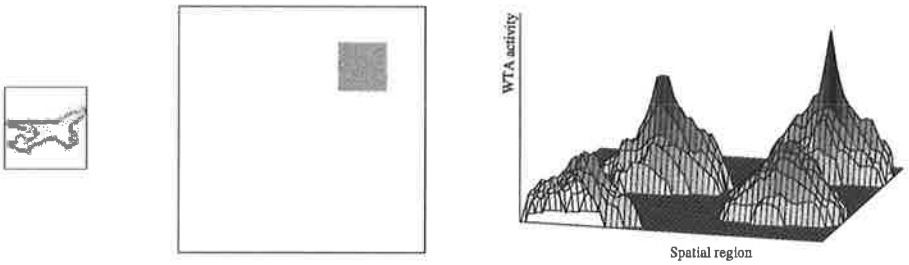


Figure 5.7: Translation invariance: Simulation II results.

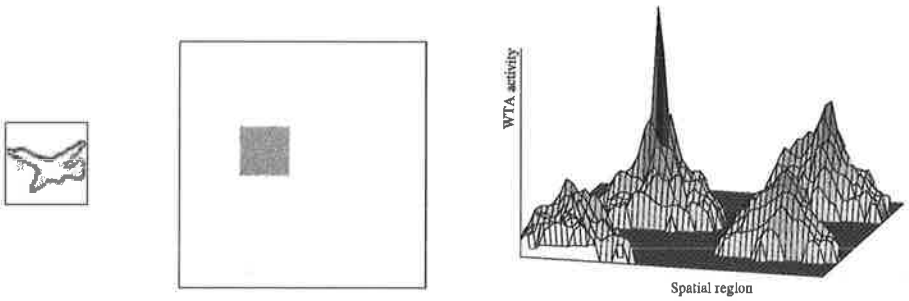
field are shown in Figure 5.8. The figure has four parts, each showing a LTM pattern, a 2D and a 3D spatial alignment WTA layer. The 3D profiles are correlation outputs between the input field and the LTM patterns. If we restricted ourselves just to any one of the 3D layers, we see that the peak location is where we have the most common features between the input scene and the kernel (LTM pattern). However, if we consider all four 3D layers, we see that layer 3 has the largest peak, thus is chosen ahead of the others. This should not be interpreted as Aircraft III has a better match with its selected location than the other pairs, simply this is caused by the stronger weight size of Aircraft III. A similar situation is discussed in ART3 [30] under "Trade-off between weight size and pattern match", in which a category may be selected ahead of another category that offers a better match, simply because it has larger weights. The paper suggests such a problem can be solved by normalising all weight patterns at all times. If normalisation was applied throughout, then there would be no difference between the peak values in the WTA fields, in that case, memory activation is based on the order of the LTM patterns. Although there is a clear advantage in applying normalisation in terms of searching, it can also be useful to not have normalisation in certain cases. For example, a high contrast object should be recognised ahead of a low contrast one.



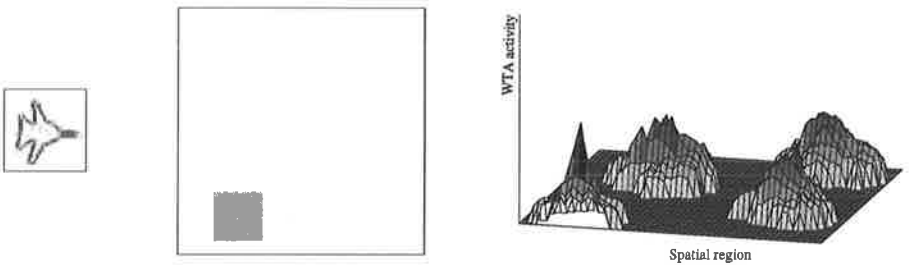
(a) Layer 1



(b) Layer 2



(c) Layer 3



(d) Layer 4

Figure 5.8: Translation invariance: Simulation II - 3D WTA field profiles.

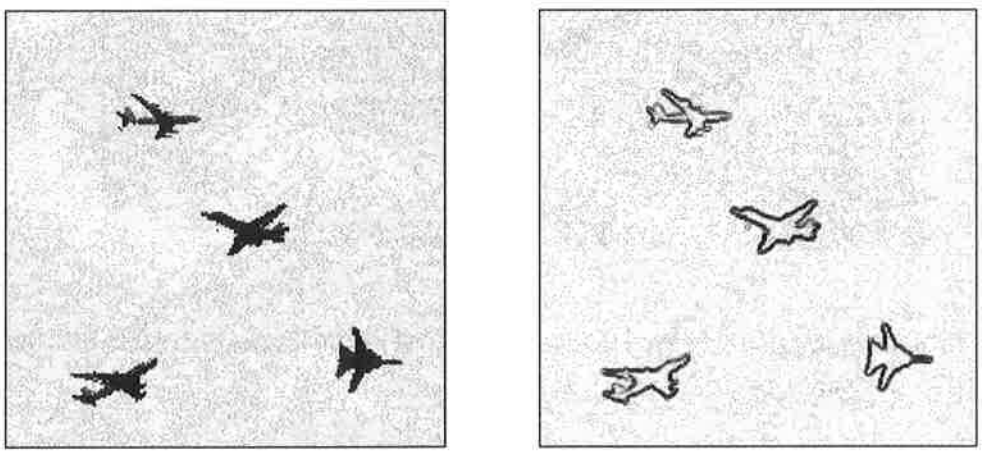
5.4 Recognition in Cluttered Images

This section demonstrates the proposed system's ability to handle object recognition in a variety of complex backgrounds, through the use of presynaptic facilitation.

The proposed system is expected to perform translation invariant object recognition in complex cluttered scenes in which objects may be subject to minor partial occlusion. So, three scenarios are to be considered: i) a non-clear visual scene; ii) a complex cluttered scene; and iii) a complex cluttered scene with minor occlusion. These cases are designed to firstly clarify the theoretical concepts used in modelling and illustrate the processing steps involved, and secondly to demonstrate the extent of the proposed system's capability to deal with such problems.

5.4.1 Simulation I

Consider Figure 5.9, which is an extension of Simulation II for translation invariance in Section 5.3.2. The figure contains four randomly placed aircrafts in a cloudy, non-clear background. The edge map appears to be reasonably clear and it is possible that top-down presynaptic facilitation may not be required to achieve recognition. Indeed, this is the case found in the simulation; the results are summarised in Figure 5.10. They correctly show that the bottom-up pattern, depicted in the central representation as well as indicated in the WTA field, is recognised as aircraft III, as given in the top-down field. The system is able to recognise the correct object because the amount of clutter is relatively low and the image contrast is high. However, the noisy background causes the degree of match to drop from 1 to 0.988. We can conclude that the model is tolerant to minor background noise without the need for presynaptic facilitation.



(a) (b)

Figure 5.9: Cluttered image: Simulation I input scene. A non-clear background. (a) Intensity map; (b) edge map.

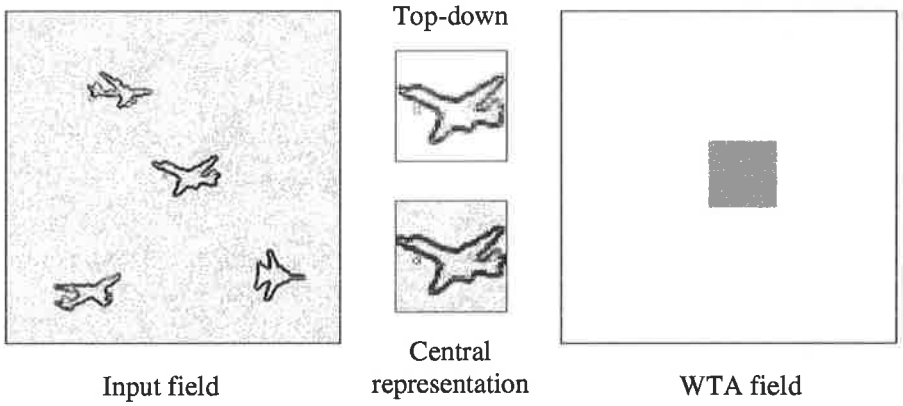
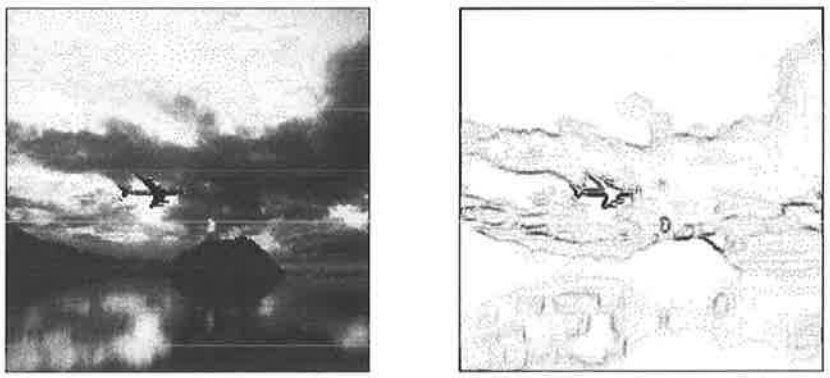


Figure 5.10: Cluttered images: Simulation I results. Results for input scene with a non-clear background. No presynaptic facilitation required.

5.4.2 Simulation II

A typical cluttered scene is shown in Figure 5.11. It shows an aircraft embedded in a complex scenic background. It can be seen from the figure that after edge detection, the target object is surrounded by clutter and irrelevant edges which have corrupted the outline of the aircraft. This, undoubtedly, would reduce the degree of match between the target object and its counterpart in memory, which would normally cause recognition to fail. This simulation illustrates the use of top-down presynaptic facilitation to increase the degree of match by strengthening the target object and reducing the amount of clutter surrounding the target object.



(a) (b)

Figure 5.11: Cluttered images: Simulation II input scene. A highly cluttered visual scene. (a) Intensity map; (b) edge map.

The input scene in Figure 5.11 was simulated using the model in Section 4.3 with simulation details as given in Table 5.3. The parameters chosen are based on ART3 [30] and SAART [113], further discussions on the design of parameters are provided in Sections 5.8.

As before, the simulation results are set out in the same format as the framework, shown in

Table 5.3: Cluttered images: Simulation II details

Vigilance	Primary: $\rho = 0.97$; Secondary: $\varsigma = 0.94$
PMCNL parameters	$A = 1$; $\bar{A} = 0.1$; $B = 1$; $\bar{B} = 0.1$; $C = 0$; $D = 0.5$; $E = 1.0$; $G = 130$; $\bar{G} = 1.8\text{E}5$; $K_u = 0.1$; $n = N_i N_j - 1$; $Y = 0$; $\alpha_u = 0.05$; $\beta_u = \beta_y = 0.01$; $\gamma = 0.5$; $\Gamma = 0$; $\rho_y = 0.05$; $\theta = 0.05$;
PMCNL initial values	$x_{ij} = 0$; $u_{ij} = 1$; $v_{ij} = 0$; $\bar{v} = 0$; $y_{ij} = 1$; $z_{ij} = 1$; $F_{ij} = T_{ij}$; $J_{ij} = C_{ij}$;
Simulation time step Δt	0.1

Refer to Section 4.3 for relevant equations and symbols. Unspecified parameters are as given in previous tables.

Figure 5.12. The main feature of the results is the addition of the selective attention field. The model behaves much like the translation invariant model, except when the bottom-up and top-down patterns failed to match and the degree of match is above the secondary vigilance, then the process of top-down presynaptic facilitation is triggered. As explained in Section 4.3, presynaptic facilitation works by enhancing object specific bottom-up synaptic signals, allowing familiar object signals to be facilitated, while non-object signals are suppressed under mutual competition. As a result, the bottom-up object is enhanced and the background clutter suppressed or removed, depending on the level of facilitation and competition.

A better illustration of presynaptic facilitation with shunting competition is shown in Figure 5.13. It depicts the bottom-up pattern in various levels of facilitation, highlighting the effects facilitation level has on a bottom-up pattern. F_s is an extra gain factor associated with the facilitatory signal and D_m is the resultant degree of match. For the results in Figure 5.12, the default facilitation level (as set out in Table 5.3) is sufficient to lift the degree of match from 0.965 to an acceptable level of 0.975. If however the primary vigilance was set at a higher level, then the facilitation level must also be set higher to improve the degree of match.

From Figure 5.13, we can see that as the level of facilitation increases, the degree of match tends to increase, albeit not linearly. Also, as the degree of match approaches a certain level, saturation occurs. Any further increases in the facilitation level have little effect on the degree of match, e.g., from $F_s = 1.6$ to $F_s = 2.0$. However, if the facilitation level gets too high, it can have a negative effect on the degree of match. $F_s = 4.0$ caused a reduction in the degree of match relative to its previous facilitation level. Thus, the facilitation level should be kept at a moderate level to avoid impacting negatively on the degree of match. Unnecessarily high levels of facilitation can also cause a higher false alarm rate.

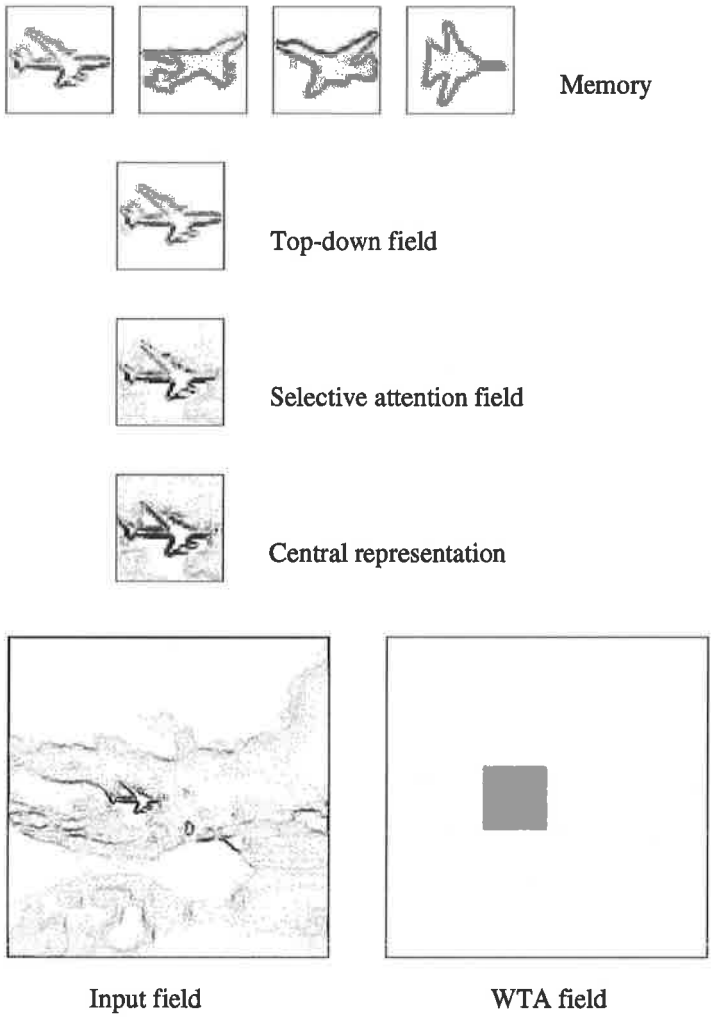


Figure 5.12: Cluttered images: Simulation II results. Presynaptic facilitation was applied, improving the degree of match from 0.965 to 0.975.

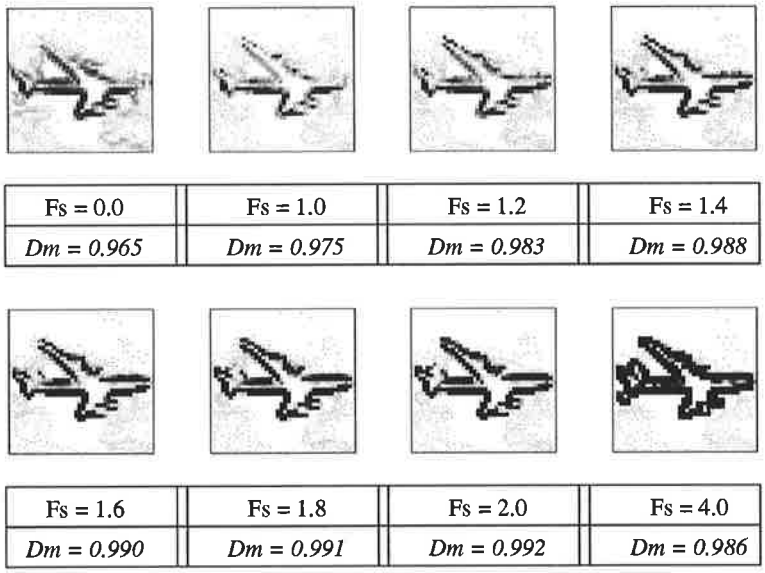


Figure 5.13: Cluttered images: Simulation II - effect of varying the facilitation level. In general, strong top-down presynaptic facilitatory signals have a positive effect on the degree of match. But as it gets too great, it could have the opposite effect. F_s is the facilitation level; D_m is the degree of match.

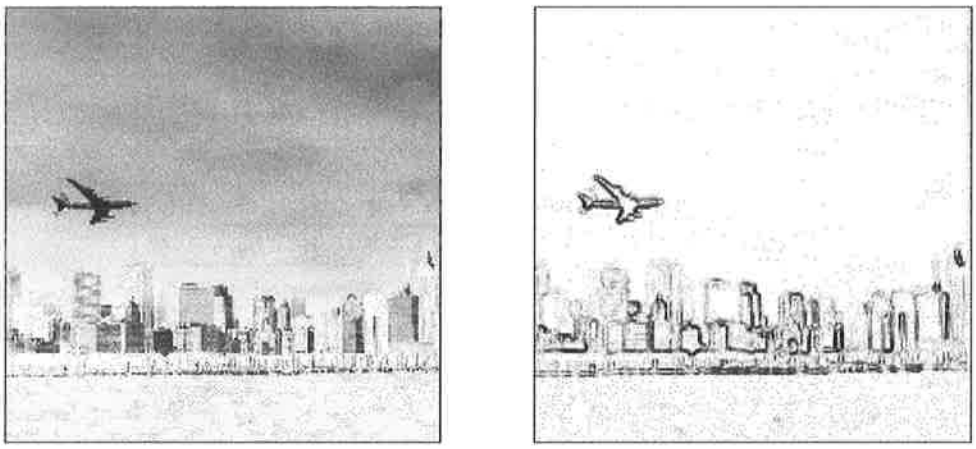
5.4.3 Simulation III

Occlusion is generally considered a more challenging problem than visual clutter. Although in both cases the outline of an object can be severely infringed upon after edge detection, the former has the problem that the occluded part is completely missing, thus provides no bottom-up information. Whereas in the cluttered case, it usually results in a degradation of its edge map, but one may use other visual features, if available, such as colour to assist recognition.

Consider Figure 5.14 which shows an aircraft flying over a city. In this simulation, we try to create a minor occlusion effect by covering the entire visual scene with vertical stripes. In real-life such problems could be caused by bad transmission of a digital image.

A common approach to occlusion is to match object fragments, rather than the whole object [86, 196]. This approach deals with occlusion in an explicit manner, but we have yet to implement object fragment matching mechanisms into the framework, therefore we use top-down presynaptic facilitation to solve the occlusion problem in an implicit fashion as shown in Figure 5.15. A more effective way for recognising object parts in occlusion is implemented in Chapter 7.

A close examination of the edge map reveals that the outline of the target object has been infringed, and is thus not complete, i.e., segments are not linked together, which greatly reduces the recognisability of the target object. This is confirmed by registering a degree of match



(a) (b)

Figure 5.14: Cluttered images: Simulation III original input scene. Prior to modifications to a highly cluttered scene with minor occlusion. (a) Intensity map; (b) edge map.

of only 0.884 without presynaptic facilitation. The recognition would surely have failed, if the parameters listed in Table 5.3 were used. However, with some simple adjustments in the secondary vigilance, the facilitation level and the shunting competition gain value, the proposed system was able to recognise the target object as shown in Figure 5.15. Obviously the secondary vigilance parameter was lowered to allow for presynaptic facilitation to proceed, and the other two were changed to provide greater facilitation and competition to enhance the critical spatial features and suppress the background clutter.

5.5 Preattentive Processing: Automatic Attentional Shift and Capture

The framework to be simulated in this section overcomes a major deficiency of the previously simulated model. Vision is a continuous process. We observe and recognise many objects in a visual scene (not all objects are recognised, only those attracting our attention), as long as our eyes remain open. To achieve this, the proposed system must be able to perform attentional capture and shift automatically, so as to recognise all familiar objects in the visual scene. As discussed in Section 4.4, this model is based on the preattentive stage of visual perception. The main function of this stage is to perform a preliminary analysis of the visual scene, from which the visual system decides where to focus. In a way, it acts as a task scheduler on the basis of the importance of various parts of the input.

With this important addition, the proposed system is expected to be able to detect and locate a

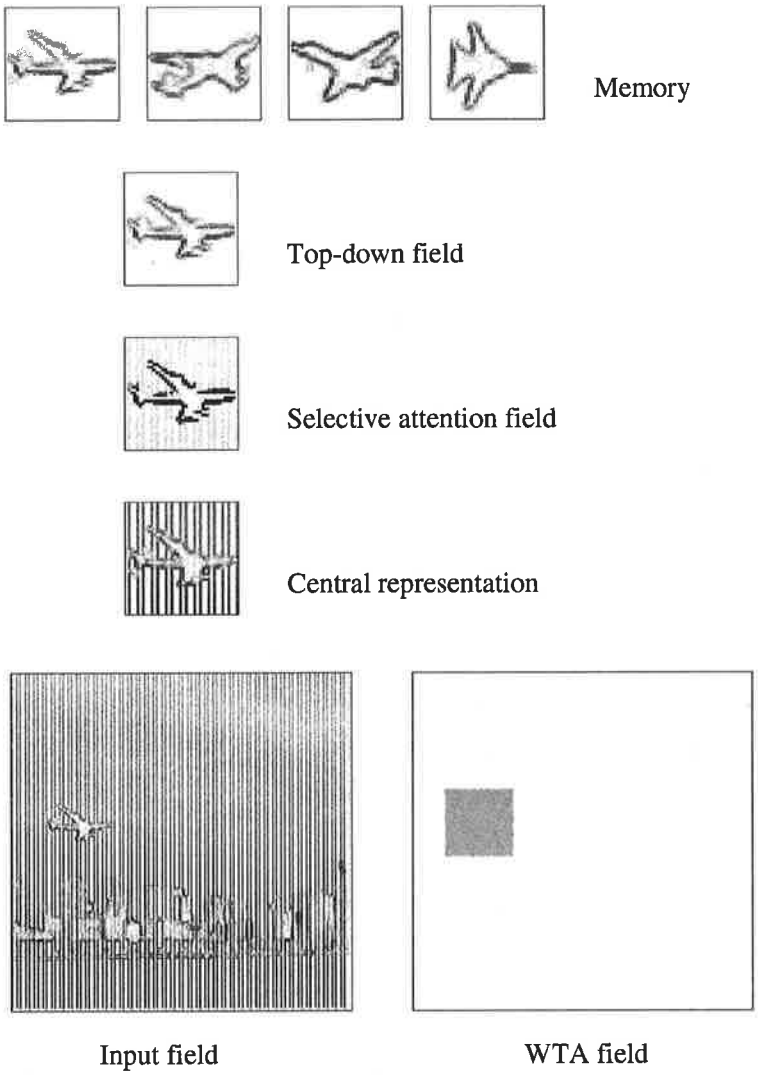


Figure 5.15: Cluttered images: Simulation III results. Simulation results for a highly cluttered scene with minor occlusion. Presynaptic facilitation was applied, improving the degree of match from 0.884 to 0.98. $F_s = 1.6$, and $\bar{G}$ is increased by a factor of 2.5.

region of interest based on some bottom-up elementary features. The preattentive mode directs the attentional focus to the region, allowing a thorough analysis of the region to proceed. The resultant region of interest can be regarded as a window of attention, which we have termed the high activity field. This window of attention is expected to shift to another region of interest upon completing the analysis of the current region of interest.

Three simulations are presented in this section for automatic attention on: i) a simple scene with familiar objects clearly visible; ii) a complex scene with background clutter and a low contrast object; and iii) a complex scene that highlights the importance of the size of the window of attention.

5.5.1 Simulation I

This simulation illustrates the entire process from initial attentional capture, which includes the detection and localisation of a region of interest, to attentional focus processing for achieving object recognition, and finally shifting of attention to another region of interest. The input scene has a very simple background and several familiar objects located randomly within it. The input scene and details of the simulation are shown in Figure 5.16 and Table 5.4.

In order to improve efficiency and speed in attentional capture, as well as modelling the coarse nature of the preattentive mode, the initial parallel sampling for the high activity WTA field is performed sparsely, with overlapping regions separated both horizontally and vertically by N_s pixels. Further discussion on the choice of N_s is provided in Section 5.8.

Table 5.4: Automatic attention: Simulation I details

High activity WTA field size	$N_a \times N_b = 60 \times 60$
Sampling skip	$N_s = 10$ pixels
Vigilance	Primary: $\rho = 0.97$; Secondary: $\varsigma = 0.92$
Gaussian receptive field:	standard deviation $\sigma = 10$
	constant $W_o = 1$
Inhibitory Gaussian receptive field:	standard deviation $\varrho = 25$
	constant $\bar{G}_o = 2$

Refer to Section 4.4 for relevant equations and symbols. Unspecified parameters are as given in previous tables.

With automatic attention, each simulation is expected to produce as many sets of results as the number of familiar objects contained in the input scene. To present results in a more efficient

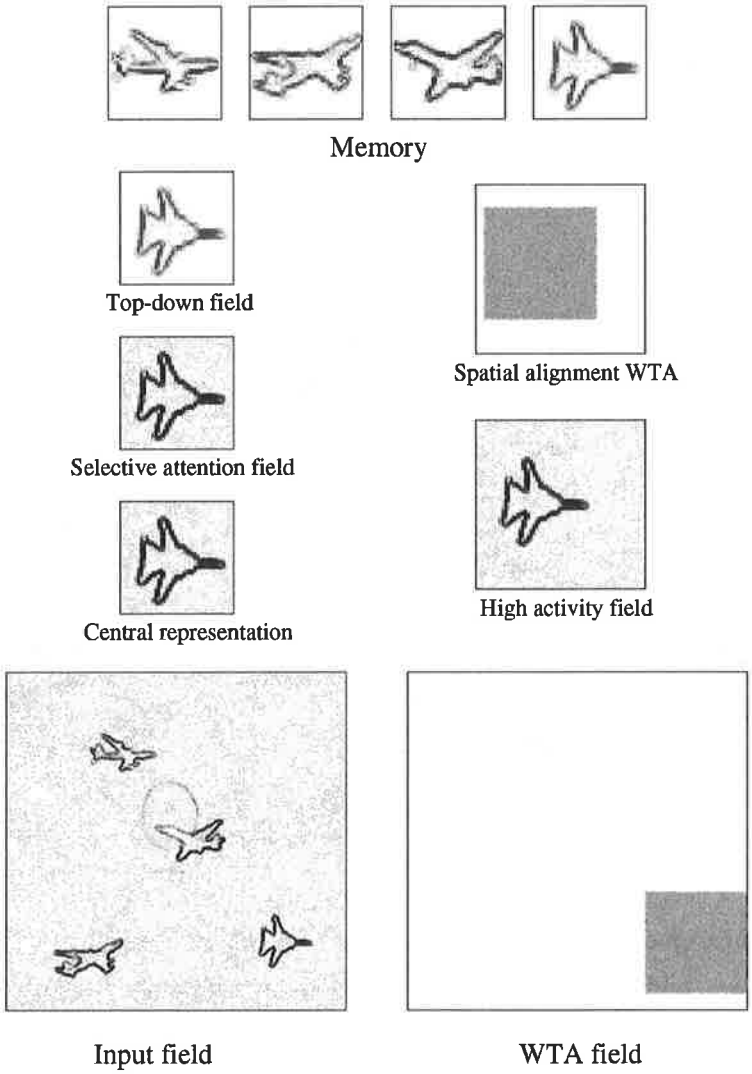


Figure 5.16: Automatic attention: Simulation I - Part 1. Shows the first recognised object, with no presynaptic facilitation, and a degree of match of 0.99.

and compact manner, only the first recognised object is presented in the framework format (shown in Figure 5.16); the rest of the results are presented with the memory and input fields omitted. Since the detection of a high activity region is based on the strength of its edge activity, this means the selected high activity regions may not contain any familiar objects at all.

Figure 5.16 shows that a high activity region is chosen by the high activity WTA field. This defines the location and contents of the window of attention as shown in the high activity WTA field and high activity field. The system then proceeds to recognise the bottom-up object in the usual manner as described in previous simulations. Once the bottom-up pattern is recognised, the system switches its attention to the next region of interest. As a result, the remaining objects in the scene are detected and recognised as shown in Figure 5.17.

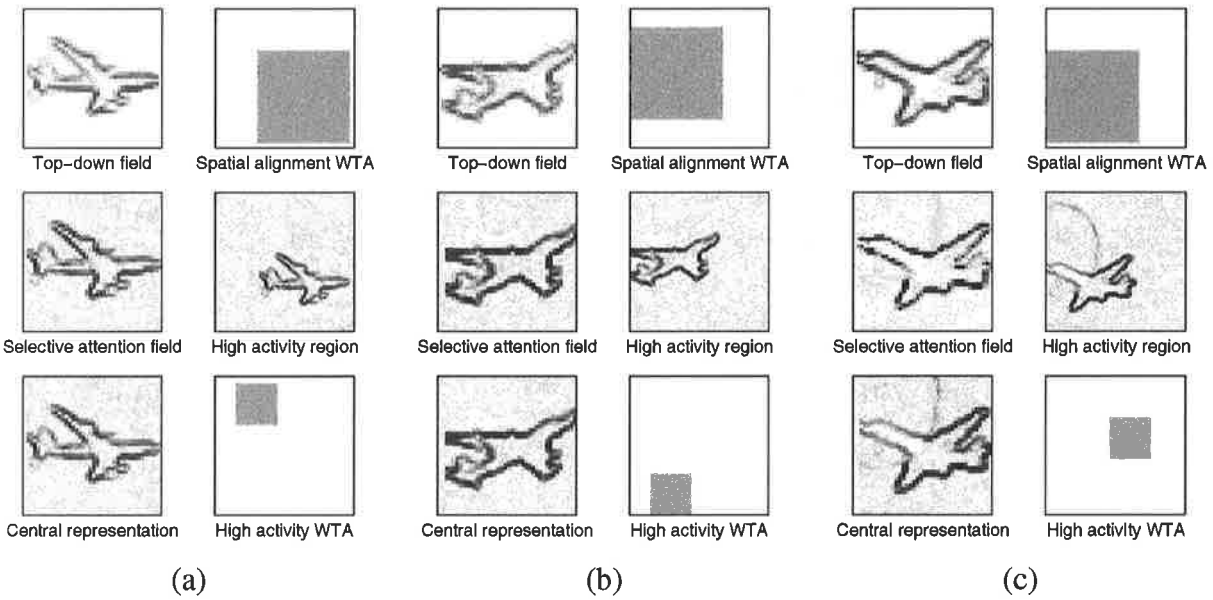
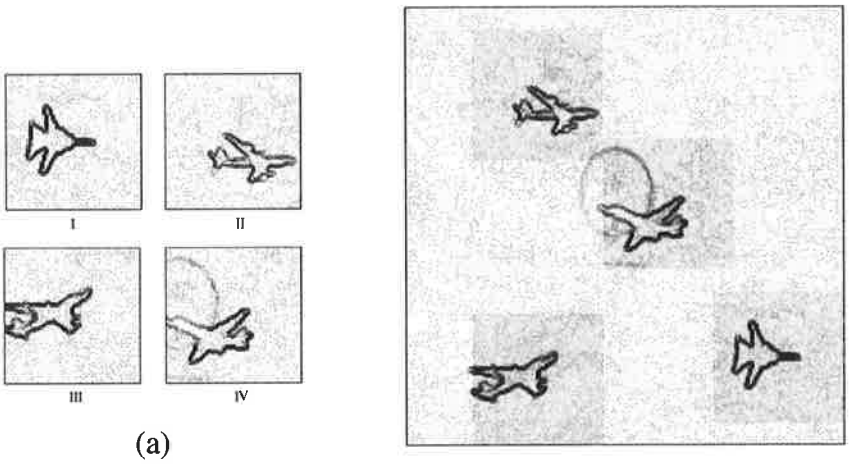


Figure 5.17: Automatic attention: Simulation I - Parts 2, 3 and 4. (a) The second recognised object, with no presynaptic facilitation and a degree of match of 0.98; (b) the third recognised object, with no presynaptic facilitation and a degree of match of 0.99; and (c) the fourth recognised object, with a degree of match of 0.965 prior to presynaptic facilitation and 0.984 after. Note that the patterns are not to scale.

Figure 5.18 highlights the locations of the windows of attention and the order of attentional shift. In particular, Figure 5.18(a) indicates aircraft IV is the first one to capture attention, followed by Aircraft I, Aircraft II and Aircraft III. Since the input scene has a very simple background, the selected windows of attention have all contained a familiar object, but we shall see in the next section this is not always the case with complex visual scenes.

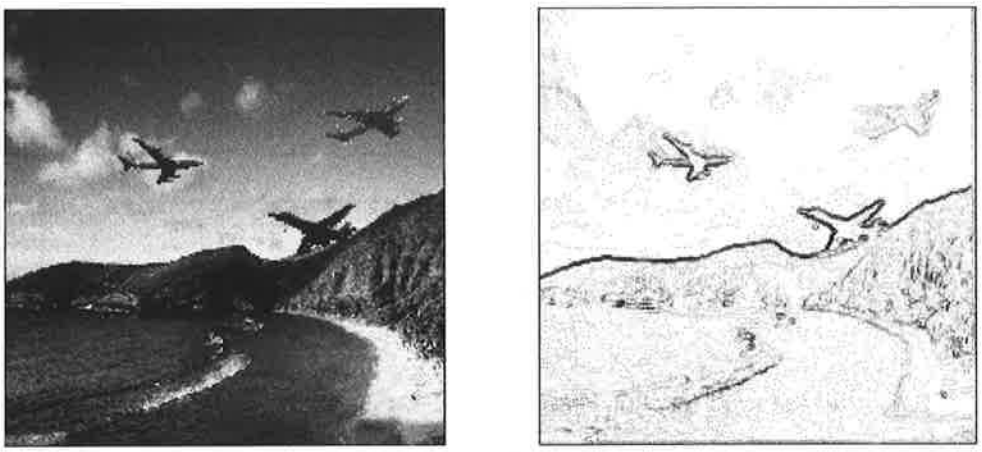


(a) (b)

Figure 5.18: Automatic attention: Simulation I - windows of attention. (a) Selected high activity regions. (b) Spatial locations of the windows of attention.

5.5.2 Simulation II

This simulation shows the elementary feature strength dependence of the preattentive mode for attentional capture. Figure 5.19 shows the input scene used in this simulation. The outline of Aircraft II is barely visible, and the outline of Aircraft III is infringed by its background. Both of these could potentially cause problems for our system. Results of the simulations are shown in Figures 5.20 and 5.21. Note that due to low contrast in the region containing Aircraft II, it takes seven cycles before attention reaches that region.



(a) (b)

Figure 5.19: Automatic attention: Simulation II input scene. (a) Intensity map; (b) edge map.

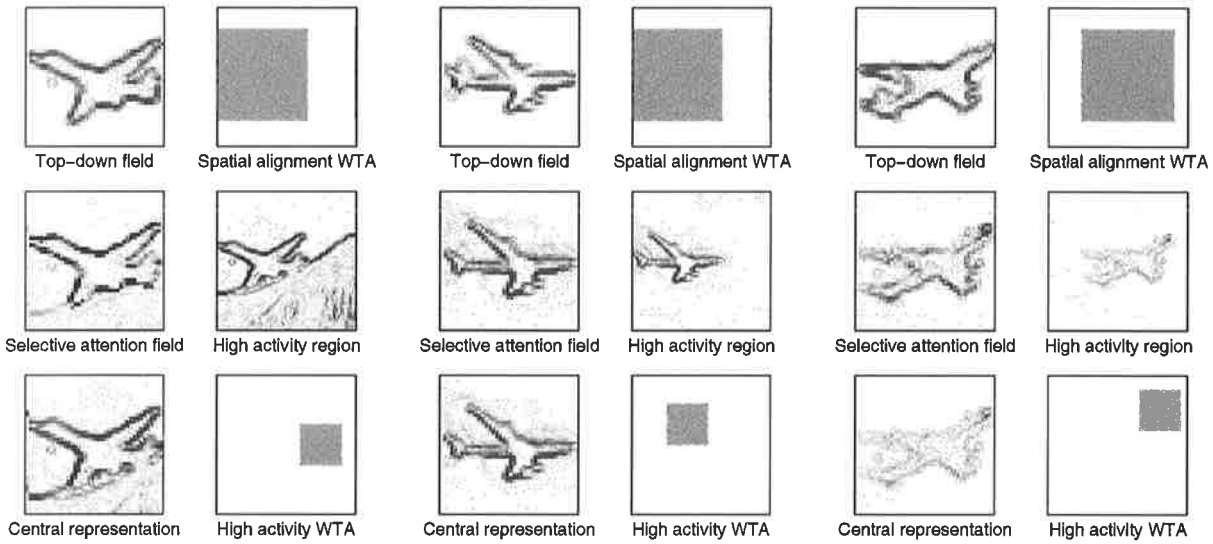


Figure 5.20: Automatic attention: Simulation II - Parts 1, 2 and 3. (a) The first recognised object, with a degree of match of 0.96 prior to presynaptic facilitation and 0.981 after; (b) the second recognised object, with no presynaptic facilitation and a degree of match of 0.98; and (c) the third recognised object, with no presynaptic facilitation and a degree of match of 0.97.

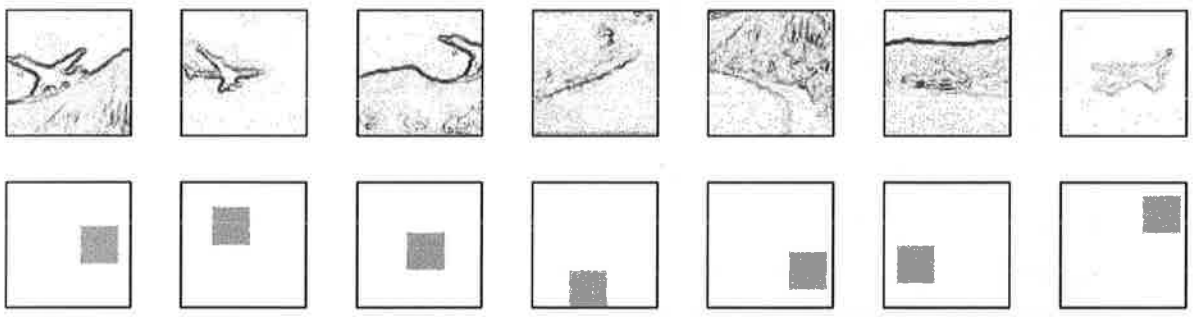
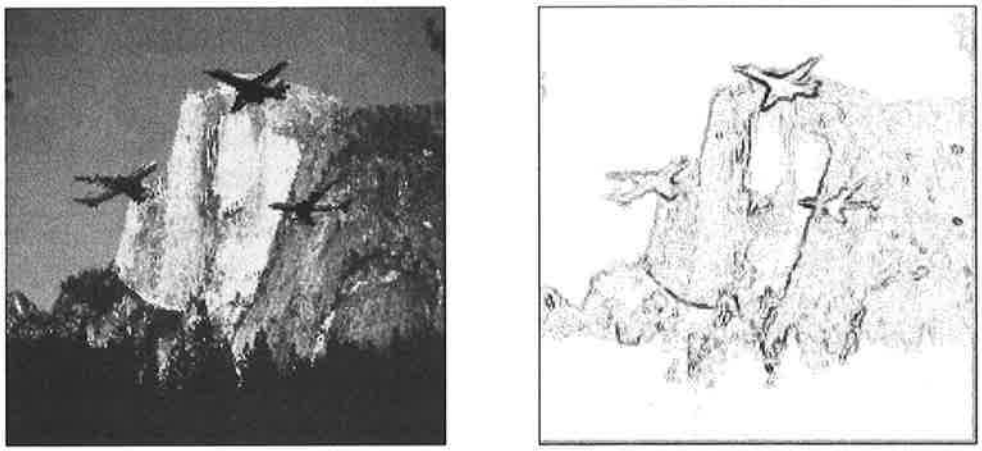


Figure 5.21: Automatic attention: Simulation II - windows of attention. Due to low contrast in the region containing aircraft II, it takes seven cycles before attentional capture is achieved over that region. Columns 1, 2 and 7 contain familiar objects. The WTA fields show the order of attentional shift.

5.5.3 Simulation III

This simulation demonstrates what might happen if the size of the window of attention is too small, and when a familiar object is near a region of high contrast. In both cases, it is possible that the target object is partially included in the high activity field, which causes the system to fail. Figure 5.22 shows the input scene and edge map used in this simulation. Results of the simulation are shown in Figures 5.23 and 5.24.



(a) (b)

Figure 5.22: Automatic attention: Simulation III input scene. (a) Intensity map; (b) edge map.

While the above situation is undesirable, we do come across similar situations in our everyday life. We tend to get attracted to novel events and objects, and it is very easy to overlook items that are less attractive, even though they are adjacent to where we have been paying attention.

One possible solution to this problem is to enlarge the size of the window of attention to provide an extra tolerance to familiar objects that are not being captured at the centre of the window of attention. Currently, the high activity field size is 1.5 times that of the object size along a single dimension, therefore it would be reasonable to use a high activity field size that doubles the object size along a single dimension. However, for processing efficiency and speed the high activity field size should be kept as small as possible. Further discussions on the high activity field size is given in Section 5.8.

After changing the high activity field size from 60×60 to 80×80 pixels, the proposed system is now able to recognise Aircraft I as shown in Figure 5.25. The proposed recognition system demonstrates its robustness in this case, as the size of the window of attention can be changed as desired without affecting the operation of the system. Note that for subsequent simulations the high activity field size is still 60×60, unless stated otherwise.

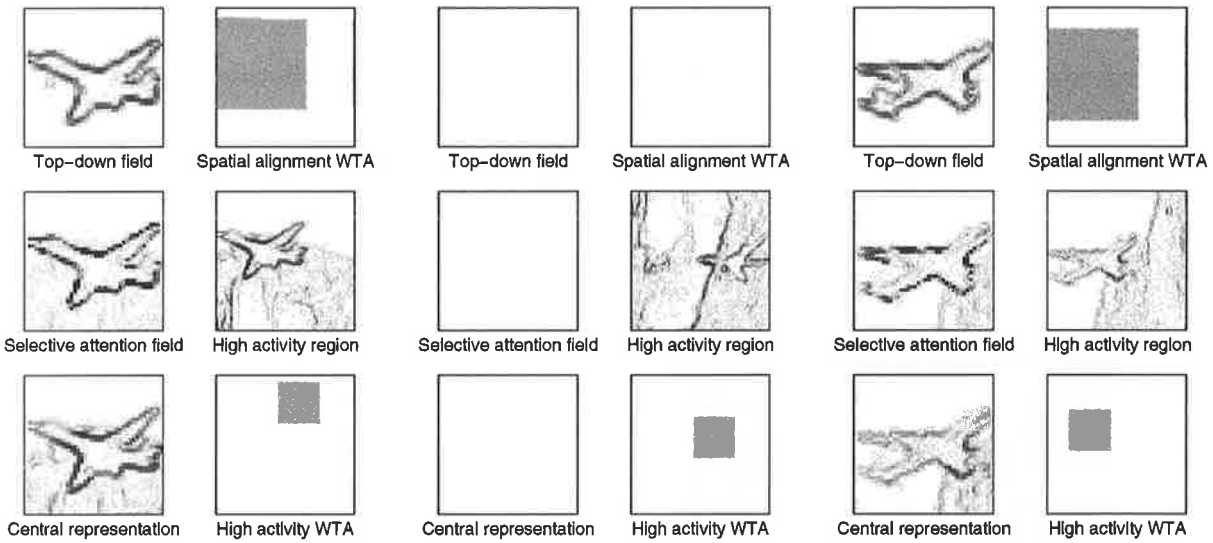


Figure 5.23: Automatic attention: Simulation III - Parts 1, 2 and 3. (a) The first recognised object, with a degree of match of 0.966 prior to presynaptic facilitation and 0.982 after; (b) no familiar object detected due to the object not completely within the window of attention; and (c) the second recognised object, with a degree of match of 0.956 prior to presynaptic facilitation and 0.973 after.

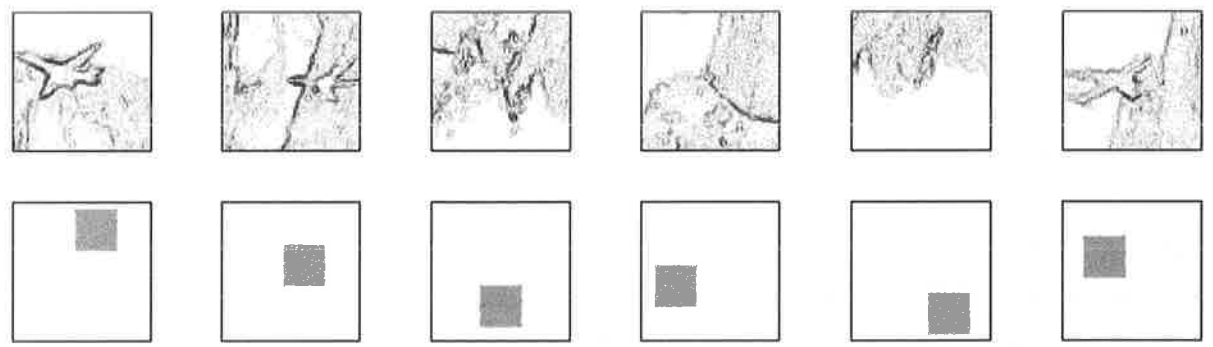


Figure 5.24: Automatic attention: Simulation III - windows of attention. High activity region 2 failed due to the missing tail of aircraft I. Note that in this simulation aircraft I is in the reversed direction and to counter that a reversed aircraft I has been learned.

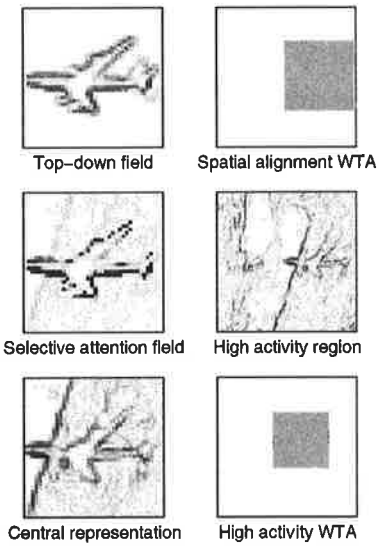


Figure 5.25: Automatic attention: Simulation III results with new window of attention size. The above high activity field size is 80×80, thus allowing the system to overcome problems encountered in Figure 5.23.

5.6 Rotation Invariance

It is common to see everyday objects in orientations other than their upright positions, thus it is very important for a recognition system to be able to recognise rotated objects. In this section, we present simulations for the model developed in Section 4.5, which is capable of recognising all familiar rotated objects located anywhere within a complex cluttered visual scene.

Rotation invariant recognition is achieved by transforming the window of attention, via mental rotation, into a number of competing parallel frames of reference. The winning reference frame is aligned with another perceptual frame of reference, the central representation, for matching.

Two simulations are presented for rotation invariance. In both cases, familiar objects are rotated and placed randomly within a complex cluttered environment. In the first simulation, results are presented in the framework format showing all the rotational templates derived from the bottom-up object location. In addition, STM patterns are also provided for Simulation I, while Simulation II results are presented in the compact format.

5.6.1 Simulation I

Results of Simulation I are summarised in Figures 5.26, 5.27 and 5.28. In each of the figures, we can see that a region of interest is first detected and located from the input field. From the resultant high activity field a large number of parallel frames of reference are generated. These

rotational templates are then gated by top-down memory patterns, generating synaptic inputs to the rotational and spatial alignment WTA field. A winner is produced from the WTA field, allowing us to determine the location and orientation of a potentially recognisable object, as well as the most likely match in memory. The chosen bottom-up and top-down patterns are compared. The outcome of which decides whether top-down presynaptic facilitation is required to improve matching.

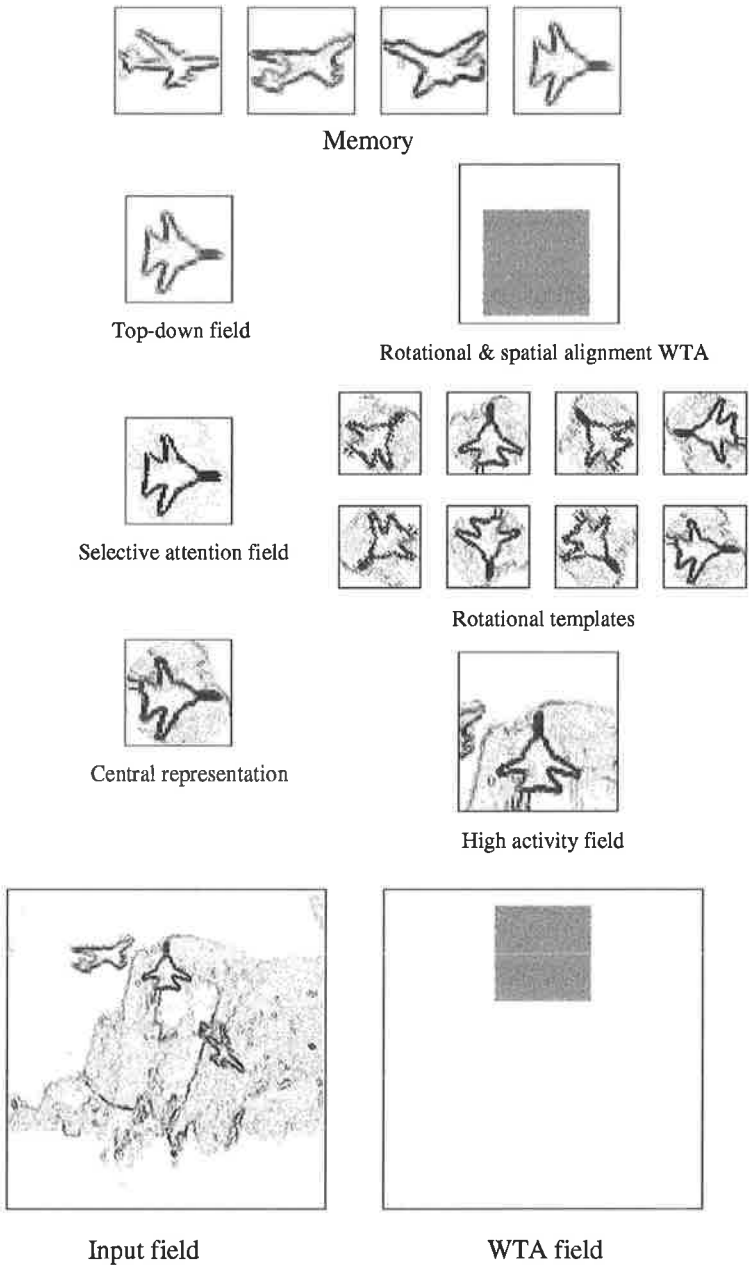


Figure 5.26: Rotation invariance: Simulation I - Part 1. The first recognised object, with a degree of match of 0.944 prior to presynaptic facilitation and 0.990 after.

In addition to the main results, STM patterns for each of the recognised objects are shown in

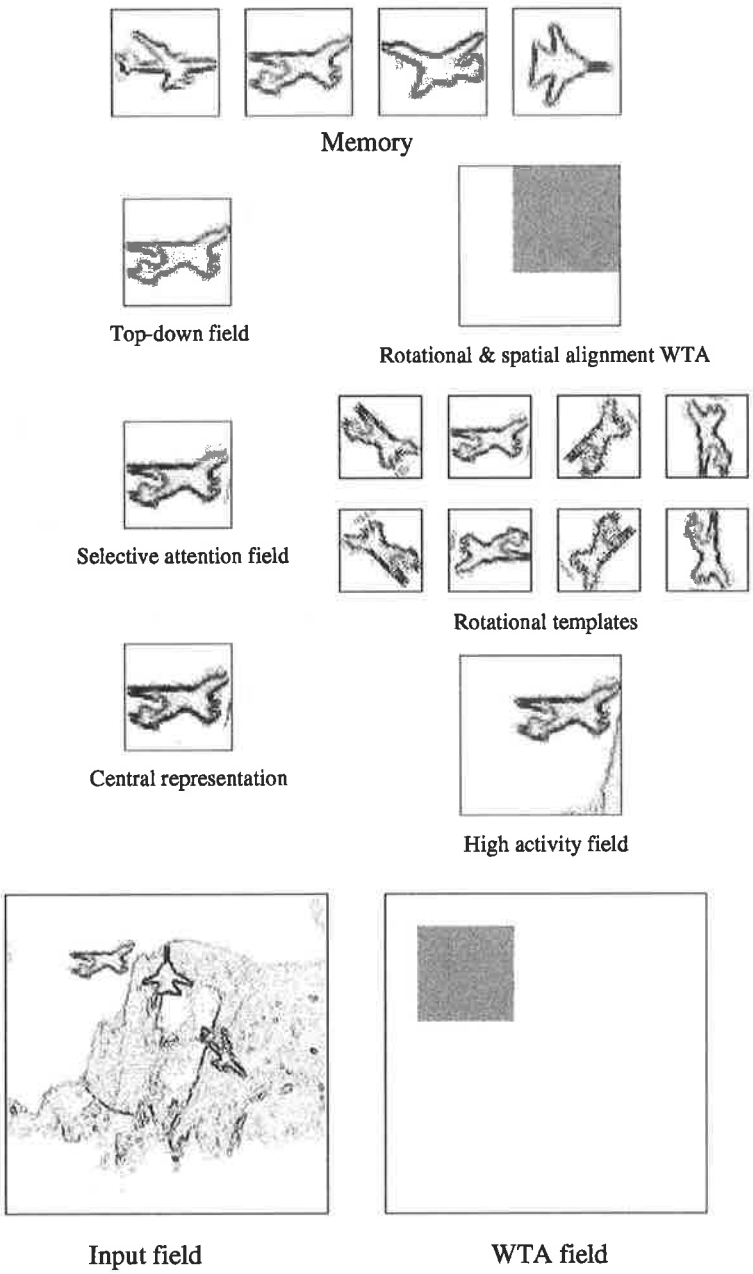


Figure 5.27: Rotation invariance: Simulation I - Part 2. The second recognised object, with no presynaptic facilitation and a degree of match of 0.993.

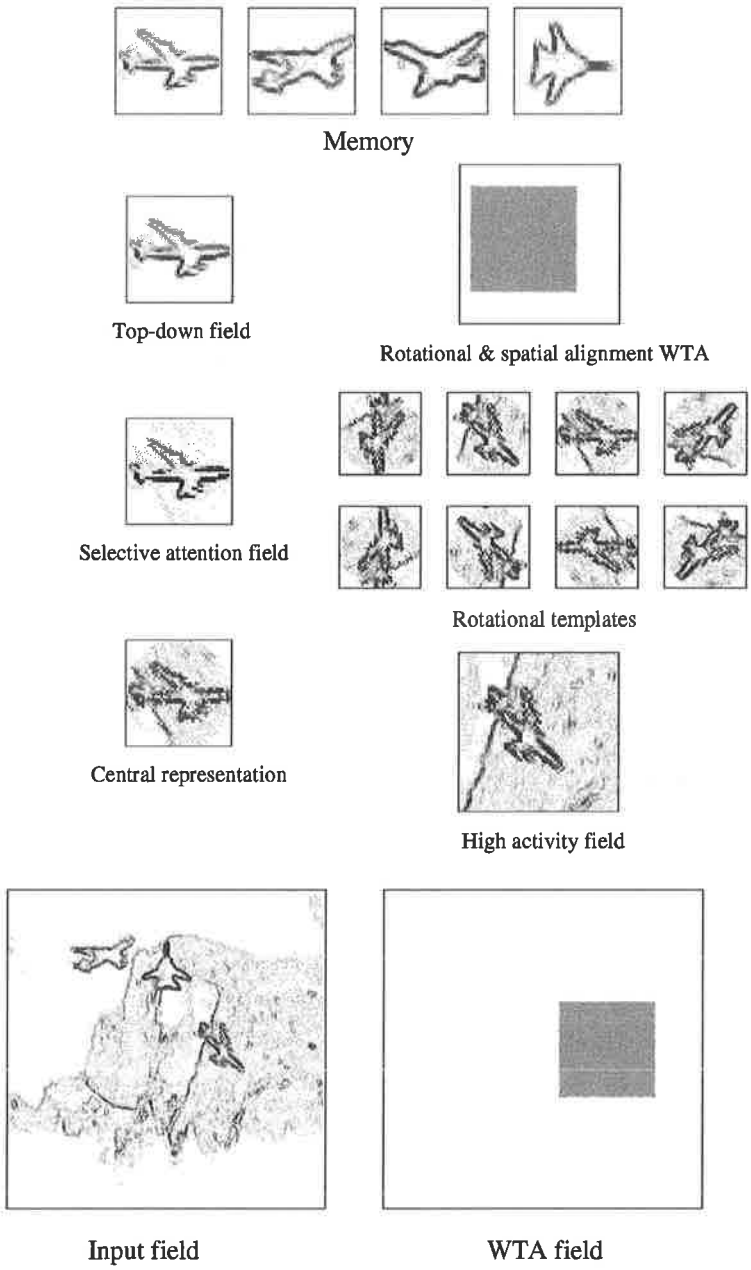


Figure 5.28: Rotation invariance: Simulation I - Part 3. The third recognised object, with a degree of match of 0.961 prior to presynaptic facilitation and 0.989 after.

Figure 5.29. It demonstrates the use of STM loops as a stabilising mechanism. Most importantly, it illustrates how a bottom-up pattern is transformed as it progresses upwards.

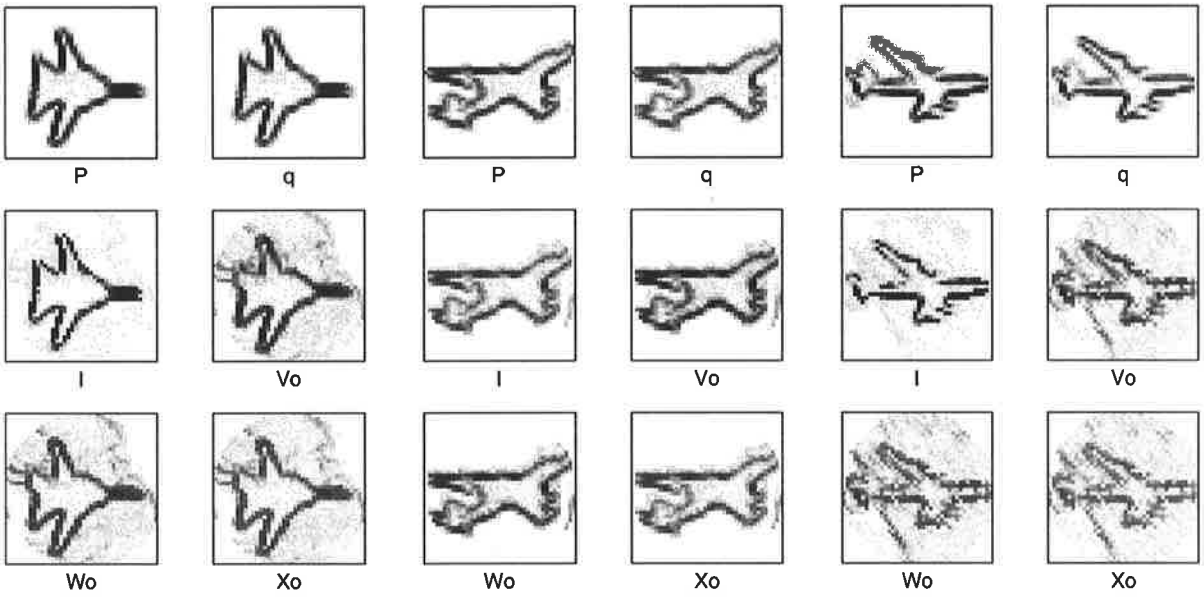


Figure 5.29: STM patterns for rotation invariance simulation I. (a) First object; (b) second object; and (c) third object.

The recognised objects need not be the first detected high activity regions. However, it is pointless to show regions that report no recognisable objects. On the contrary, any false alarm is reported.

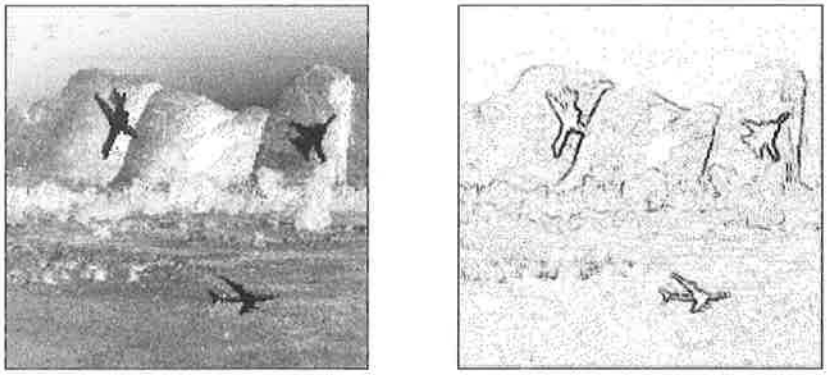
5.6.2 Simulation II

Simulation II is another example of rotation invariant object recognition by the proposed system. It also has a very complex background which can be seen in Figure 5.30.

All recognised objects are summarised in Figure 5.31. Rotational templates have been omitted, since they are the same as the central representation depicted in various orientations.

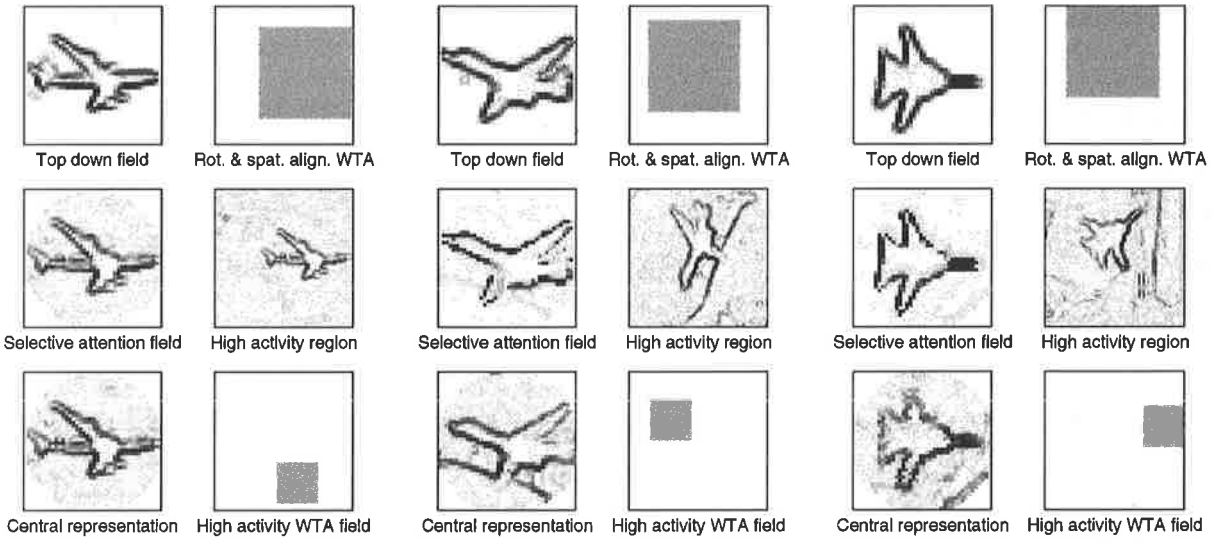
5.7 Distortion Invariance

The framework simulated in this section represents the complete model for the proposed system, capable of performing translation, rotation and distortion invariant object recognition in complex visual scenes with automatic attentional capture and shift.



(a) (b)

Figure 5.30: Rotation invariance: Simulation II input scene. (a) Intensity map; (b) edge map.



(a) (b) (c)

Figure 5.31: Rotation invariance: Simulation II - Parts 1, 2 and 3. The first recognised object, with no presynaptic facilitation and a degree of match of 0.983; (b) the second recognised object, with a degree of match of 0.935 prior to presynaptic facilitation and 0.979 after; (c) the third recognised object, with a degree of match of 0.925 prior to presynaptic facilitation and 0.987 after.

Two simulations have been devised to highlight the distortion invariant property of the system. In particular, the use of band transformed memory patterns to detect and locate arbitrarily distorted 2D object patterns, and the use of shape attraction to reshape the distorted patterns into a recognisable form. The two cases considered are: i) a clear background visual scene; and ii) a complex cluttered background visual scene. Additional results are provided in Appendix A.

Distorted visual scenes used for the simulations are generated using computer graphics tools from normal scenes. The software package used is GIMP[1]. The distortion effect is achieved through the use of the **waves distort filter** under GIMP, which simulates the effect of throwing a stone in a pond.

5.7.1 Simulation I

Consider the input field in Figure 5.32. It contains three randomly placed objects on a clear background. As the objects are quite severely distorted, it would present some degree of difficulty even for humans to visually recognise them against the learned memory patterns.

Encouraging results have been achieved using the proposed recognition system to deal with distorted objects, and are shown in Figures 5.32 and 5.33. From Figure 5.32, we can see that the system operates in very much the same way as before. It begins by detecting and locating a region of interest, from which rotational templates are generated to activate a stored model in memory for comparison with the bottom-up pattern. It is only when the matching fails, that the system considers whether to use shape attraction to reshape distorted patterns to a recognisable form. If the conditions for shape attraction are met (see Section 4.6), then band transformed memory patterns, as shown in the top of Figure 5.32, are used in place of normal memory patterns for bottom-up memory activation and initial matching. This enables the system to determine the spatial location of a potentially recognisable object, and thus shape attraction can be applied.

Figure 5.33(a) shows that it may not be necessary to apply shape attraction to recognise distorted objects. A distorted version of Aircraft I was recognised using only top-down presynaptic facilitation. Of course, this is only possible if the distorted object is sufficiently similar to the original object as in Figure 5.33(a), while in Figure 5.33(b), the object can only be recognised using shape attraction.

[1]This program is available at www.gimp.org

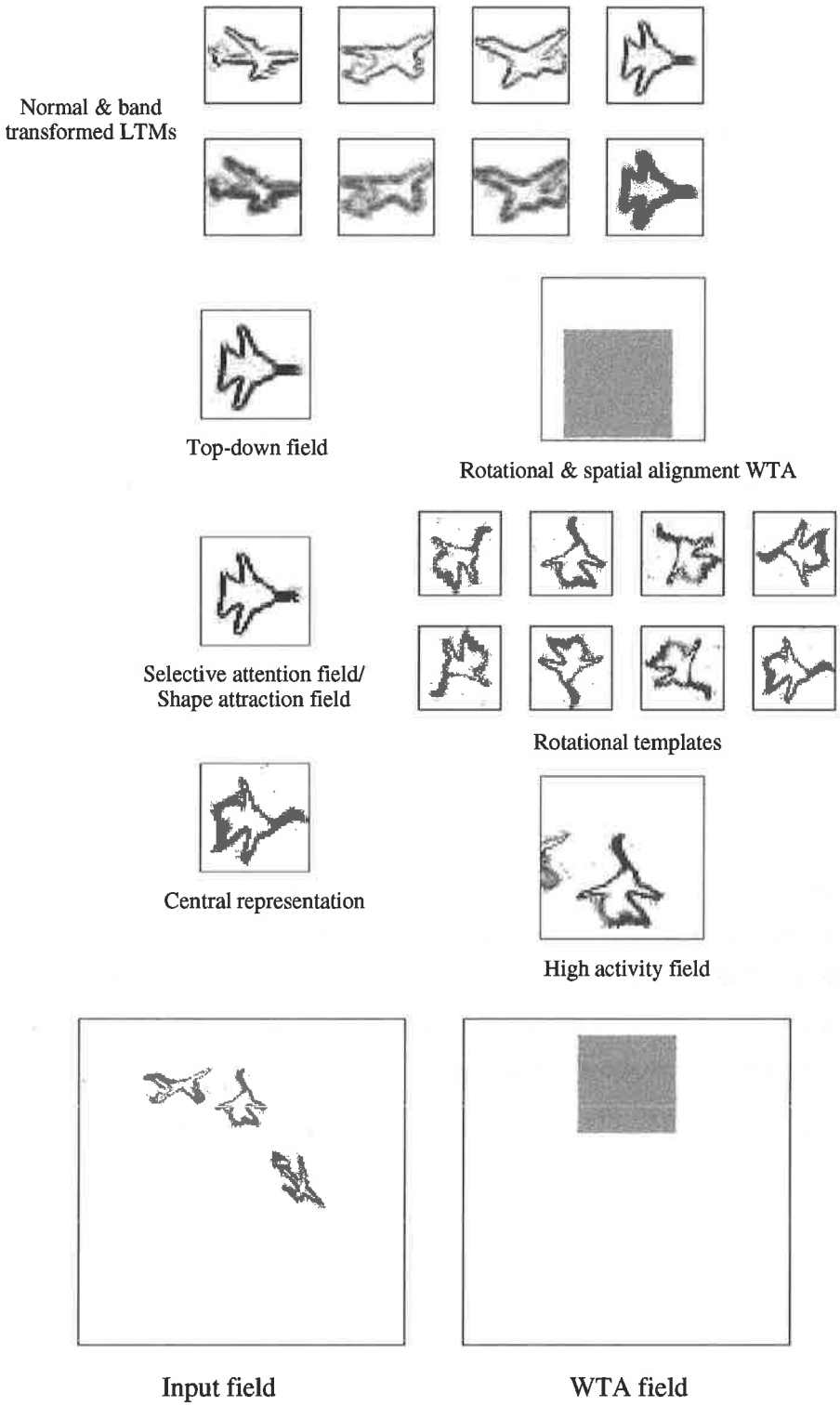


Figure 5.32: Distortion invariance: Simulation I - Part 1. The first recognised object, with a degree of match of 0.94 prior to shape attraction and 0.992 after.

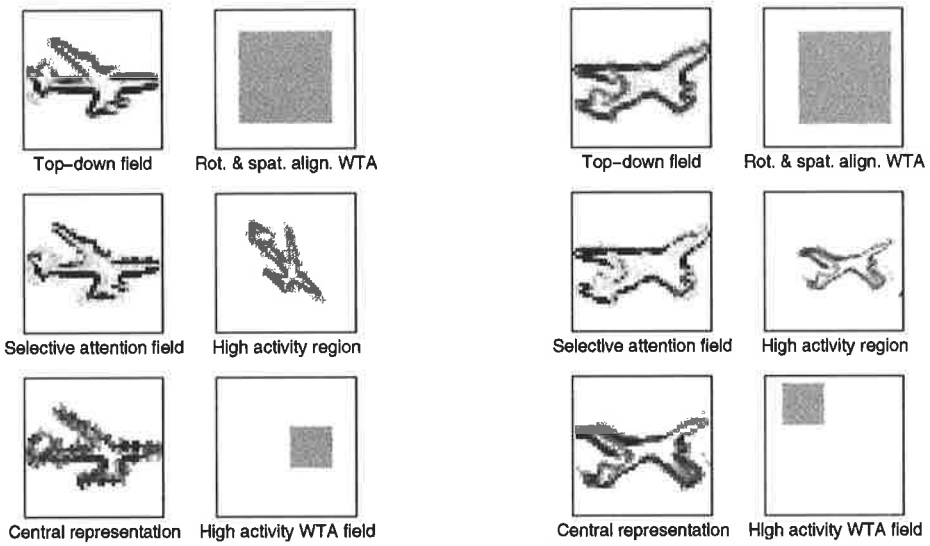


Figure 5.33: Distortion invariance: Simulation I - Parts 2 and 3. (a) The second recognised object, with a degree of match of 0.964 prior to presynaptic facilitation and 0.988 after; and (b) the third recognised object, with a degree of match of 0.941 prior to shape attraction and 0.988 after.

5.7.2 Simulation II

Performing distortion invariant object recognition under complex cluttered environments is a much more challenging problem than the one with a clear background as in Simulation I. This can be attributed to the fact that background clutter can be mistakingly treated as part of a distorted pattern. Inevitably, this increases the false alarm rate of the system, because a meaningless pattern could be interpreted as a familiar object pattern.

Figure 5.34 shows the input scene for this simulation. It has basically the same distorted aircrafts as before, except it now has a complex cluttered background. With the experience of Simulation I, we would expect Aircraft II to be recognised without much trouble, however it is not so easy to predict whether the other two aircrafts would be recognised.

From the simulation, four recognised objects were reported by the system. The four sets of results are shown in Figures 5.35 and 5.36. Three out of the four sets have correctly identified its bottom-up object, the remaining one confirms our earlier prediction that complex background scenes can dramatically increase the false alarm rate of the system. Due to the vortex nature of shape attraction, care must be exercised in choosing the vigilance parameters and other associated parameters. If the vigilance parameters are set too high, then we risk lowering the recognition rate. On the other hand if too low, a high false alarm rate is likely to occur.

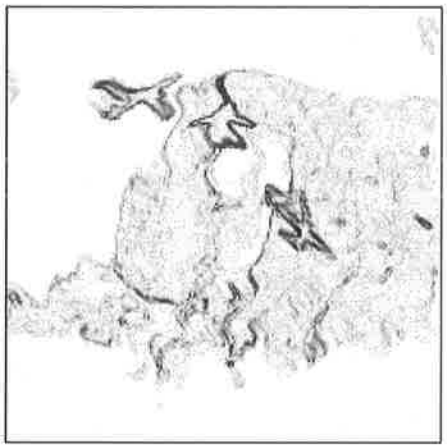


Figure 5.34: Distortion invariance: Simulation II input scene.

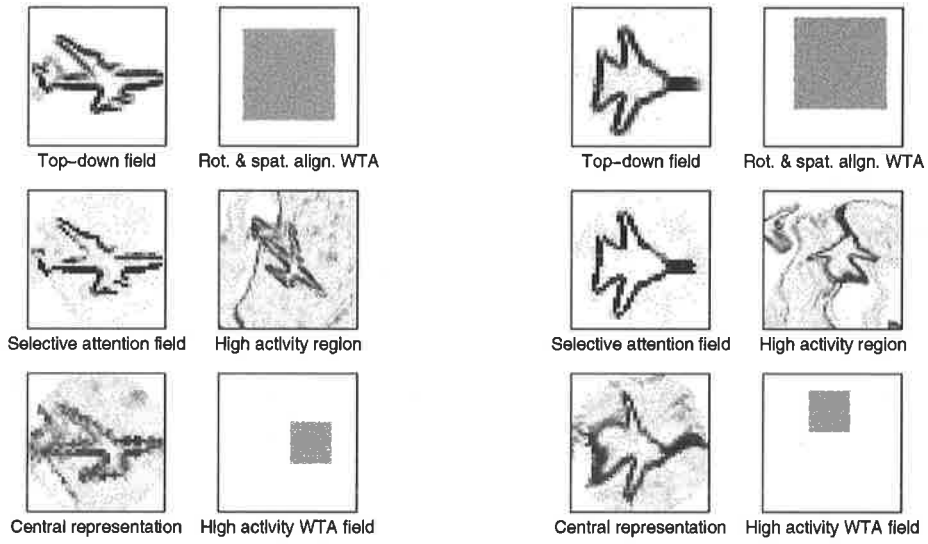


Figure 5.35: Distortion invariance: Simulation II - Parts 1 and 2. (a) The first recognised object, with a degree of match of 0.957 prior to presynaptic facilitation and 0.989 after; and (b) the second recognised object, with a degree of match of 0.931 prior to shape attraction and 0.993 after.

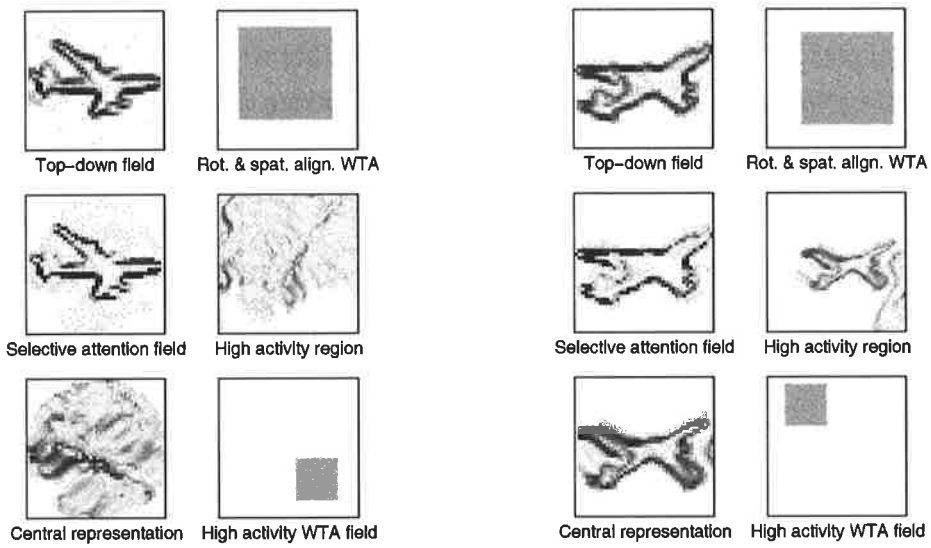


Figure 5.36: Distortion invariance: Simulation II - Parts 3 and 4. (a) false alarm; (b) the third recognised object, with a degree of match of 0.950 prior to shape attraction and 0.989 after.

The proposed system obtains a recognition rate of 90% and a false alarm rate of 27% when simulated using 8 distorted input scenes with a total of 24 distorted familiar objects. The high false alarm rate is indicative of the difficulty associated with the recognition of distorted patterns in complex cluttered scenes. Some of these results are provided in Appendix A. It should be pointed out that the rates are highly dependent on the values used for the primary and secondary vigilance parameters. In the simulated cases, these were determined empirically to achieve the best recognition rate. A fixed choice for the vigilance parameters would produce different recognition and false alarm rates. The above figures represent the optimal rates achieved by the proposed system.

Note that the high activity regions shown in Figures 5.35 and 5.36 are not necessarily the first four regions of interest, but the four that the system reported containing familiar objects.

5.8 Design of System Parameters

The proposed visual scene analysis system requires a considerable number of parameters to perform object recognition effectively and efficiently. As with most engineering systems there is not a single set of parameters that will work in all situations. For that reason, adjustments are often required for complex scenes. Some of the major parameters that require frequent adjustments are discussed in this section.

There are many factors that may affect the choice of system parameters, some of which are based on personal preference, and so can be arbitrary but within a reasonable range. Others are mostly image and performance dependent, such as the input scene and object sizes, the nature of the image background, illumination conditions and contrast levels, locations of target objects, and false alarm rates, etc.

The best way to illustrate the design process is through an example. In this section we provide a design that is used in our real-world imagery simulations in the next section, in which real-world scenes captured by a digital camera are analysed by the proposed system.

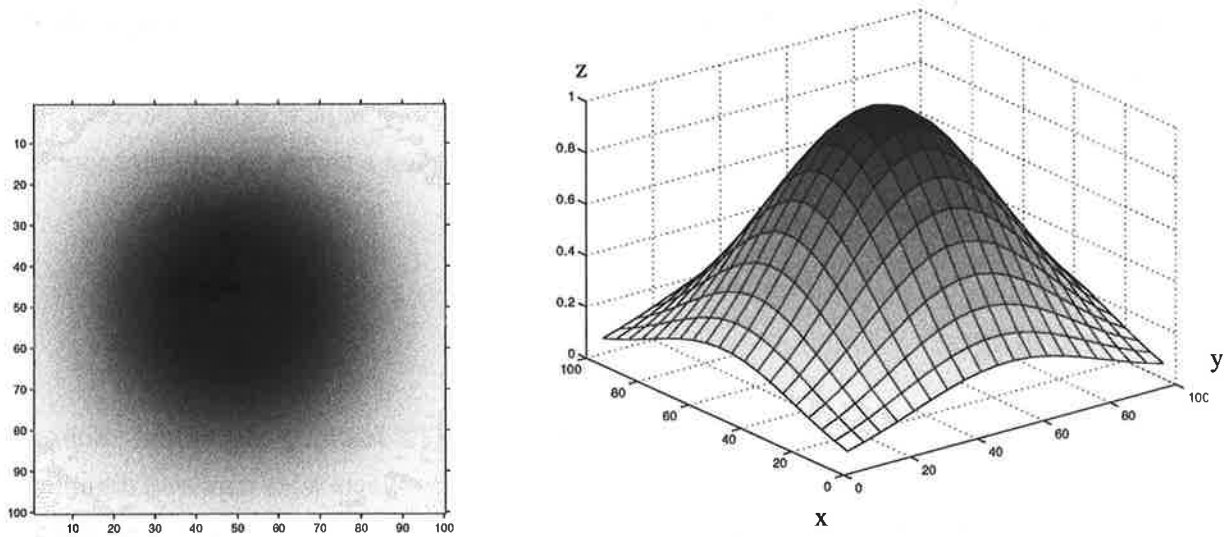
Beginning with the size of the object fields, $N_p \times N_q$, the enclosed object pattern should occupy as much space as possible in order to minimise potential problem with objects near image borders. Once we have assigned a convenient size, e.g., the minimum size that maintains a reasonable resolution, we can train the system with objects of the same size. In the real-world imagery case, we let $N_p \times N_q = 64 \times 64$. As in Section 5.2, learning and STM equations are based on ART2 neural networks [29], therefore parameters required in learning can be derived accordingly.

With the lack of explicit size invariant mechanism, the target objects contained in the visual scene must be similar in size to those learned (minor size variations are allowed). The input scene is scaled so that its objects approximately match those in memory. Due to changes in camera angle, height and position, visual scenes captured by the camera will have slightly different sizes after scaling, resulting in $N_i \times N_j = 208 \times 166$ to 250×200 pixels, as given in Section 5.9. Next the size of the high activity field must be chosen. Obviously, this field must be large enough to include the entire target object and yet be kept as small as possible to reduce processing time. It has been found that a high activity field size that is 1.5 times of the object field's size is a good starting point, thus we let $N_a \times N_b = 100 \times 100$ pixels. Since the initial attentional capture is based purely on contrast (edge information), it is possible for objects situated near strong clutter to be partially included in the high activity field, which of course leads to non-detection, thus no recognition is possible. To provide more tolerance we may set the high activity field size to twice the object field size at the expense of additional computational time. Ways to improve the robustness of the automatic attention stage are proposed in Chapter 7.

Coarse sampling is generally acceptable with a large window of attention. It allows the preattentive mode to process information rapidly, and therefore capture attention at a relatively short time. But as pointed out above a large window of attention is very inefficient, so a balance between preattentive sampling and focussed processing must be reached in order to optimise the system performance. It has been empirically determined that a sampling skip, N_s, between 5 to 10 pixels is effective in achieving attentional capture at a relative fast speed. It should be noted that a large sampling skip can result in the same problem as a small window of attention

- the non-detection of targets due to partial inclusion. Thus, it is advisable to begin with a small sampling skip first.

Another quantity that needs to be selected carefully for attentional capture is the Gaussian receptive field size, i.e., its standard deviation σ. This determines the origin of each region of interest, therefore it is particularly important to get this right. Given $N_a \times N_b = 100 \times 100$ and $N_p \times N_q = 64 \times 64$, the ideal scenario would be to have detected a region of interest containing a familiar object at the centre of the window of attention. For that to happen, we need to set σ such that the receptive field draws contribution mainly from the centre 64×64 pixel area of the high activity region. An intuitive way to choose σ is to use graphical means to display the receptive field as shown in Figure 5.37. By inspecting the Gaussian profile of the receptive field, we can see that it covers most of the centre 64×64 area. However in extremely cluttered visual scenes, re-adjustment of the receptive field size may be needed.



(a) (b)

Figure 5.37: Graphical parameter determination. Region of coverage by a receptive field with a standard deviation $\sigma = 40$. (a) Effective region in 2D; (b) in 3D.

The graphical approach can be applied equally well to the design of the post-attention inhibitory signal, since it also has a Gaussian profile. The size of the inhibitory signal should be smaller than the receptive field so as not to accidentally suppress information from neighbouring regions or closely located objects. For the real-world imagery simulations in Section 5.9, it has been empirically determined that a value between 15–35 for ϱ in (4.31) is effective.

The choice of the primary and secondary vigilance parameters is arbitrary in nature, and usually depends on the nature of the input, or the outcomes of a preliminary simulation which provides indications on the adjustment that is required. In general, increasing vigilance parameters low-

ers the false alarm rate but it compromises the detection rate. Useful ranges are 0.96–1.00 and 0.92–0.96 for the primary and secondary vigilance, respectively.

In certain cases, it is desirable to increase the effect of top-down presynaptic facilitation. There are two ways in which this may be achieved. First, we can simply increase the facilitatory signal as shown in Figure 5.13. The other is by increasing the competition, $\bar{G}$, between cells so that cells without facilitation are suppressed more readily. The rest of the parameters used in the PMCNL are based on SAART, which was covered in Chapter 3.

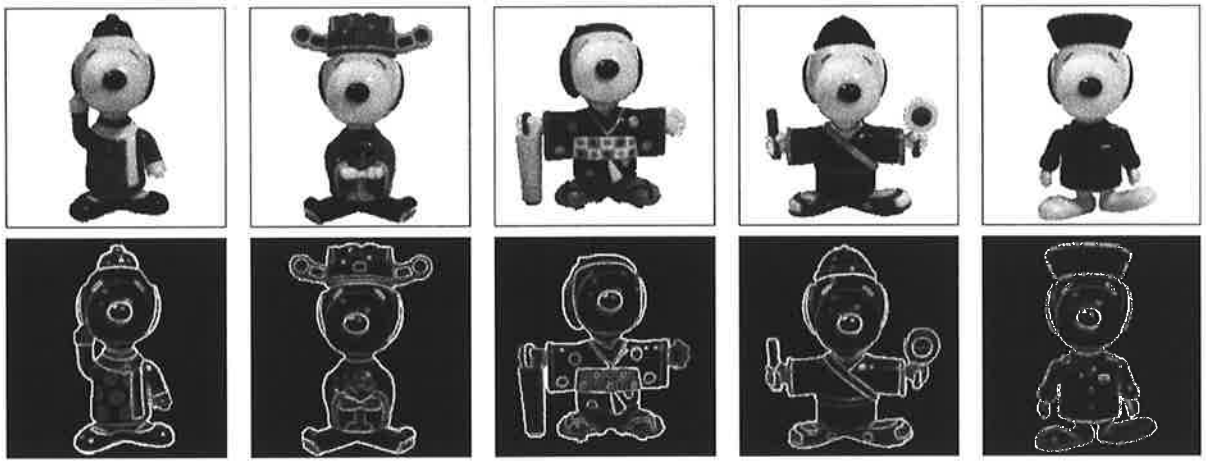
5.9 Real-World Imagery Simulations

In this section, three real-life scenes captured by a digital camera are simulated by the proposed system. The input scenes were designed to highlight specific properties of the system. In particular, they allow the system to exhibit translation and rotation invariances, automatic attentional capture and shift, and recognition in cluttered environments. Although it is preferable to simulate real distorted objects as well, distortions on rigid real-life objects cannot be obtained easily, thus we have adopted the same approach as before by using a computer graphics tool to artificially distort the scenes.

5.9.1 Learning

Five input objects were learned by the system using the ART2 learning algorithm. Subsequently, real-life complex cluttered scenes featuring these input objects were captured digitally. A number of preprocessing steps must be performed before an input object can be learned. First, each object was captured against a clean background. The object image was then resized and cropped to an appropriate size, followed by grayscale conversion and edge detection. In certain cases where illumination conditions are inconsistent, contrast enhancement was necessary in order for the important features to become visible. Similar steps were applied to the scenes to transform them to the appropriate edge map input format.

Figure 5.38 shows the input objects for the simulations. The input objects are toy figures of "Snoopy" dressed in a variety of national costumes. For easy reference, the figures have been labelled as "Snoopy China", "Snoopy Hongkong", "Snoopy Japan", "Snoopy Korea", and "Snoopy Russia", respectively. These objects were chosen because they all belonged to the same class of object having a very similar body shape and sharing many features. Hence, under a complex cluttered environment, they pose a challenging problem to our visual scene analysis system.



(a) (b) (c) (d) (e)

Figure 5.38: Real-world imagery simulation input objects. (a) Snoopy China. (b) Snoopy Hongkong. (c) Snoopy Japan. (d) Snoopy Korea. (e) Snoopy Russia.

The edge maps of the input objects are also shown in Figure 5.38. These edge maps were learned by the ART2 algorithm and the results of the learning process are shown in Figures 5.39 and 5.40. As for the aircraft simulations, during the learning process, an uncommitted node is selected, whose weight vector is adapted over time as shown in Figure 5.40. Figure 5.39 indicates the fluctuation of the degree of match for the duration of the learning process. The vertical dotted lines are instances where a new input object is presented. While the horizontal line is the vigilance parameter level.

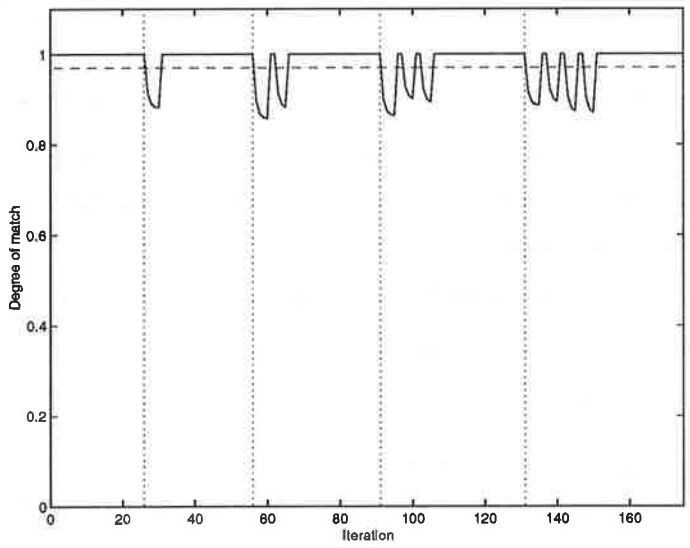


Figure 5.39: Degree of match during learning of real-world input objects. Vertical dotted lines are instances where a new input object is presented.

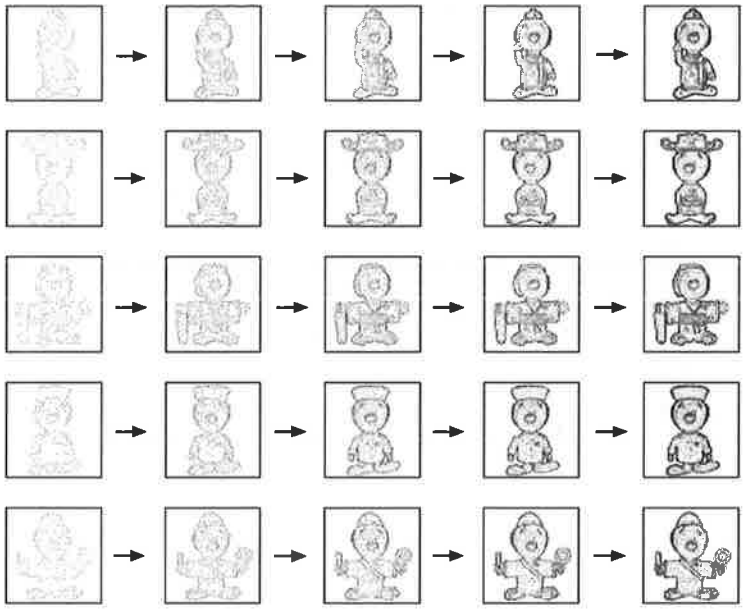
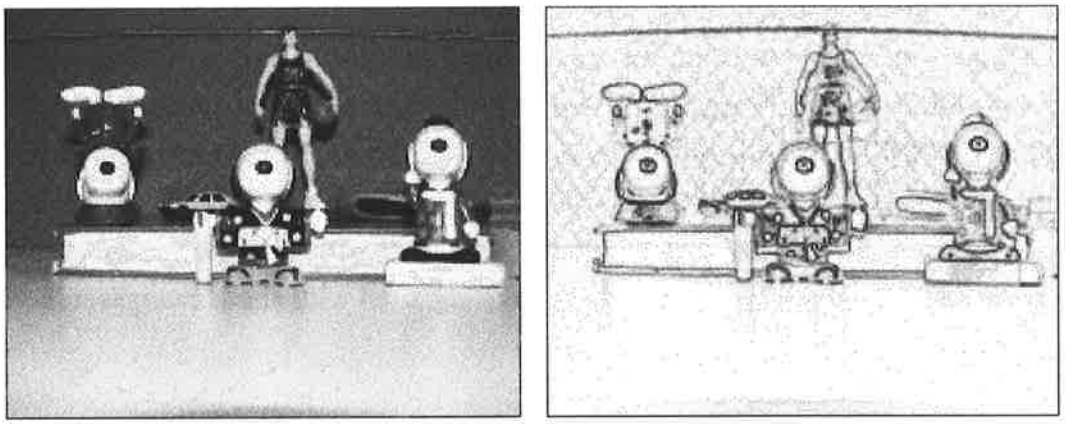


Figure 5.40: Adaptation of LTM weight patterns. LTM weight vectors are strengthened during the learning phase.

5.9.2 Simulation I

The first real-world visual scene to be simulated, shown in Figure 5.41(a), features three learned objects that are placed randomly on a cluttered background containing several everyday items: a toy car, a toy basketball player figure, a book, a mobile phone and an eraser. Two of the three learned objects are erected in their upright positions, and the remaining object is placed upside down. In order to recognise all the familiar objects in the input scene, our system must be able to detect, locate and identify them correctly in spite of changes in position and orientation, and influences from background clutter. The original scene was captured in an ordinary office environment using an *Olympus* digital camera, model C-1400L, as a 640×512 24-bit true colour image, with items arranged as shown in Figure 5.41(a). Due to the lack of size invariance mechanism in the system, the input scene was scaled to an appropriate size such that the objects contained are approximately the same as those learned. Further processing was performed to convert the image to the standard edge map input format as shown in Figure 5.41(b).

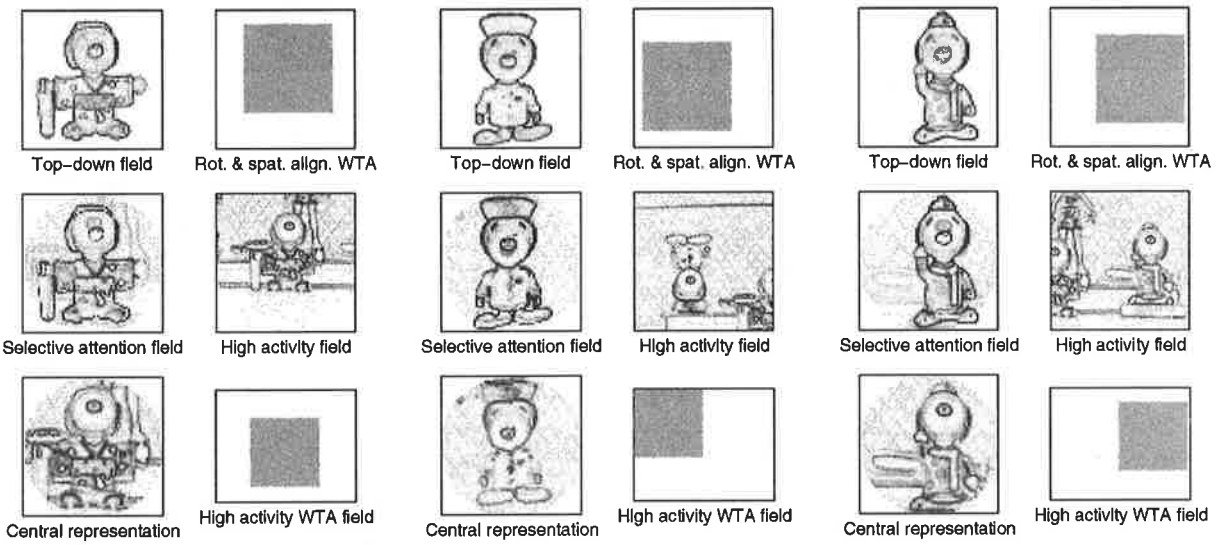
The simulation results are structured in the same compact format as before with the LTM patterns, rotational templates and input field omitted. The recognised objects are shown in Figure 5.42 in order of their recognition from left to right. As before, the high activity WTA field indicates the location of the window of attention in the input field. The rotational and spatial alignment WTA field further pinpoints the exact location of a potential target within the region of interest. Patterns from top-down and bottom-up are matched, and if required presynaptic



(a) (b)

Figure 5.41: Real-world imagery: Simulation I input scene. (a) Intensity map. (b) Edge map.

facilitation or shape attraction may be applied to enhance the degree of match between the two.



(a) (b) (c)

Figure 5.42: Real-world imagery: Simulation I - Parts 1, 2 and 3.

In addition, STM patterns of the recognised objects are shown in Figure 5.43 to illustrate the steps involved in transforming the bottom-up pattern to become a better match with its top-down memory. This real-world imagery simulation has presented a problem previously not encountered with the synthetic imagery simulations. Although we have already scaled the input scene so that the contained objects are similar in size to those learned, the 2D projections captured are different to the learned ones. As real-life objects are 3D in nature, any slight change in the camera angle or distance to an object can result in a slightly different 2D projection. It can also affect the relative size of the scene objects. It should be pointed out that no scaling on the

individual LTM patterns was performed, only the overall scale of the input scene was changed. Hence, the system was found to be robust against slight changes in size, as well as changes in 2D projection. For example, Figure 5.43(b) shows the bottom-up Snoopy Russia was captured at an angle deviated from its horizontal position, therefore Snoopy Russia in the input scene is noticeably shorter than the one learned. It is also clear that the two 2D projections of the same object are not exactly the same. Further evidence on the robustness of the system is given in subsequent simulations where the input objects were captured with slight changes in both the azimuth and elevation angles.

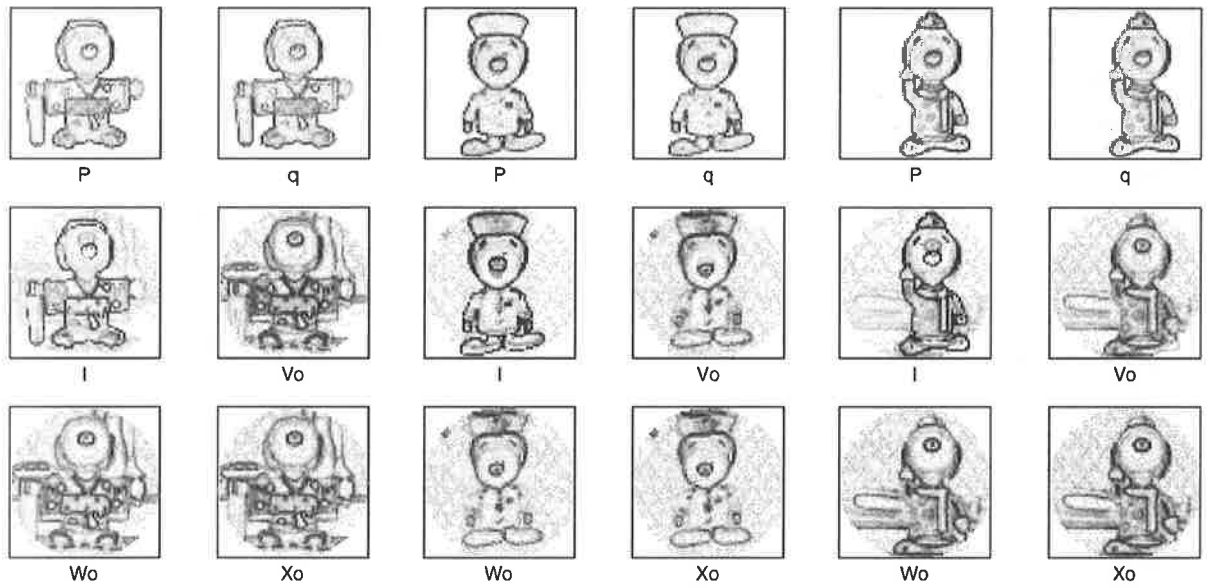


Figure 5.43: Real-world imagery: Simulation I STM patterns.

Another problem encountered is that when a captured object is too close to the image border - Snoopy China in Figure 5.42(c). Because the object pattern is contained within an area that lacks a full set of neighbours, and convolution cannot proceed smoothly through the area. Common image processing techniques employed to solve this problem include: (a) extending the image size by repeating the border rows and columns, (b) wrapping the image such that the first column comes immediately after the last, and (c) padding the image with zero rows and columns. As a quick-fix, we have shifted the object pattern slightly right to compensate for the lack of a full set of neighbours. A subsequent simulation employing method (c) above has allowed recognition to be performed successfully without any changes to the LTM patterns.

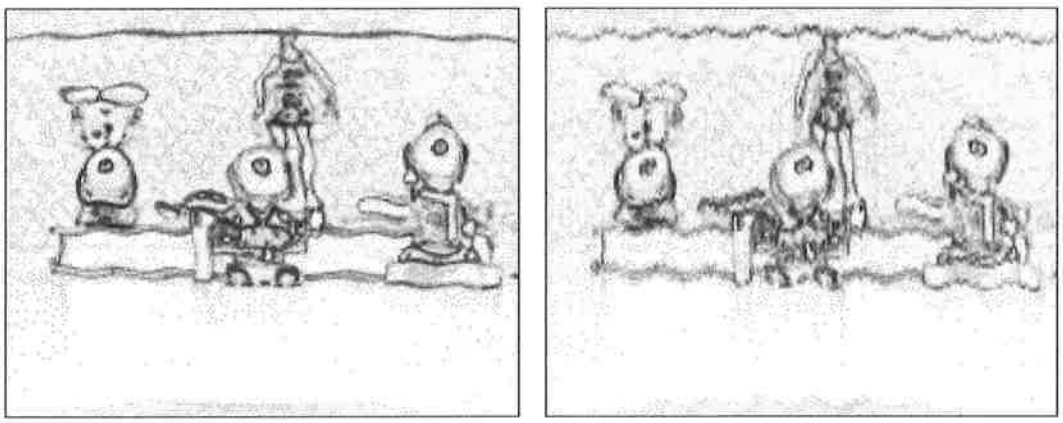
Note that the STM patterns are blank at their corners, this is due to the lack of one-to-one mapping under rotation transformation. We have padded those locations with zeros, but other methods can also be used. One can extend the area to be rotated by including extra rows and

columns from the input scene.

Distorted Real-World Scene 1

We present simulations for two distorted versions of the scene in Figure 5.41. The first is what we call a *waves distortion* which simulates the effect of throwing a stone in a pond. The second is called a *ripple distortion* which displays the image in ripples as to achieve the effect of a disturbed water surface. These two distortions are shown in parts (a) and (b) of Figure 5.44.

The two distortions are similar in nature, with the ripple case appearing to have an added blurring effect on the image. There are other distortion filters available but for the purpose of illustration the two examples are sufficient. Certainly, there is a limit to how much distortion the system can handle. A severely distorted image would challenge the human visual system.



(a) (b)

Figure 5.44: Distorted images of real-world imagery scene 1. (a) Waves distortion. (b) Ripple distortion.

The results of the two distortion simulations are shown in Figures 5.45 and 5.46. The system was able to correctly identify the bottom-up objects in each case through the use of band transformation and shape attraction. The use of shape attraction has provided an added degree of robustness to the system in terms of tolerance towards slight changes in size and 2D projection. Furthermore, the similarity in body shape between, say, Snoopy Russia and Snoopy China has confirmed the system's ability to distinguish and identify similar objects, despite distortion and background clutter.

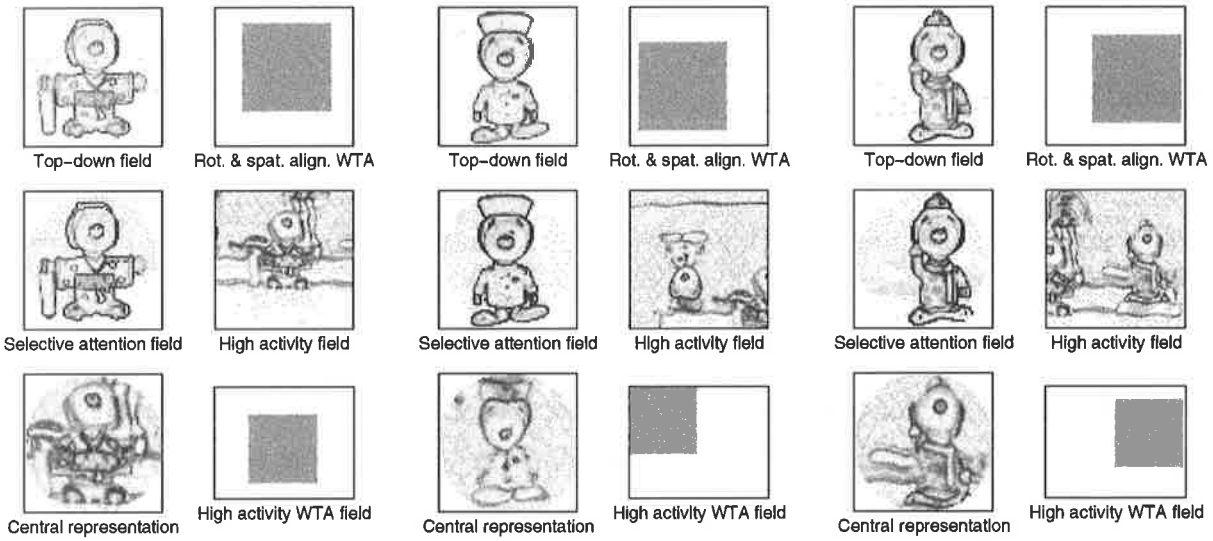


Figure 5.45: Waves distorted real-world imagery scene 1 simulation results.

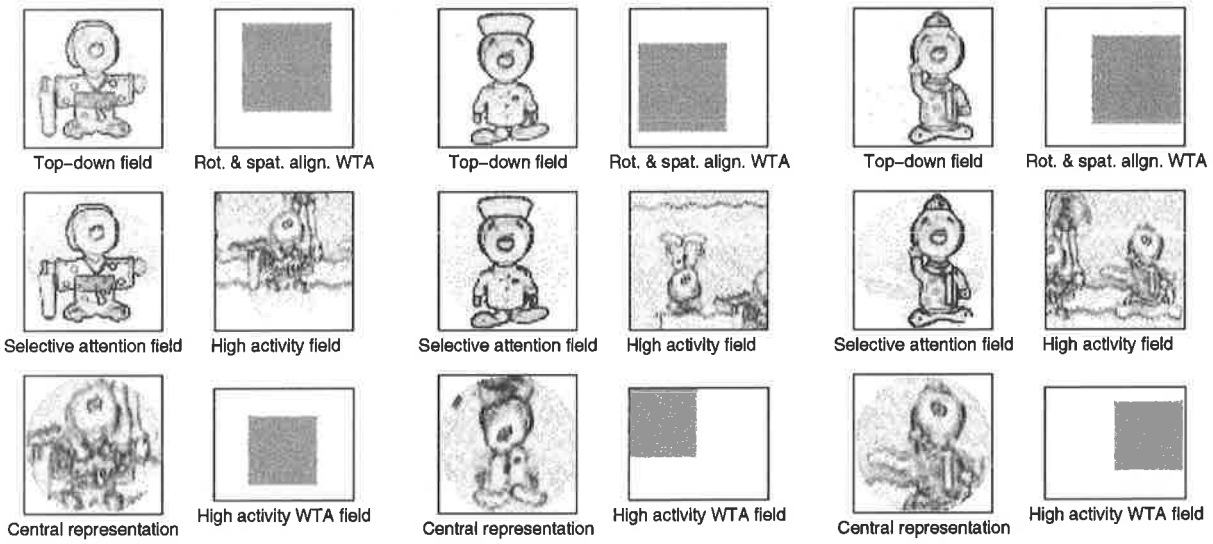


Figure 5.46: Ripple distorted real-world imagery scene 1 simulation results.

5.9.3 Simulation II

Real-world imagery scene 2 is similar to scene 1. The main difference is scene 2 contains a greater amount of background clutter which is strategically placed to interfere with the target object edges. From Figure 5.47 we can see that some of the edges from Snoopy Hongkong blend into the edges of the toy basketball player in the background. Similarly, the mobile and the pen appear to be natural extensions of Snoopy Japan in the edge map.

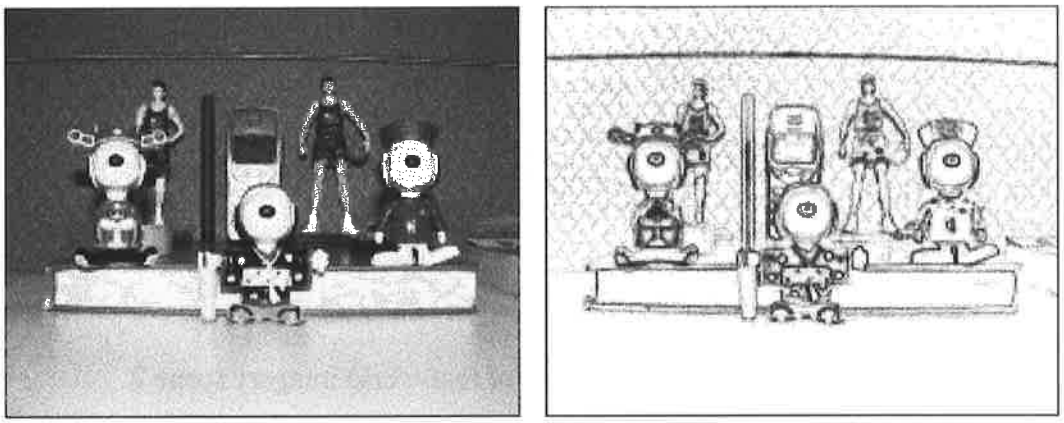


Figure 5.47: Real-world imagery: Simulation II input scene and its edge map.

Results from the simulation on real-world scene 2 are shown in Figure 5.48. The system exhibited no difficulty in identifying the objects correctly, and the greater amount of background clutter did not appear to have much negative effect on the system's performance.

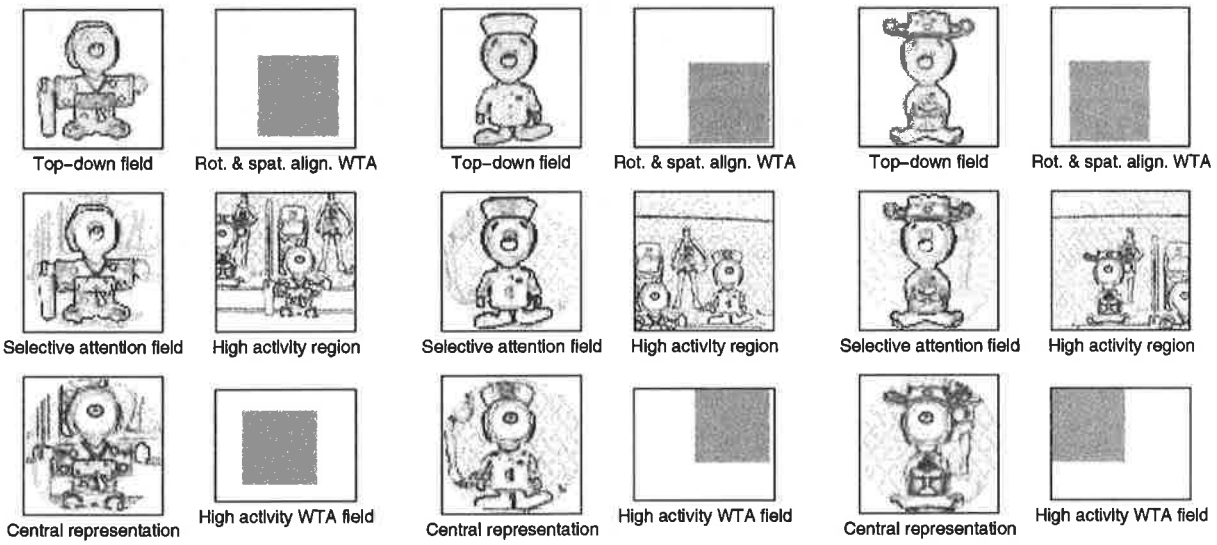


Figure 5.48: Real-world imagery: Simulation II - Parts 1, 2 and 3.

Distorted Real-World Scene 2

As before, real-world scene 2 is distorted and simulated. The distorted image is shown in Figure 5.49. Once again the system was able to correctly recognise all familiar objects in the scene, despite distortion and background clutter, as shown in Figure 5.50.

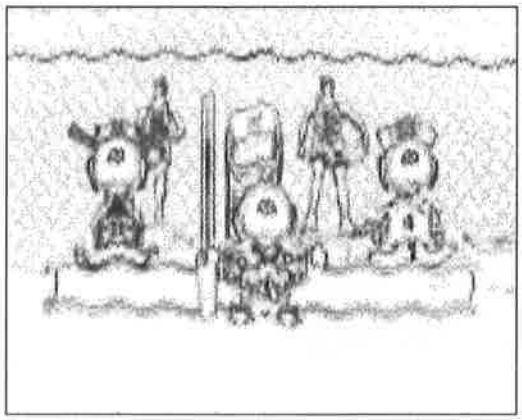


Figure 5.49: Ripple distorted image of real-world imagery scene 2.

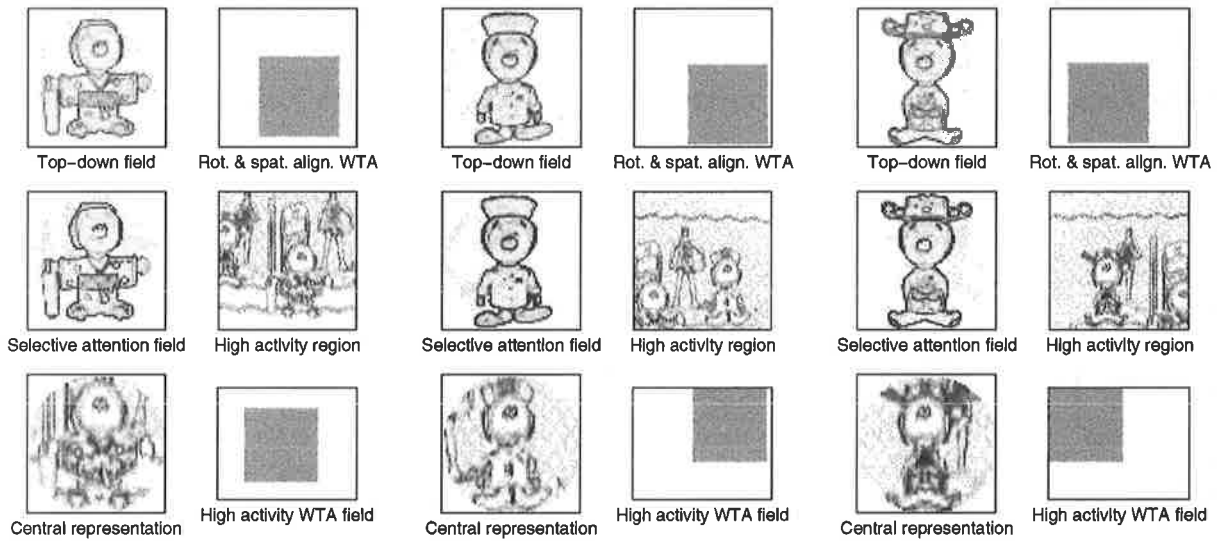


Figure 5.50: Ripple distorted real-world imagery scene 2 simulation results.

5.9.4 Simulation III

The third real-world scene is a much more challenging scenario. It can be seen from Figure 5.51 that the input scene is low in contrast but high in clutter. The problem is further compounded

by bad illumination conditions. A close examination of the input scene also shows the 2D projections of the target objects are noticeably different from the learned 2D projections. For example, the heads of Snoopy Japan and Snoopy Korea are facing away from their normal directions, caused by variations in the camera's azimuth and elevation angles.

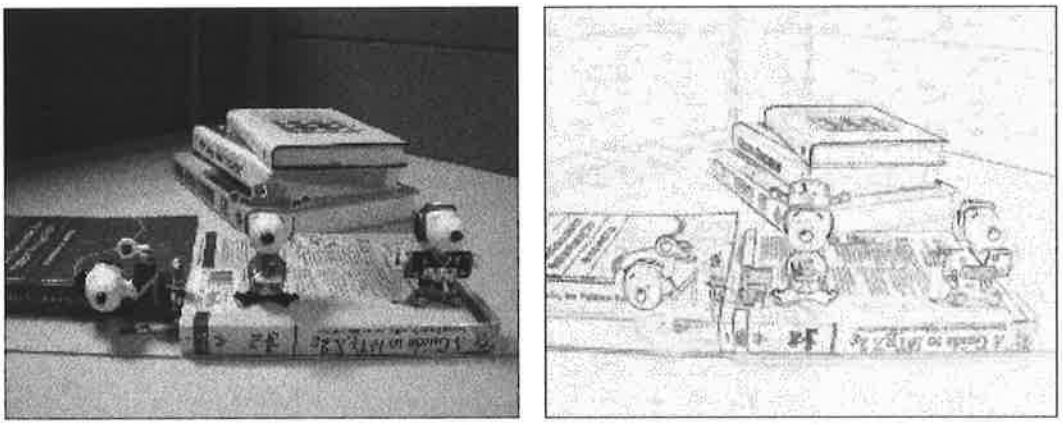


Figure 5.51: Real-world imagery: Simulation III input scene and its edge map.

Results for the simulation are shown in Figure 5.52. All familiar objects in the scene are recognised, demonstrating the system's robustness in scene analysis. It also shows the system's ability to handle extreme visual conditions such as bad illumination, strong background clutter, and slight changes in size and 2D projections.

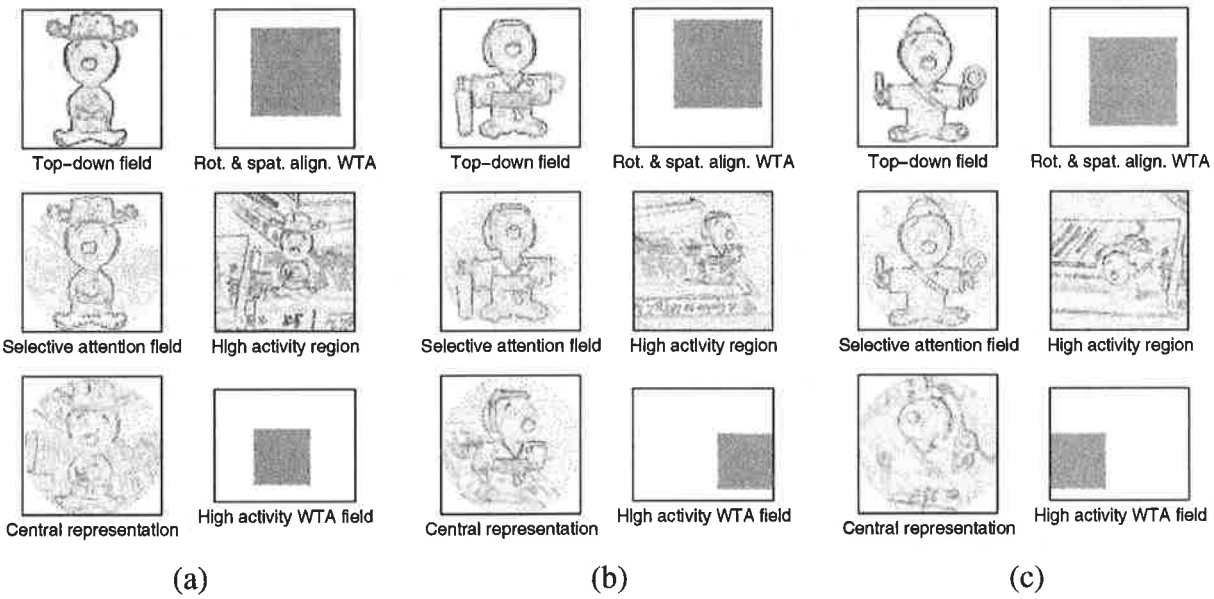


Figure 5.52: Real-world imagery: Simulation III - Parts 1, 2 and 3.

Distorted Real-World Scene 3

The same distortion is applied to real-world scene 3, as shown in Figure 5.53. However, this time the system has falsely recognised Snoopy Hongkong in the visual scene as Snoopy Japan due to a combination of extreme background clutter and distortion, as shown in Figure 5.54.

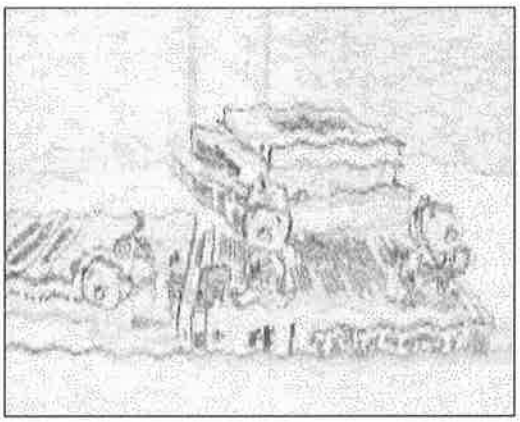


Figure 5.53: Ripple distorted image of real-world imagery scene 3.

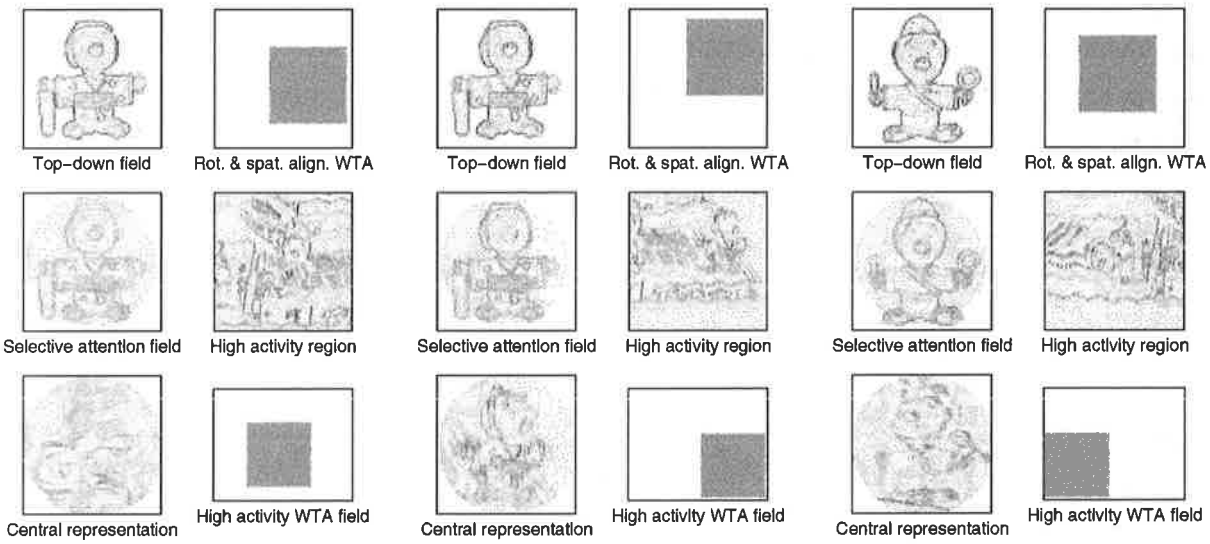


Figure 5.54: Ripple distorted real-world imagery scene 3 simulation results.

Real-World Imagery Simulation Parameters

A summary of the major parameters used in the real-world imagery simulations is shown in Table 5.5. The values shown in the table only serve to provide an indication on the useful range for

each parameter. In many cases the choice of parameters depends very much on the input object and scene sizes, as well as the nature of the input scene. For example, a highly cluttered scene will require extra care in the selection of the vigilance parameters, and whether the competition and facilitation provided in PMCNL is strong enough to filter out the undesirable elements from the input. For input scenes with closely placed objects, the post-attention inhibitory signal must not be too strong otherwise it can suppress important information from neighbouring objects. As discussed in Section 5.8 the size of the Gaussian receptive is particularly important given that a window of attention that does not cover the entire target object can lead to recognition failures. However an overly large window of attention is inefficient and time consuming.

Table 5.5: Real-world imagery simulation parameters

Input scene size $N_i \times N_j$	208×166 to 250×200 pixels
High activity region size $N_a \times N_b$	100×100 pixels
Input object size $N_p \times N_q$	64×64 pixels
Number of learned objects	5
Sampling skip	5 - 10 pixels
Primary vigilance parameter ρ	0.96 - 1.00
Secondary vigilance parameter ς	0.92 - 0.96
STM equation parameters	$a = 1$; $b = 0.2$; $c = 1$; $d = 0.9$; $e = 10^{-6}$; $\theta = 10^{-3}$
PMCNL parameters	$A = 1$; $\bar{A} = 0.1$; $B = 1$; $\bar{B} = 0.1$; $C = 0$; $D = 0.5$; $E = 1.0$; $G = 130$; $\bar{G} = 100$; $K_u = 0.1$; $n = N_i N_j - 1$; $Y = 0$; $\alpha_u = 0.05$; $\beta_u = \beta_y = 0.01$; $\gamma = 0.5$; $\Gamma = 0$; $\rho_y = 0.05$; $\theta = 0.00$;
Simulation time step Δt	0.1
Gaussian receptive field:	standard deviation $20 \leq \sigma \leq 40$
	constant $W_o = 1$
Inhibitory Gaussian receptive field:	standard deviation $15 \leq \varrho \leq 35$
	constant $\bar{G}_o = 2$

Where a range of values are given indicates the choice for that parameter is case dependent but within the given range. Parameters not given above are the same as before.
Refer to Chapter 4 for relevant equations and symbols.

5.10 Limitations of the Model

So far, we have shown that the proposed system is capable of performing well, but as with any engineering system there are imperfections and weaknesses. Some of the more obvious shortcomings are caused by visual functions that have yet to be implemented, such as 3D object recognition and handling large variations in size. However, the system has been shown to be tolerant to minor changes in 2D projections and size, as illustrated by the real-world imagery simulations in Section 5.9.

One main limitation discovered during the simulations is the need for adjustment and fine tuning of system parameters for new images. This is especially true for complex cluttered scenes. While simple scenes of similar nature usually only require adjustment once, it has been found that for complex scenes the success of the system is highly sensitive to the vigilance parameters. For example, a familiar object surrounded by background clutter requires top-down presynaptic facilitation to strengthen object elements while inhibiting non-object elements. But applying presynaptic facilitation indiscriminately can have adverse effects on the false alarm rate, as it attempts to extract a similar pattern out of the cluttered background. Similarly, if shape attraction is not applied carefully, it could attract unrelated elements to form a familiar pattern, thus it is important that prior to their application the bottom-up and top-down patterns must be of an acceptable degree of match. This problem of finding suitable system parameters has also plagued other neural architectures. The performance of the neocognitron [64] is shown to be strongly dependent on the choice of parameters [9, 111].

Another process that often requires adjustment is attentional capture. For low contrast objects or objects surrounded by strong background clutter, it is possible that they may never capture attention, resulting in their non-detection. In practice, the problem is more likely to be that the target object is near the border or partially captured by the window of attention. In either case, the object cannot be detected due to the lack of a full set of neighbours, the region of interest is then suppressed by the post-attention inhibitory signal, therefore the object becomes undetected. This problem is usually solved by changing the receptive field and post-attention signal sizes. The size of the high activity field may also be changed to accommodate for these objects.

Instead of changing the system parameters, the attentional capture problem may be solved by applying a top-down expectation. Currently, the preattentive mode captures attention based on bottom-up information only by selecting the region with the highest contrast. However, the preattentive mode can also use top-down information for attentional capture. Consider the toy figures used in the real-world imagery simulations; it depends whether we are consciously looking for those toys or simply studying a visual scene containing those toys. If it is the latter

case, then we would simply recognise all objects within the scene as modelled in our system. On the other hand, if we were trying to look for the toys, then we would have some knowledge from top-down on what the toys would look like. So instead of using a Gaussian receptive field to capture the most salient region, we would use an object body shape receptive field to capture regions containing object patterns similar to the toys' general shape. Preliminary simulation results have proved that the use of body shape receptive field is feasible and effective, and are shown in Appendix A.

In the following, we provide a list of situations where the proposed visual scene analysis system might fail.

- Extremely cluttered scenes, e.g., newspaper as background.
- Large part of an object is occluded.
- Poor edge maps from low contrast scenes or bad lighting conditions.
- Object size differs significantly from memory.
- Two dimensional projections that differ significantly from their counterparts in memory.
- Severely distorted scenes.

Solutions to some of these problems will be proposed in Chapter 7.

5.11 Conclusions

In this chapter, the proposed visual scene analysis system was applied to a number of digital images, consisting of both synthetic and real-world scenes, for visual analysis. The simulations showed that the proposed system is capable of detecting, locating and recognising all familiar objects within a visual scene, regardless of their positions, orientations and background complexity in an automatic fashion. The system is also robust against minor changes in an object's shape caused by distortion or different viewing angles, i.e., different 2D projections of the object.

We have chosen to use both synthetic and real-world images because the former are useful during model development and testing. While the latter provide us a more challenging problem with many practical considerations. Camera angle and distance, lighting conditions, non-uniform object size and differences in shape between stored models and input objects are just some of the issues encountered only in real-world imagery simulations.

The real-world imagery simulations undertaken involved capturing a number of real-life scenes featuring several input objects by using a digital camera. The captured images were pre-processed into suitable edge maps. Appropriate system parameters were chosen for the input scenes and objects. The actual system began by learning the input objects using the ART2 learning algorithm. The input scenes were simulated by the system after some parameter adjustment and fine tuning.

By and large, with the aid of adjustment, the proposed system was successful in locating and recognising all familiar objects in the input scenes despite difficult visual conditions. Several minor problems encountered during simulations are highlighted and possible solutions proposed, and in some cases verified via further simulations. In particular, the problems associated with changes in camera angles are discussed. The results demonstrated that the system is robust against minor arbitrary changes in an object's shape.

An example of system parameter design is provided. It describes how the system can be adjusted to produce better performance. Limitations of the system are discussed and possible solutions are offered in some cases.

In conclusion, the proposed recognition has been demonstrated to be effective in performing translation, rotation and distortion invariant object recognition with automatic attentional capture and shift, in the presence of background clutter and occlusion.

Chapter 6

Recognition of Moving Objects

6.1 Introduction and Overview

Motion is an important part of seeing. To see is to connect inferences about motion, colour and patterns into a unified explanation of the visual scene. Elementary motion is involved in an early stage of object recognition [23, 195], since the same pattern must be located over a spatial and temporal range, in order to perceive motion. Movements of an object have been identified as one of the elementary features used in attentional capture [19, 98]. Furthermore, Lu and Sperling [117] have shown that voluntary attention can determine the direction of perceived visual motion.

Besides luminance contrast, motion would be an important addition for the proposed visual scene analysis system for bottom-up attentional capture. Such an addition would allow the system to detect, locate and recognise any familiar moving objects, complementing and enhancing the existing capabilities for static object recognition. It can also demonstrate the model's extendibility as a framework for visual scene analysis.

This chapter presents a neural architecture for the detection of elementary motion direction. The approach is based on the human motion perception that motion is a visual inference called "apparent motion". Modelling is primarily inspired by interactions along the motion pathway, using similar neural equations as for the static system. The chapter describes how this motion detection architecture fit into the overall framework for recognising moving objects.

By modelling the motion detection architecture with the same building blocks, it acquires the ability to perform presynaptic facilitation, thus it possesses attentional mechanisms. A useful application of the attentional mechanisms is the modelling of directional bias, such that the system can favour movements in a particular direction via a top-down cognitive signal. Selective

attention can even determine whether motion is perceived at all [117]. An attentional modulated motion detection module may allow the system to selectively follow and interpret one moving object, whilst avoiding being distracted and confused by other moving objects or background.

This chapter is organised in the following manner. A brief introduction of the motion pathway and apparent motion is given in Section 6.2. The proposed neural architecture is presented in Section 6.3. Simulations and analyses of the motion module are provided in Section 6.4, and finally, we describe how the motion module can be incorporated into the framework in Section 6.5.

6.2 The Motion Pathway

Strong physiological evidence exists to indicate that motion is processed by a discrete visual portion of the visual pathway. The discovery of direction-sensitive neurons, which are relatively insensitive for other features such as colour and orientation, in cortical areas is an important step in understanding motion perception [53, 208]. Direction-selective neurons are only found in certain layers of the cortex and are quite rare in other areas. Besides area V1 (striate cortex), a principle projection site of the LGN (lateral geniculate nucleus), they are also located in area MT (middle temporal). The direction-selective neurons in area V1 appear to send their output mainly to area MT, which leads to the suggestion that the path from area V1 to area MT plays an important role in motion perception. Based on the direction-selective receptive field property in these areas, this particular visual stream is called the *motion pathway* which is summarized in Figure 6.1.

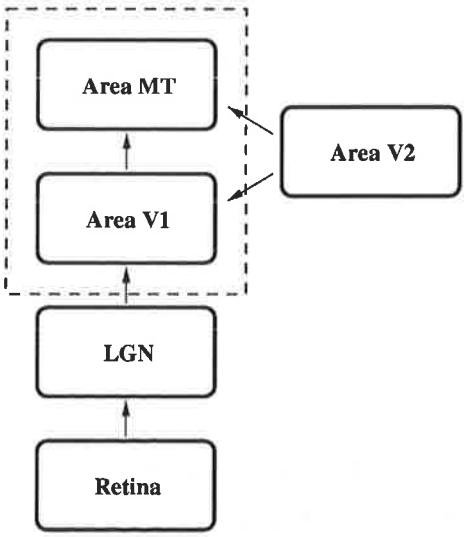


Figure 6.1: Anatomy of the motion pathway.

As we move up the motion pathway, the receptive field properties of the neurons within the motion pathway respond to increasingly sophisticated stimulus properties, showing the hierarchical nature of information processing in the visual system. For example, Movshon *et al.* [132] discovered direction-sensitive neurons that are also pattern-sensitive.

The motion detection module is inspired by the receptive field properties in the areas V1 and MT, as indicated by the dotted box in Figure 6.1. Although the vast majority of neurons in area MT are direction-selective, there are neurons with other receptive field properties.

6.2.1 Apparent Motion

Motion perceived by our eyes is only an inference of the retinal image by the visual system, not a description of the physical world. A demonstration of this inference is apparent motion, which is a phenomenon we see in our everyday life. You may have noticed apparent motion when sitting in a stationary bus that is side by side with another bus. If the other bus moves forward and yours remains, occasionally you get the sensation that your bus is moving backwards. This sensation is apparent motion.

Understanding apparent motion can help explain motion perception, because our visual system responds as well to apparently moving stimuli as to real moving stimuli. Apparent motion can be thought as a form of visual illusion, and by exploiting the imperfections of our visual system we can be tricked into seeing something that is not real. Motion pictures are a common application of apparent motion, in which still frames are rapidly altered in time, thereby inducing motion perception, even though there is no actual movement of stimuli. So the motion module will be simulated using apparent motion.

Because motion is an inference process, humans are good judges of relative speeds but tend to perform poorly in judging absolute speeds. Also, due to the physiological structure of the visual system, perceived speed is affected by stimulus contrast. The effect of contrast on speed perception will be discussed further in a later section.

6.3 Neural Architecture for Motion-Direction Computation

Three areas have been identified from the motion pathway that are required for modelling the motion module. The first is image formation, which concerns the encoding of the retinal image. The second is related to how the encoded information is represented by the neural responses within the early visual pathways. Finally, an interpretation of the neural representation

is achieved. In this case, the interpretation is usually the detection of motion and its direction.

Several broad principles regarding the representation of input stimulus by the neural response within the peripheral and early cortical visual pathways have emerged from vision research. Two of which are fundamental to our approach to the implementation of the neural motion detector:

Principle 1. Anatomical studies show that neurons in the visual pathway are segregated into different visual streams, with each responsible for a specialized visual function. For example, area V4 is believed to be responsible for colour perception [207] and area MT for movement [53]. The functional role of individual visual streams is identified by inferring from the anatomical properties along with the way the neurons in these separate streams respond to light stimulation.

Principle 2. Electrophysiological experiments suggest that the most important information represented by the visual pathways is the image contrast, not the absolute stimulus level. Neurons in the early stages of the visual system are most sensitive to image contrast and respond best to highly contrasted input stimuli. The perceived speed of an object is contrast dependent. High-contrast targets appear to move faster than low-contrast targets when both have the same physical speed [179].

Principle 1 allows us to treat the motion pathway as an independent module of the visual system model. The modular approach is useful for modelling, since individual visual processes can be modelled and analyzed independently before integration. This is the approach we have adopted to incorporate the motion module into the framework. The second principle indicates that the proposed motion detection model should utilize contrast information to generate motion signals, and be able to explain why perceived speed is contrast dependent.

The three identified areas are translated into the following processing stages: light adaptation, contrast-gain control, motion-direction computation, and top-down selective attention. To implement these processes, we propose a five-layer neural architecture, comprising of the *input layer*, the *photo-sensitive layer*, the *transient layer*, the *direction-selective layer*, and the *selective attention layer*, as depicted in Figure 6.2.

The layers are modelled using the same or variant of the differential equations for the chemical synapse and neurotransmitter models, reviewed in Section 3.4, and used in modelling the framework in Section 4.3.

Although the visual stream shown in Figure 6.1 is termed the motion pathway, it consists of neurons of unknown function. This suggests that the pathway may have functions beyond

motion perception, and that other portions of the visual pathways may also be important for motion. So the motion module described herein is merely a model inspired by the motion pathway, and is far from a model of the motion pathway. Furthermore, the model covers only the motion pathway features that are responsible and essential for motion-direction computation.

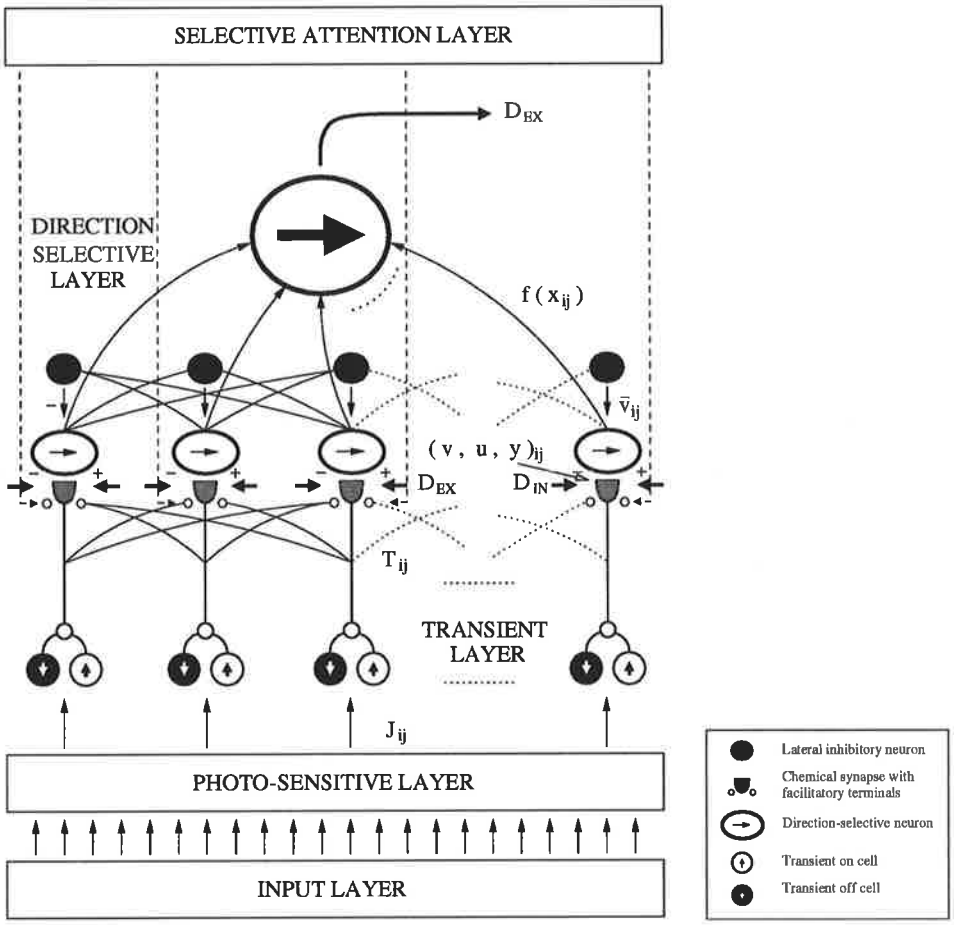


Figure 6.2: Direction-selective neural architecture.

6.3.1 The Input Layer

This layer is for the acquisition of external visual stimuli in 2D space. Equivalent to the formation of the retinal image. The input is represented by I_{mn} denoting the luminance of the input at position mn. We can view this as an image formation layer.

6.3.2 The Photo-Sensitive Layer

This is a preprocessing layer, consisting of models of the photoreceptors, for space-averaging and contrast enhancement of the input image. In other words, this layer performs the initial encoding of light to neuronal responses. This layer is modelled by a shunting on-centre/off-surround competitive neural network [27, 73]. As a preprocessing stage, this layer is not always required, except when the input image is very large or low in contrast. The outputs of this layer are denoted as J_{ij}.

6.3.3 The Transient Layer

The early visual pathways in primate retina consist of sustained ganglion cells and transient ganglion cells [55], which are believed to be responsible for processing information for form perception and motion. Sustained cells are responsive to local contrast of image, while transient cells give transient responses to light onset or offset, thus are sensitive to movement or changes in contrast pattern. So it is appropriate for us to model the early stage of the neural motion detector with a layer of transient cells.

Two types of transient cells are needed, transient *on*-cells and *off*-cells, for input increments and decrements, respectively. These two cell types are based on [76, 145] and have the same structure which is shown in Figure 6.3.

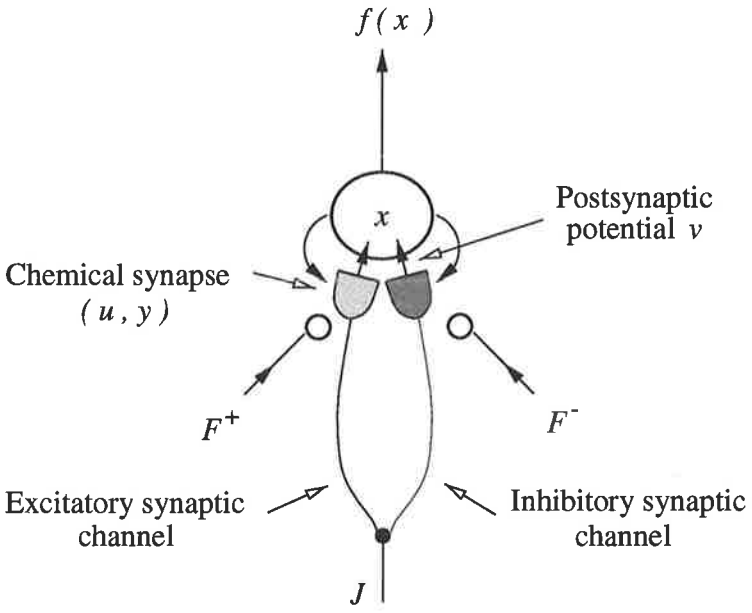


Figure 6.3: A single neuron scheme to model a transient response cell.

A transient cell is characterised by its two synaptic connections: the excitatory channel and the inhibitory channel. Both of these channels respond to the input luminance level J, except that the inhibitory channel acts to prevent the cell from firing while the excitatory channel does the exact opposite. Hence, the cell fires whenever the input from the excitatory channel exceeds the inhibitory one. The transient pair are modelled by the basic synaptic equations as follows:

- Postsynaptic cellular activity

$$\frac{dx}{dt} = -Ax + (B - x)G[v^+ - v^-]^+ - (C + x)G[v^- - v^+]^+ \tag{6.1}$$

  where A is the passive decay rate, B and C are the saturation limits for the upper and lower bounds respectively, and both G and $\bar{G}$ are amplification factors. This equation represents shunted competition of a layer of neurons with the on-centre off-surround anatomy whose cellular activity is restricted to the range $(-C, B)$. $[y]^+ = \max(y, 0)$ is a threshold function. The middle term $[v^+ - v^-]^+$ provides excitation to the transient cell, while the last term $[v^- - v^+]^+$ inhibits it.

- **Excitatory postsynaptic potential**

$$\frac{dv^+}{dt} = -D^+v^+ + J[y - Y]^+(\rho_v^+ + E^+f(x)) \tag{6.2}$$

- **Inhibitory postsynaptic potential**

$$\frac{dv^-}{dt} = -D^-v^- + J[y - Y]^+(\rho_v^- + E^-f(x)) \tag{6.3}$$

  where $D^{\pm}$, $E^{\pm}$, and $\rho_v^{\pm}$ are constants, J is the input, Y is the threshold for transmitter release, and $f(x) = \max(x - \theta, 0)$ is the thresholding function. The excitatory/inhibitory postsynaptic potential acting on a cell is due to the bound transmitter on the postsynaptic cell.

- **Stored transmitter**

$$\frac{du}{dt} = \alpha_o(z - u) - (\beta_o + K_uJf(x))(u - y) \tag{6.4}$$

  where α_o and β_o are tonic adaptation constants, z is the transmitter production rate, and K_u is a constant. This equation says that the transmitter storage rate in the excitatory synapse is depleted by the correlated firing of the input signal J and the postsynaptic feedback signal $f(x)$.

- **Mobilized transmitter**

$$\frac{dy}{dt} = (\beta_o + F)(u - y) - J(\rho_y + K_yf(x)[y - Y]^+ - \gamma y \tag{6.5}$$

where F is the facilitatory signal, ρ_y and K_y are constants, and γ is the decay constant. The equation says that the transmitter mobilization rate is increased by the facilitatory signal F and that the transmitter is released by the input signal J and by the correlated firing of J and the postsynaptic feedback signal $f(x)$.

How does the cell work given that both synaptic channels are modelled by similar equations? The answer lies in the rates at which the two channels react. If one channel reacts faster than the other, then it will always reach steady state first upon a change in input. This creates a difference between the two channels during the transient period. Our model exploits this period to provide activation to the cell. For a transient on-cell the excitatory channel must be faster than the inhibitory one, so that whenever there is a rise in the luminance level the excitatory channel will react first and the inhibitory one will try to catch up, the difference between these two triggers the on-cell as in Equation (6.1). When the two finally reach steady state they negate each other's input to the cell, thus the on-cell is shut off. The off-cell works on the same principle, but the inhibitory channel is faster instead. A drop in the input luminance level will cause both channels to decay, but with a faster decaying rate the inhibitory channel will reach the steady state first, and the off-cell remains excited until the two channels settle at their steady state.

These cells can be regarded as change detectors, they respond whenever changes occur to the input luminance. The transient on- and off-cell outputs converge to a maximal activation neuron which outputs the larger one of the two signals, denoted as $T(i, j)$. In Figure 6.2 the off-cells are represented as dark circles with bright arrows and the on-cells as bright circles with dark arrows.

To illustrate how the excitatory and inhibitory synaptic channels interact with each other to give rise to transient on and off signals, two computer simulation results are included in Figure 6.4 and Figure 6.5. The top graph in the respective figures is the actual transient response, the middle graph has plots of the excitatory postsynaptic potential v^+, in solid line, and the inhibitory postsynaptic potential v^-, in dashed line. The bottom graph is the input which is the same for both on- and off-cells. Figures 6.4 and 6.5 show clearly that transient cells are driven by the difference between the two opposing potentials which negate the effect of each other when in steady state.

6.3.4 The Direction-Selective Layer

The motion pathway is basically inferred from areas with direction-selective neurons. From electrophysiological experiments on the detection of such neurons and the signals leading up to

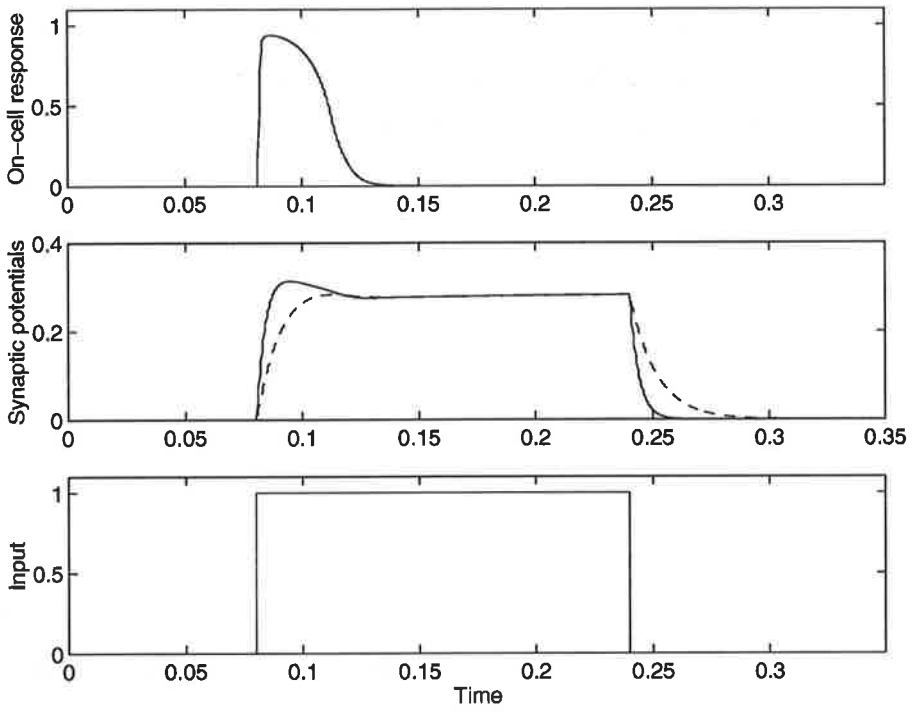


Figure 6.4: On-cell response to step input.

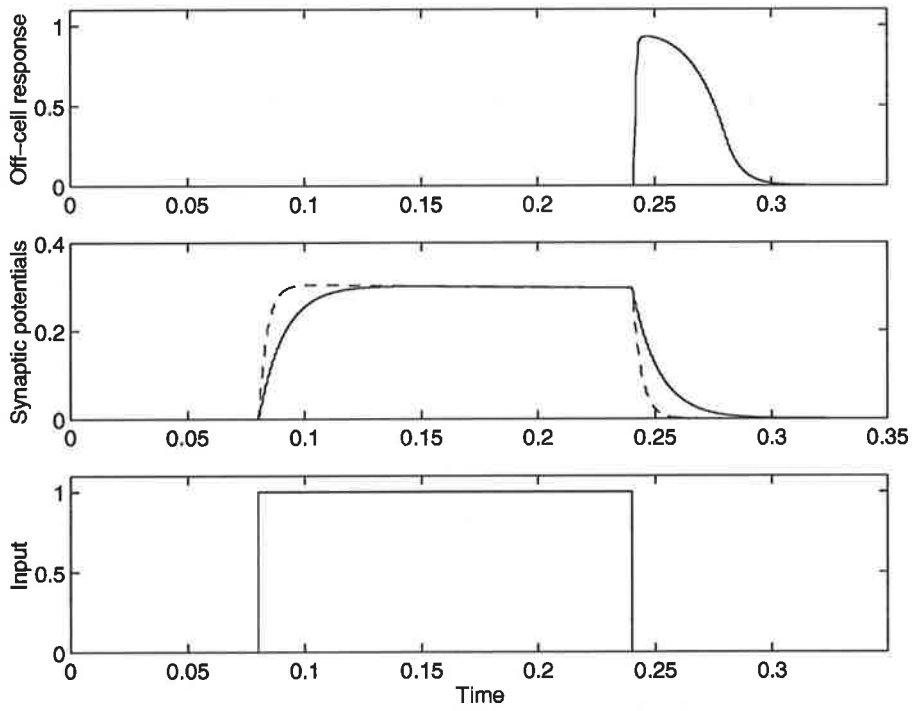


Figure 6.5: Off-cell response to step input.

them, we can speculate on how the transformation takes place and propose hypothetical models which utilize the same input signals and generate direction-selective outputs. The outcome of this approach is a computational model. However, a computational model that is inspired by biological findings can help explain and predict physiological phenomena.

The direction-selective layer takes transient signals as inputs and outputs directionally sensitive neural responses. Each direction-selective neuron has a *preferred direction* $\vec{D}$ [170] of motion, to which the cell responds best. Movements to the motion module is simply the shifting of a contrast pattern from one spatial location to another in time. Based on these ideas we derive a structure, shown in Figure 6.6, for one-dimensional motion detection. As an example, the depicted structure's preferred direction is to the right.

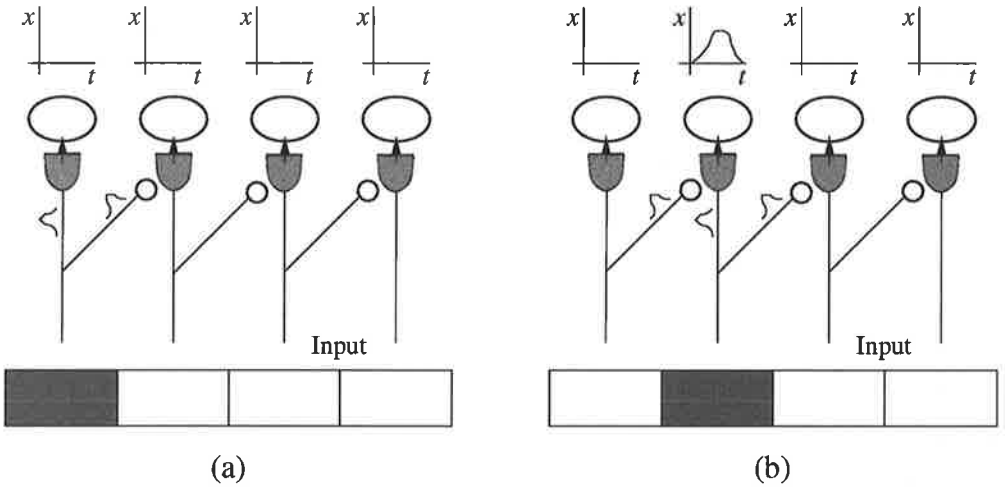


Figure 6.6: Simplified 1D direction-selective layer.

The main feature in Figure 6.6 is the interconnections between synaptic pathways, whose role is to facilitate the synaptic pathway to the right of itself. This has the effect of allowing information to flow from one spatial location to the next. Also note that the synaptic input in each pathway is a transient signal, which means it can only persist for a certain period of time. If a change in the input, as in Figure 6.6, takes place before the transient signal is completely decayed away, then the pathway with the change is facilitated by its neighbour. More importantly the facilitation as shown is from left to right in agreement with the preferred direction, resulting in the activation of the direction-selective cell, as shown in Figure 6.6(b). The excitation is generated by transient signals across two spatial locations, which can be regarded as a temporal contrast pattern and the resultant transformation as a spatio-temporal receptive field. To cover movements beyond the next spatial location in the preferred direction, each pathway can either facilitate all or some of the pathways to its right, depending on the desired effective range. This forms the basis of the direction-selective layer.

The structure is only capable of detecting movements to the right. So for 1D motion detection,

another similar structure with synaptic pathways interconnected from right to left is required for detecting movements to the left.

Several design issues are addressed below prior to the extension of the layer to 2D:

- The model is such that without facilitation the cellular activities will be very minimal, as in the case of Figure 6.6(a), where a single change occurs in the input stimulus.
- A sudden change, such as the presentation or removal of a contrast pattern spanning several spatial locations, will cause all corresponding direction-selective neurons to fire. This is of no concern as the cellular activity in each will be of the same magnitude, and under the mutual inhibition and competition no overall movement will be detected.
- At the facilitatory node of each pathway, the synaptic input is facilitated by neighbouring transient signals. This allows movements in the stimulus pattern to be detected. If the synaptic pathway has no input, then the facilitatory signals simply increase the rate of accumulation of neurotransmitters at the synapse, so the pathway is more responsive to a future input.
- There exists two types of direction-selective cell. The first, $x_{\vec{D}_{ij}}$, responds locally to movements in its preferred direction as in Figure 6.6. The pattern formed by local direction-selective cells is called the *directional field*. The second, $\mathcal{X}_{\vec{D}}$, provides a global measure of the overall movement in the preferred direction. $\mathcal{X}_{\vec{D}}$ is obtained by summing up individual local direction-selective cell activities, shown as a large direction-selective cell in Figure 6.2.

Figure 6.7 shows a 2×2 example for a 2D direction-selective layer. Each 2×2 field represents an input stimulus, consisting of four pixels. Next to each field is a cell which represents a global direction-selective neuron in the direction indicated by the arrow encircled. The arrows in the fields represent the facilitatory connections required, thus the synaptic pathways connections that are required to form the directional field. The number of synaptic pathways and facilitatory connections required by each direction-selective neuron varies. It is impractical to model each direction-selective cell with a different structure, because changing the direction requires rewiring the neural circuitry. In the following, we propose an architecture for computing movement directions by wiring all synaptic pathways together.

A more general model that is suitable for any input stimulus size and motion direction is needed if this were to become a useful engineering model. Inspired by the fact that complex webs of interconnections [193] are very common in visual pathways, all the synaptic pathways are connected together as in Figure 6.2. Given that all pathways can communicate with each other,

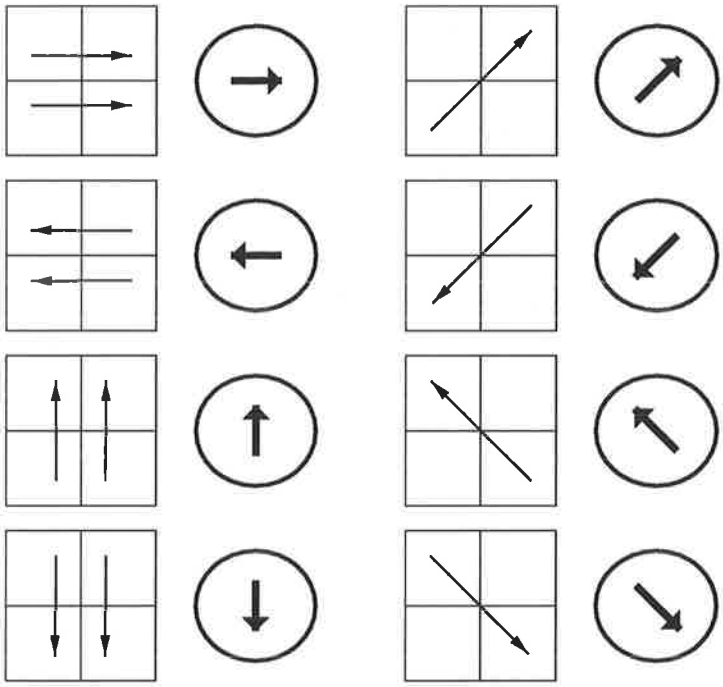


Figure 6.7: 2D direction-selective layer design.

we need to pick out the contributions in the preferred direction while ignoring the rest of the signals. Figure 6.8 is another representation of the interconnecting nature of the layer, which should give us some idea how this can be done. It shows that the preferred direction is to the right, and according to Figure 6.7 only two of the twelve connections (shown as arrow heads) are in agreement with it. The two connections can be selected by performing simple dot product operations between the preferred direction vector and the vectors formed by the interconnections, denoted as $\vec{D}$ and $\vec{p}$, respectively.

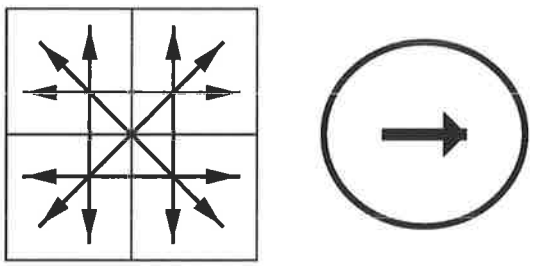


Figure 6.8: Combined design approach.

Architecturally, the resultant design of the direction-selective layer consists of four sub-layers as shown in Figure 6.2. The first is an interconnected synaptic layer, (v_{ij}, u_{ij}, y_{ij}), for communicating spatial and temporal information between synaptic pathways. Next is a directional field layer, $x_{\vec{D}_{ij}}$, formed by a group of direction-selective cells. Followed by a lateral feedback inhibition layer, $\bar{v}_{ij}$, for providing local competition. Lastly, a global direction-selective layer,

$\mathcal{X}_{\vec{D}}$, is required for determining the overall direction of movement. Note the number of local direction-selective cells depends on the input stimulus size, and the number of global direction-selective cells depends on the number of directions required. For example, for computing 2D motion with a 45^o resolution, it will require eight global direction-selective cells, thus eight preferred directions.

Mathematically, the layer is completely described by the dynamic equations in (6.6)–(6.10). $x_{\vec{D}_{ij}}$ is the postsynaptic cellular activity which is a measure of the neuronal activity for the preferred direction $\vec{D}$ at location (i, j). v_{ij}, y_{ij} and u_{ij} are the postsynaptic potential, the mobilized transmitter and the stored transmitter for location (i, j) for modelling the behaviour of a chemical synapse [113]. The main feature of this layer is the cooperative and competitive interactions of the network. Within each directional field, the individual cells $x_{\vec{D}_{ij}}$ are engaged in an intra-competition provided by the lateral feedback inhibition in (6.8) to contrast enhance the movement locations. At the same time these competing cells cooperate to form an overall activity, $\mathcal{X}_{\vec{D}}$ (represented by the large cell in Figure 6.2), which in turn forms the directional excitatory feedback signal D_{EX} in (6.11). This feedback is then used against an inhibitory feedback D_{IN} in an inter-competition between opposing directional fields in (6.7). D_{IN} is formed by the opposing directional field $\vec{d}$ for inhibiting $x_{\vec{D}_{ij}}$. The facilitatory signal F_{ij} is determined by summing up transient signals facilitating location (i, j) from location (k, l) and multiplied by a dot product between the normalised vector $\vec{p}$ and the preferred direction $\vec{D}$. So that only transient movements in agreement with the preferred direction of the directional field can provide excitations to the directional cells as in Equation 6.12. Note all parameters have the same meanings as before.

$$\frac{x_{\vec{D}_{ij}}}{dt} = -Ax_{\vec{D}_{ij}} + (B - x_{\vec{D}_{ij}})Gv_{ij} - (C + x_{\vec{D}_{ij}})(\bar{G}\bar{v}_{ij} + \Gamma) \tag{6.6}$$

$$\frac{dv_{ij}}{dt} = -Dv_{ij} + T_{ij}y_{ij}(\rho_v + Ef(x_{\vec{D}_{ij}}) - D_{IN} + D_{EX}) \tag{6.7}$$

$$\frac{d\bar{v}_{ij}}{dt} = -\bar{A}\bar{v}_{ij} + \frac{1}{N_i N_j}\bar{B} \sum_{(k,l)\neq(i,j)} f(x_{\vec{D}_{kl}}) \tag{6.8}$$

$$\frac{du_{ij}}{dt} = \alpha_o(z_{ij} - u_{ij}) - (\beta_o + K_u T_{ij} f(x_{\vec{D}_{ij}}))(u_{ij} - y_{ij}) \tag{6.9}$$

$$\frac{dy_{ij}}{dt} = (\beta_o + F_{ij})(u_{ij} - y_{ij}) - T_{ij}(\rho_y + K_y f(x_{\vec{D}_{ij}})y_{ij}) - \gamma y_{ij} \tag{6.10}$$

$$D_{EX}, D_{IN} = \frac{1}{N_i N_j} \sum_{i,j}^{N_i,N_j} f(x_{ij})_{\vec{D},\vec{d}} \tag{6.11}$$

$$F_{ij} = \vec{D} \cdot \vec{p} \sum_{(k,l)\neq(i,j)} T_{kl} \tag{6.12}$$

Computationally, a 2D direction-selective layer of size $N_i \times N_j$ can be summarized as follows:

1. at each synaptic pathway (i, j), $i = 1 \ldots N_i$ and $j = 1 \ldots N_j$, test if the facilitatory signal from pathway (k, l) $k = 1 \ldots N_i$ and $l = 1 \ldots N_j$ is of different transient signal type to that of (i, j). As no facilitation between pathways with the same type of transient signal is allowed;

2. calculate the normalised spatial vector $\vec{p}$ from location (k, l) to location (i, j) and the dot product $\vec{D} \cdot \vec{p}$, thus the facilitatory signal, F_{ij}, as given in (6.12);

3. compute D_{EX} and D_{IN};

4. iterate v_{ij}, $\bar{v}_{ij}$, u_{ij} and y_{ij}; and

5. iterate $x_{\vec{D}}$ in (6.6). The subscript $\vec{D}$ indicates this cell is directionally selective in the direction of $\vec{D}$. So for a system requiring detection of motion in n directions, there will be n sets of the direction-selective layer.

6.3.5 The Selective Attention Layer

Selective attention is incorporated at the facilitatory terminals located alongside the synapses. It allows attentional modulation - a gain control mechanism which governs the flow of charge particles from the synapses to their cells modelled in (6.10), thereby adjusting the postsynaptic potential in (6.7), available to the cells. The facilitatory signal F_{ij}, shown in (6.13), is a control signal which carries out decisions from top-down. This particular implementation of top-down selective attention allows the system to favour movements in a chosen direction, while ignoring or inhibiting the others, i.e., a directional bias. The bias, $\vec{b}$, is modelled as a dot product with the preferred direction $\vec{D}$:

$$F_{ij} = \vec{D} \cdot \vec{p} \sum_{(k,l)\neq(i,j)} T_{kl}\, \vec{D} \cdot \vec{b}. \tag{6.13}$$

Neuropsychological studies [152, 155] support the existence of the directional bias and its effect in determining directional cell responses.

6.4 Simulations and Analysis

There are three parts in the analysis of the motion module. The first seeks to verify the model's ability to detect and compute motion direction. Next, the effects of stimulus contrast and temporal frequency on motion signals are investigated. Lastly, we illustrate how to use top-down attentional modulation to achieve directional bias.

6.4.1 Motion-Direction Detection

To demonstrate the effectiveness of the proposed model, two computer simulations of apparent motion based on greyscale image sequences are described. The images in each simulation are fed to the neural model at a rate of twenty-five frames per second. Two sets of image sequence have been chosen to illustrate different aspects of the model's properties. The first set is a moving tennis ball sequence, which is good for visualizing the process. The other set is a sequence of simple shapes at pixel level, such as a bar moving to the right, up or diagonally. This allows us to examine individual cell locations, and is useful for analysis as this shows how activities build up over successive frames.

For computational convenience, we have excluded the photo-sensitive layer which has the functions of image compression and contrast enhancement. This is a relatively unimportant layer in terms of motion detection. For large or low contrasted images, this layer must be incorporated for the model to be efficient. The exclusion has the effect of making the image pixel at location (i, j) as the transient layer input $J(i, j)$.

The results of each simulation are organised and presented in a number of grid-like representations, as shown in Figure 6.9, for easy interpretation. Each grid consists of nine boxes with the centre box being the actual source input of the model, and the remaining eight representing the activity in their respectively direction-selective cell, as indicated by the arrows. The plots shown can be considered as snapshots of cell activities after $1/25$ seconds (period between frames) of iteration of each source input.

Simulation I

In this simulation we present six greyscale images of size 32×30 to the neural model, and the results are summarized in Figures 6.10-6.12. An initial presentation of a contrast pattern would cause all direction-selective cells to fire, but due to their strength being equal, no cell would emerge as winner under competition, which is the case in Figure 6.10. Movement of

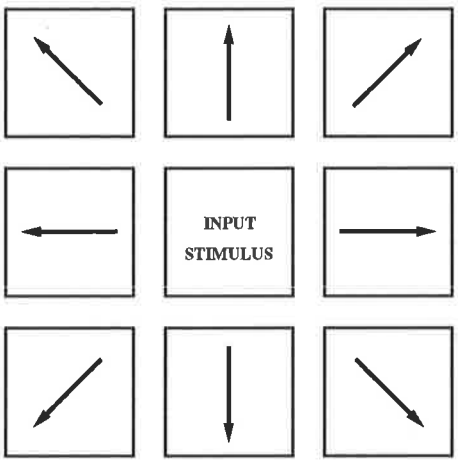


Figure 6.9: Motion direction representation.

the tennis ball in successive frames causes the activities in those direction-selective cells that have preferred directions close to its movement direction to build up. All plots have been scaled uniformly to display better contrast, this however does not affect the model's behaviour in anyway. The relative strength of the fields is an indication which directional cell is winning. It is difficult to tell in Figure 6.12, whether the "R" cell or the "DR" cell is stronger, but this can readily be determined by a winner-take-all competition. Presenting results in this "fuzzy" way allows us to observe the activity in each directional field.

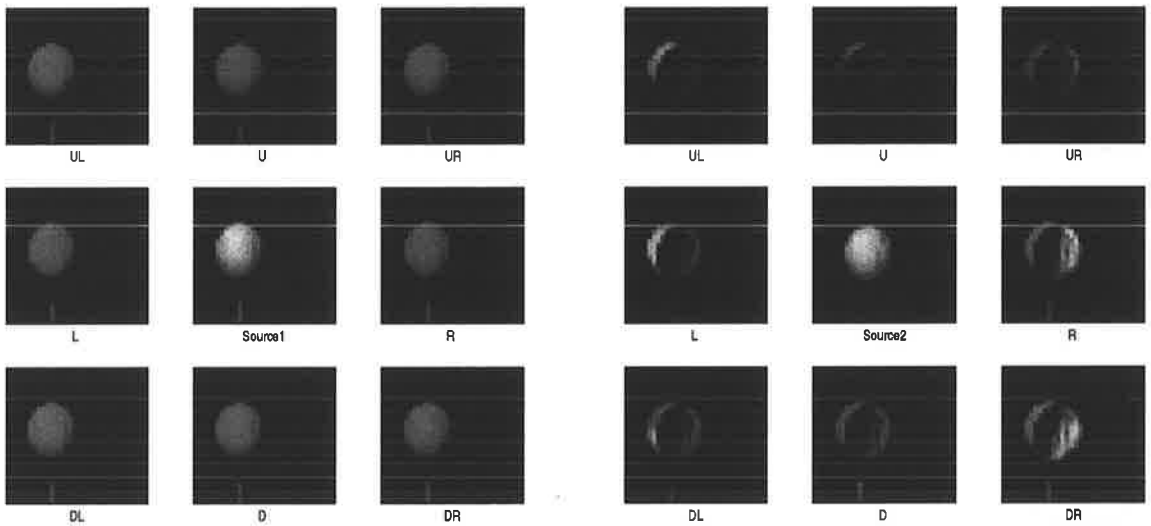


Figure 6.10: Ball frames 1 and 2. In the above U stands for up or upper; L for left; R for right; D for down.

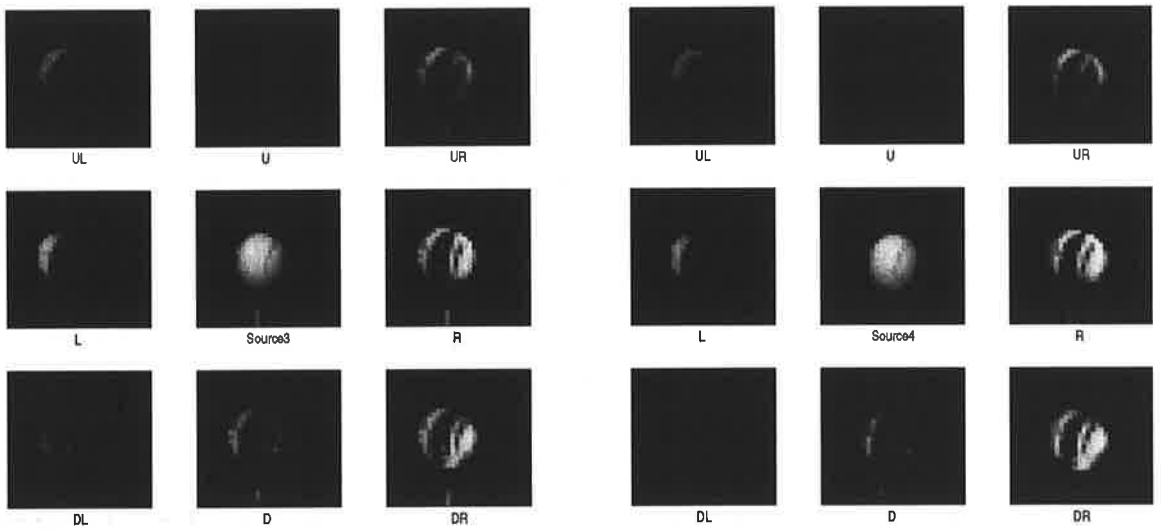


Figure 6.11: Ball frames 3 and 4.

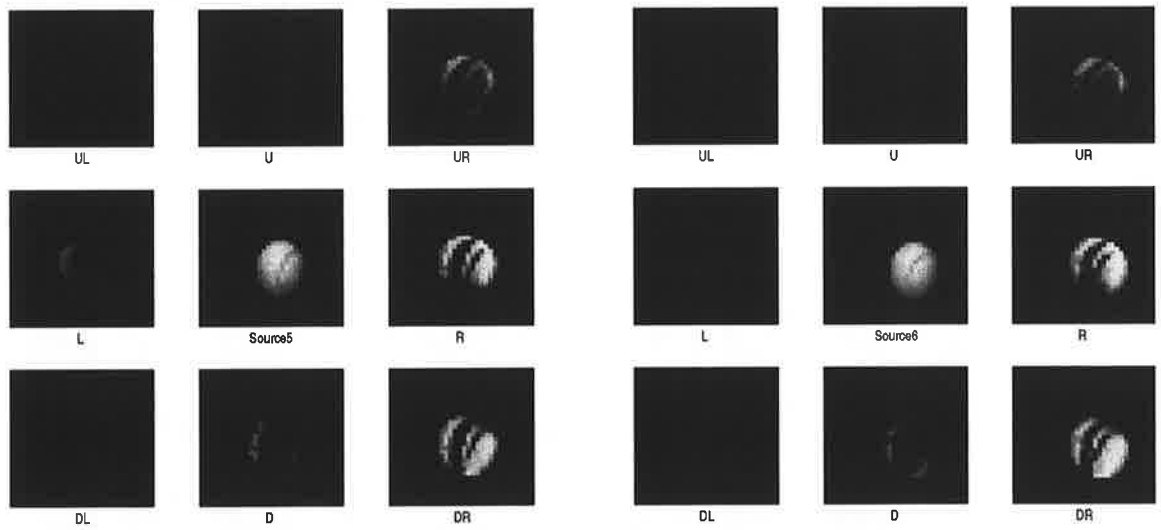


Figure 6.12: Ball frames 5 and 6.

Simulation II

This simulation allows us a close examination of individual pixels of the input frames. We have sixteen input frames of size 4×4 pixels, starting from Figure 6.13 to Figure 6.16. The first few frames contain a vertical bar moving from left to right, and the corresponding results show that the "R" cell is the strongest (by inspection). Source 6 in Figure 6.14 illustrates that model displays motion persistence property when the source is a blank image, the "R" cell is still the most active. From frame 7 to frame 9, a diagonal line of pixels is made to appear moving either up or to the right. The model responds by showing both "U" and "R" as possible winners. The rest of the results display similar characteristics. It is important to note that the results could be drastically different by changing the parameters in the dynamic equations as they govern the behaviour of the model. For example if the system is slow to react to changes, or motion signals are allowed to persist due to slow decaying rates, then inputing images with an object moving to the right and then down, will result in the activation of the "DR" cell. If this kind of residual effect is dominant, then the motion signals are either incorrect or delayed.

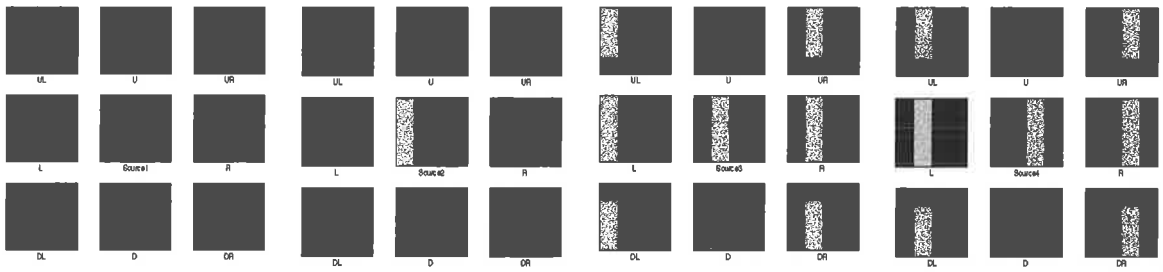


Figure 6.13: Pixel frames 1, 2, 3 and 4.

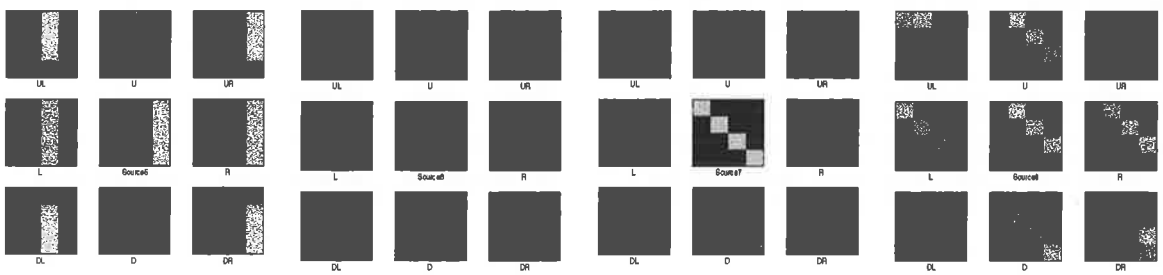


Figure 6.14: Pixel frames 5, 6, 7 and 8.

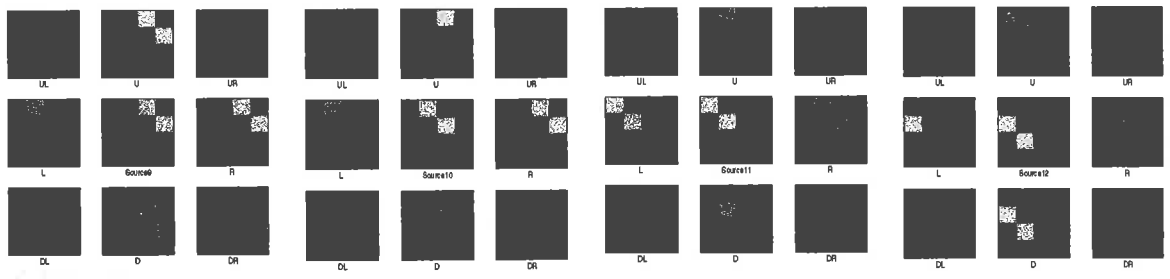


Figure 6.15: Pixel frames 9, 10, 11 and 12.

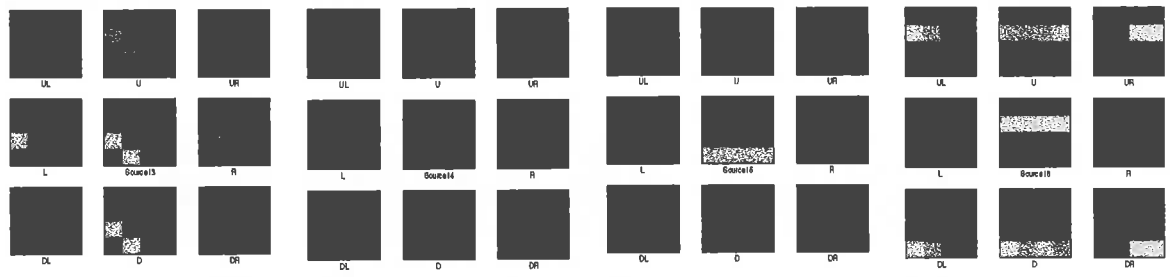


Figure 6.16: Pixel sequence 13, 14, 15 and 16.

6.4.2 Effects of Stimulus Contrast and Temporal Frequency

Contrast plays an important role in the coding of speed and direction in the visual system. It is thus a major concern to both physiologists for the understanding of the visual system, and engineers requiring to design models to account for its effects. Stimulus properties such as spatial frequency, temporal frequency, contrast and duration have a strong influence on the coding of motion signals in the human visual system [24, 26, 144, 173, 179, 184, 185, 205]. In particular, Stone and Thompson [179] have shown that human speed perception is contrast dependent. However motion sensitive cells in the primate's visual cortex do not detect speeds, but rather are directionally selective [124]. The contrast dependence of the motion sensitive neuronal response poses a serious problem to any speed-coding scheme in that any motion perception model based on the direction tuning property of the motion pathway will have neuronal responses partly arising from the contrast and partly from the speed of the stimulus.

This section intends to analyse the effects of stimulus properties such as contrast and temporal frequency on the motion module. Since the model is inspired by its biological counterpart, we need to examine the effects these stimulus properties have on the detected speed and direction of motion.

Effect of Input Stimulus Contrast

This simulation shows the effect of contrast on the model behavour, and demonstrates the role of facilitatory signals in enhancing the responsiveness of the model.

We have a number of gray-scale image sequences as inputs, each frame is of size 30×30 pixels, as shown in the rows of Figure 6.17. Each sequence depicts a tennis ball moving from left to right in the form of apparent motion at the same speed, but the contrast level of the sequences decreases from top to bottom. So in theory, if contrast played no part in the model's motion signal, then the responses stimulated by the inputs of Figure 6.17 would be the same. Basically this is an apparent motion simulation of the same sequence performed over a number of times with the contrast level varied each time.

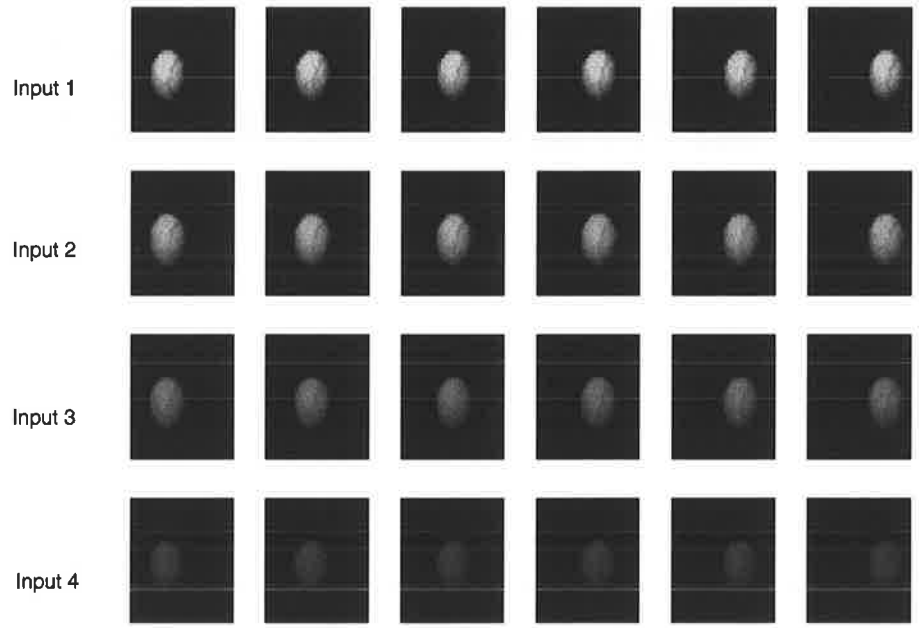


Figure 6.17: Inputs for contrast simulation.

The resultant global direction-selective cellular activities are plotted in Figure 6.18 with plots corresponding to rows of the inputs in Figure 6.17. It shows that a cell is excited whenever there is a change in the input, i.e., a movement occurred. Movements are characterised by the dips in the plots, which coincides with the change over of frames. The decaying behaviour of the curves is also indicative of the transient nature of a temporal system, such that if the frame rate is too slow the model will fail to respond. This is analogous to exceeding the ISI (interstimulus interval) limit, the model is unable to associate the two events as being motion. By comparing the cellular activities in Figure 6.18, we can see that the cellular activity drops as the contrast reduces. However plots for inputs 1 and 2 show, as expected, that contrast has minimal effect on the response when it has reached a certain level of contrast.

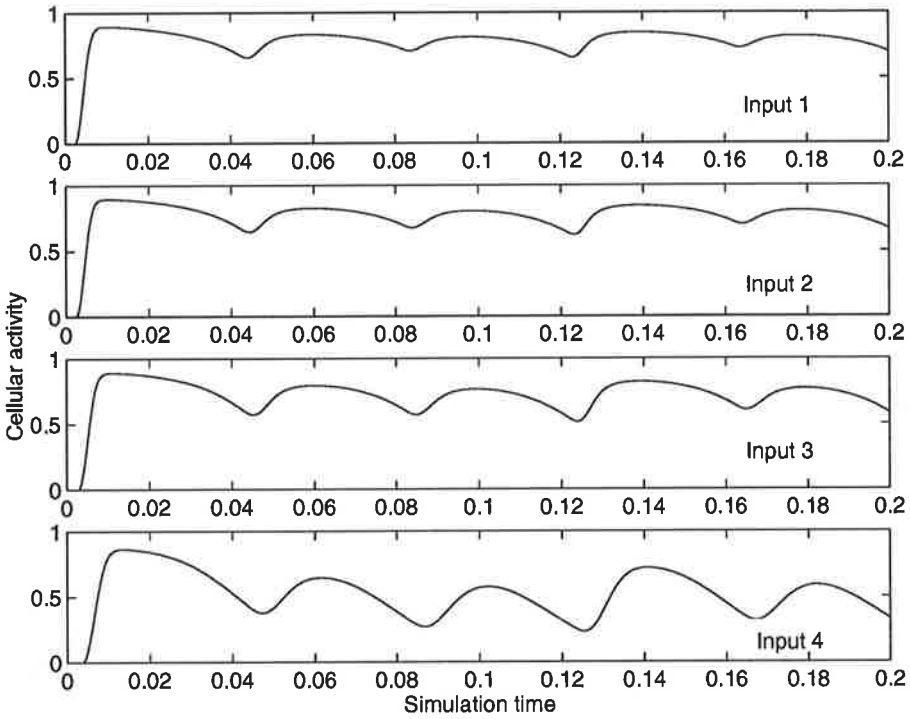


Figure 6.18: Motion cellular activities for various levels of contrast.

The same simulation was repeated for six contrast levels. The resultant cellular activities were converted to *root-mean-square* values and plotted against the input as a percentage of intensity of the original input sequence, shown in Figure 6.19. This percentage can be regarded as a form of contrast since the transient cells take the temporal difference as being the input, so by reducing the intensity level we reduce the contrast level. Figure 6.19 shows that the model output is significantly affected by the contrast of the stimulus. It also agrees with Stone and Thompson's result that higher contrast inputs yield faster perceived speeds as represented by the cellular activities here. The plot can be viewed as having three distinct regions. Along the y-axis between $0 - 0.1$ reflects the weak responses of low contrasted stimuli, $0.1 - 0.5$ is a relatively linear region characterised by its sharp rise in cellular activity, and the remaining region indicates that contrast is not having as a noticeable effect on the response as it saturates towards the limit. It must be pointed out that the exact location of the curve is strongly influenced by the parameters used in the model's dynamic equations, however this should not have too much effect on the general shape of the response.

Facilitatory inputs One of the main features of the motion module is the incorporation of facilitatory terminals [113] to all synaptic connections, enabling the modulation of transmitted signals for gain control purposes. The effect of providing facilitatory signals is illustrated by Figure 6.20 for facilitation levels of 0.9, 0.5 and 0.3 from top to bottom, respectively.

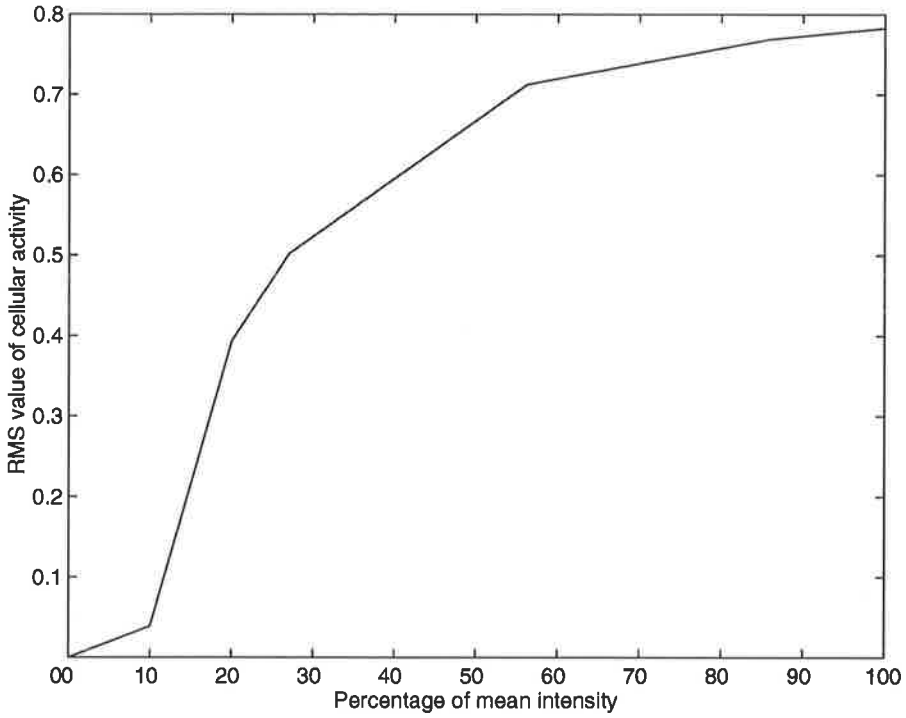


Figure 6.19: Effect of contrast level on cellular activity.

It shows that modulation of synaptic signals via facilitation can cause the model to behave more responsively for a range of contrast levels. In particular, outputs of input sets of mean intensity above 10% are significantly enhanced by facilitatory inputs. For a given input stimulus, with known initial conditions of the model, the facilitatory signals can be used to tune the model to provide an appropriate output.

Tuning Characteristics of Transient Cells

The transient layer is vital to the functioning of the system. Understanding of its characteristics is crucial to further extensions of the model. The transient layer as mentioned before is responsible for temporal changes in the input stimulus. These changes can be either in continuous or discrete form as described in the following sections.

Continuous Input Stimulus This section examines the effect of gradual continuous change in the rising rate of the input stimulus on the transient layer. Input stimuli shown in the bottom graph of Figure 6.21 are presented to the transient layer to produce transient signals as shown in the top graph of Figure 6.21. Each input is characterised by the angle at which the signal rises from zero to one, thus this angle can be used as a measure of the rate of increase in the intensity level. From the graph, the angle can vary between 0° to 90° with 0° being the slowest

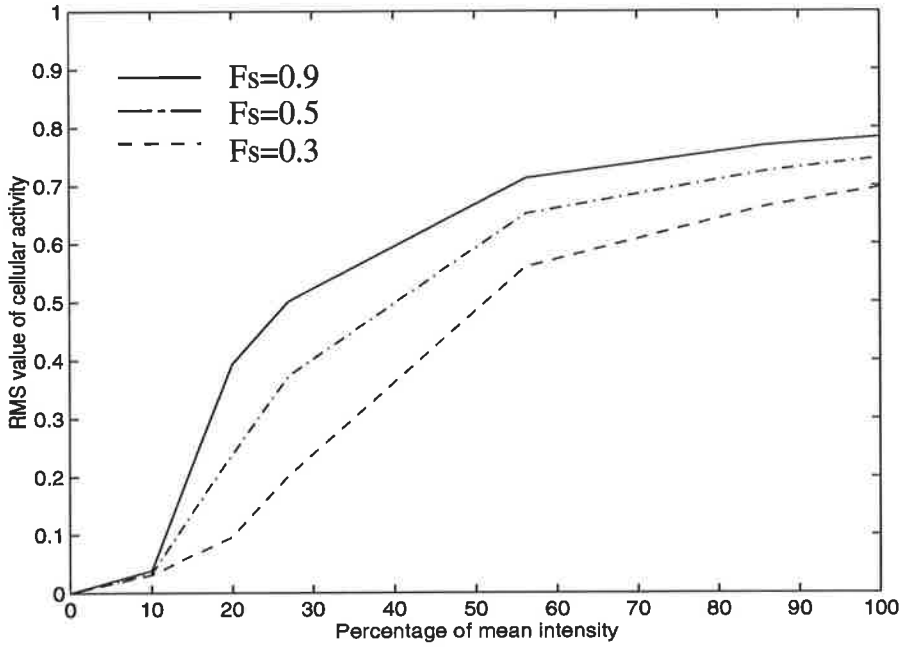


Figure 6.20: Effect of facilitation on contrasted inputs.

rate of increase and 90° the fastest, representing no rise in intensity and an instantaneous rise, respectively.

The transient cell outputs produced are a series of pulses with their shapes determined by the input stimulus rising rate. Slow rising stimuli have small and flat pulse shaped outputs, in contrast fast stimuli have tall and sharp pulse outputs. While fast stimuli produce large output signals, they tend to decay away in a short period of time. In order to maximise the final motion signal the model requires the transient signals to be both large in magnitude and persistent in time. The results are therefore replotted in another representation by taking the root-mean-square value of the transient response to accommodate for time and magnitude as a function of the input rate, shown in Figure 6.22.

Figure 6.22 consists of four plots of transient output as a function of input stimulus rate for four different levels of facilitation. In descending order of magnitude the facilitation levels are 0.9, 0.6, 0.3 and 0.2. They all exhibit the same general characteristics that no output is generated when the angle is small, and the signal peaks before 90°. The effect of reducing facilitation can be seen as shifting the transient curve down and slightly to the right. This is expected since the role of the facilitatory signal is to provide a mechanism under which the transmitted signal can be modulated to achieve gain control.

Discrete Input Stimulus However, for apparent motion the changes in intensity are sudden, non-continuous and discrete, i.e., from frame to frame. We examine how the transient layer

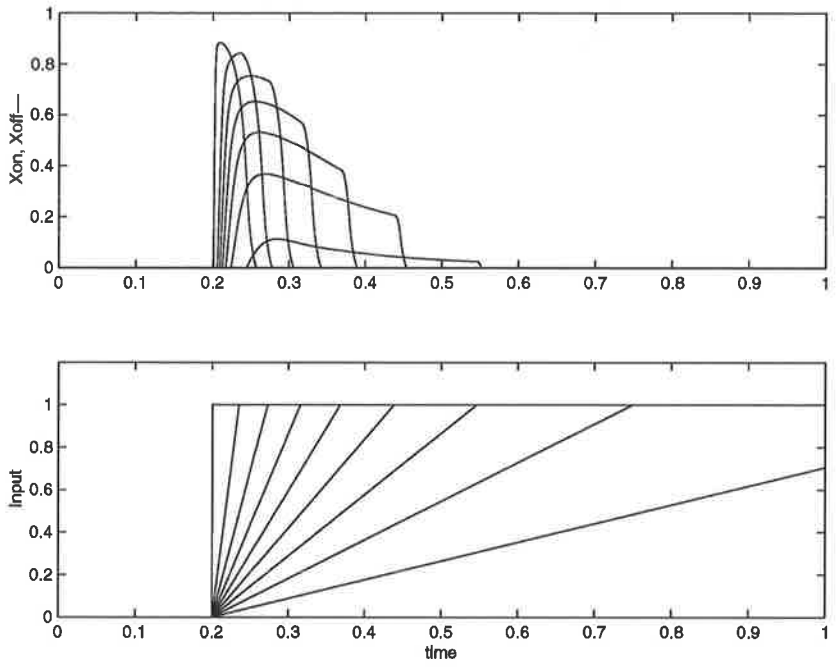


Figure 6.21: Continuous change in stimulus speed.

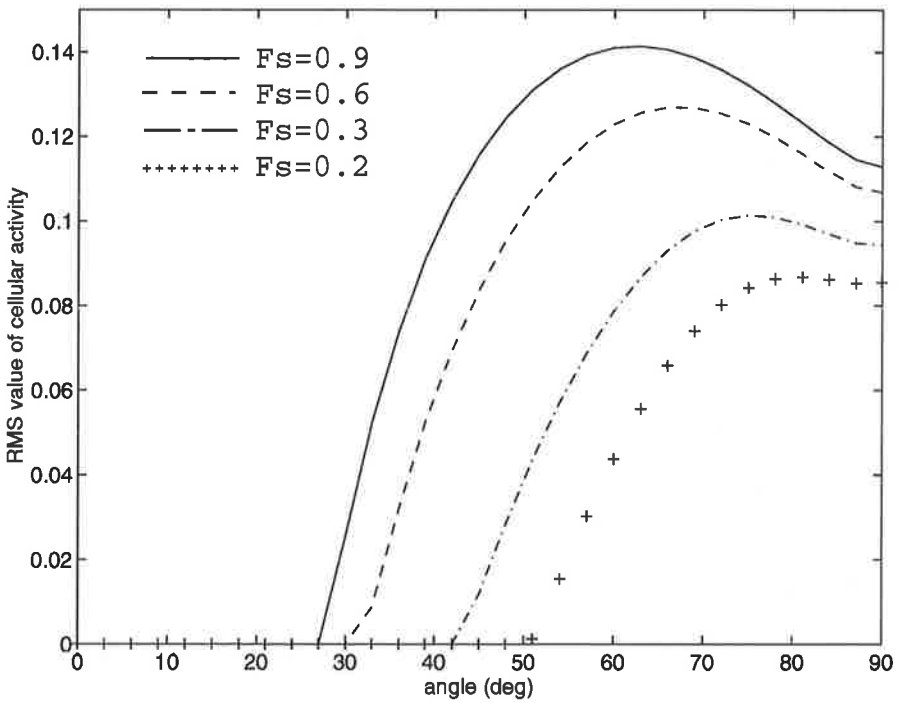
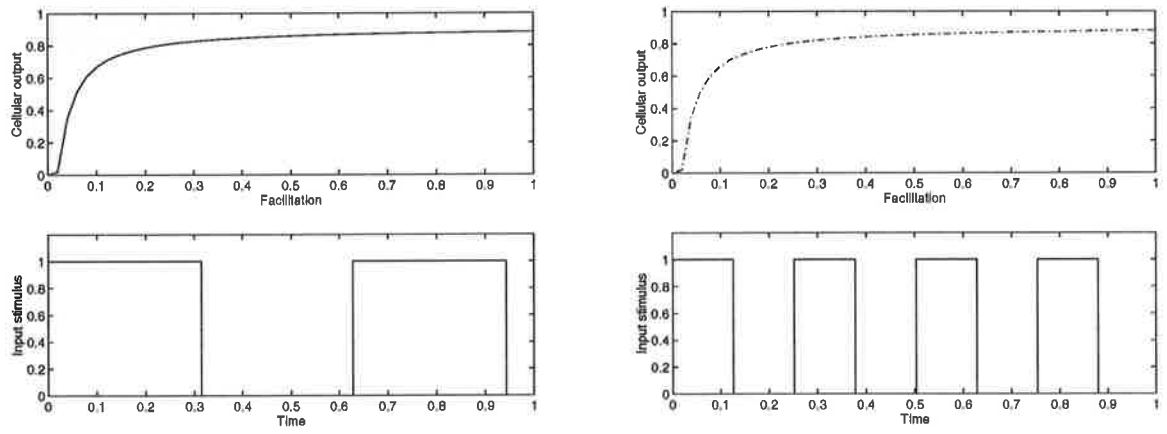


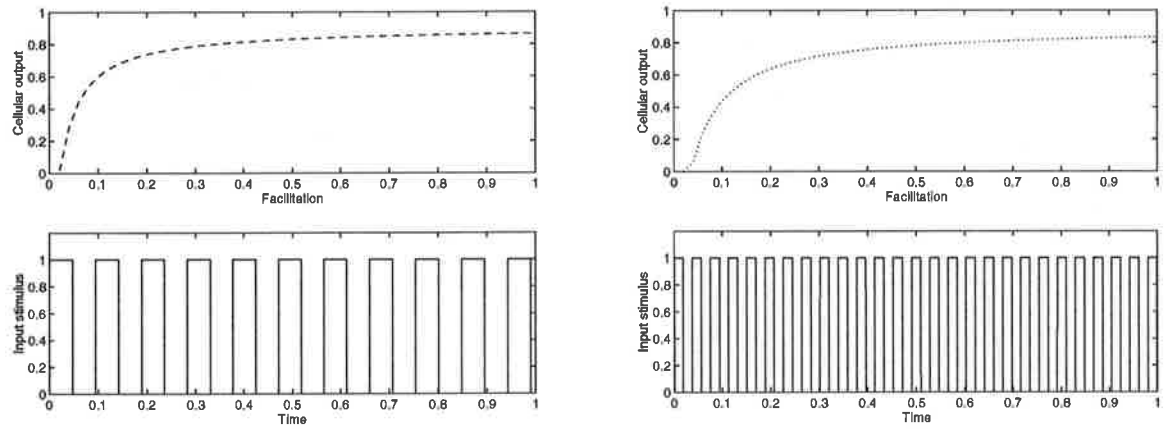
Figure 6.22: Effect of facilitation on variable rated inputs.

responds to variations in the speed of the input stimulus as a function of the facilitation level. Figures 6.23-6.25 show six sets of results. Within each set, the top graph is the maximal transient cell response at each level of facilitation to the input stimulus shown below it. The speed of the input is defined as the frequency of rectangular input pulses oscillating between zero and one. All the simulated speeds have frequencies that are multiples of the numerical time step h, which are specified at the bottom of each figure.



(a) (b)

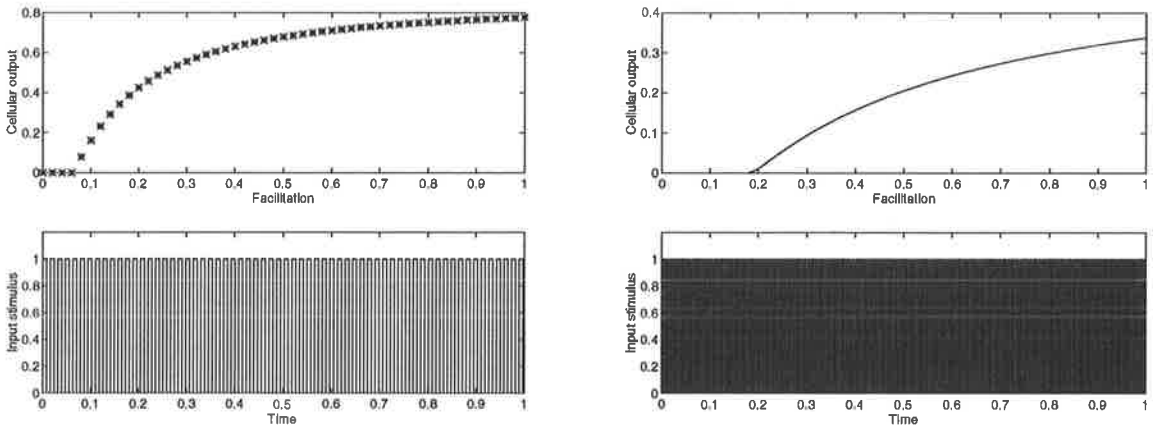
Figure 6.23: Transient cell characteristic curve with input stimulus speed dependence. (a) 1000h, and (b) 400h. The rectangular pulse input speed is measured in terms of the numerical time step h.



(a) (b)

Figure 6.24: Transient cell characteristic curve with input stimulus speed dependence. (a) 150h, and (b) 60h.

Figure 6.26 reveals that there are two limits by which the transient cell output is bounded. The upper limit is approached when the input stimulus is sufficiently slow, allowing the transient cell to respond fully to the input. The lower limit is reached when the input speed is too fast for



(a) (b)

Figure 6.25: Transient cell characteristic curve with input stimulus speed dependence. (a) 25h, and (b) 10h.

the transient cell to respond. The role of the facilitatory signal is to make the transient cell more responsive and receptive to faster input stimuli and to provide an external gain. As evident from the graphs, the transient response tends to saturate as the facilitation level increases.

These characteristic curves can be used to tune receptive fields to obtain the desired output signals. For example, given a system is tuned for movements in a particular direction, facilitatory signals can be provided to all the bottom-up inputs gated by the transient cells responsible for detecting movements in the chosen direction.

When the data are plotted in a different format as in Figure 6.27, it shows that there is an optimal temporal frequency at which over a period of time the average (RMS) transient activity is at its peak. This frequency occurs when the transient cell is allowed appropriate time to respond and decay, hence resulting in strong direction-selective signals. If the temporal frequency is too low, there are long periods between transient pulses as given in Figures 6.4-6.5. This can result in weak or no direction-selective signals. Whereas high temporal frequencies do not provide enough time for the transient cell to respond, therefore transient changes may not be detected at all, resulting in no direction-selective signals.

6.4.3 Directional Bias

Two apparent motion simulations are performed on the proposed network architecture, one with selective attention and one without, to demonstrate the effectiveness of the scheme. The test visual stimuli are 32 $\times$ 32 pixel gray-level images, each with two rectangular bars, one vertical and the other horizontal. These bars are displaced slightly in successive frames, creating movements in apparent motion, with the horizontal bar moving upwards and the vertical to the

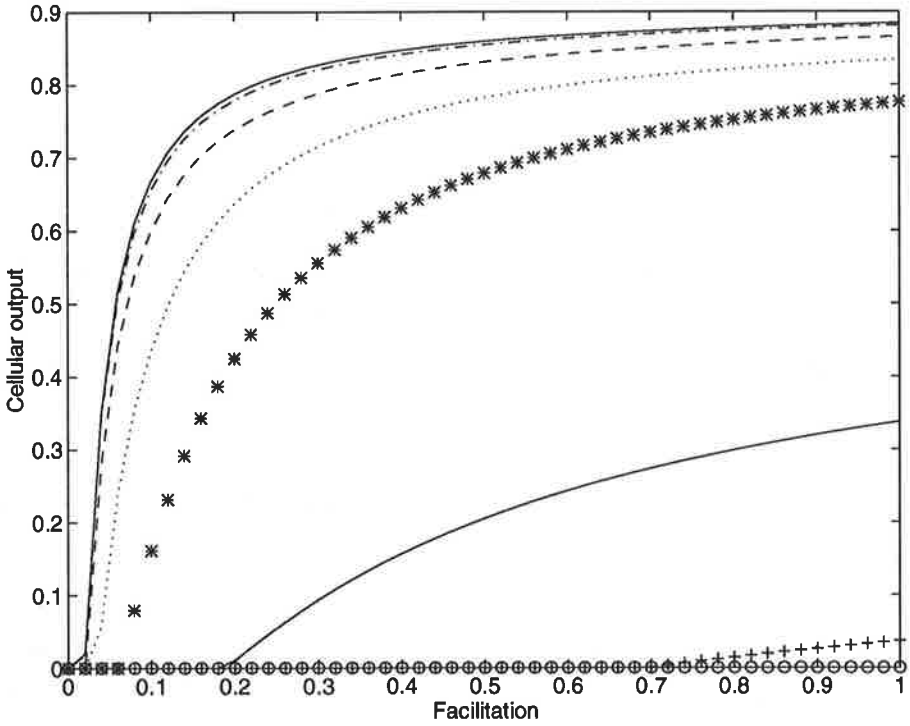


Figure 6.26: Transient cell characteristic curve with input stimulus speed dependence. The plots (from top to bottom) are for speeds of 1000h, 400h, 150h, 60h, 25h, 10h, 5h, and h.

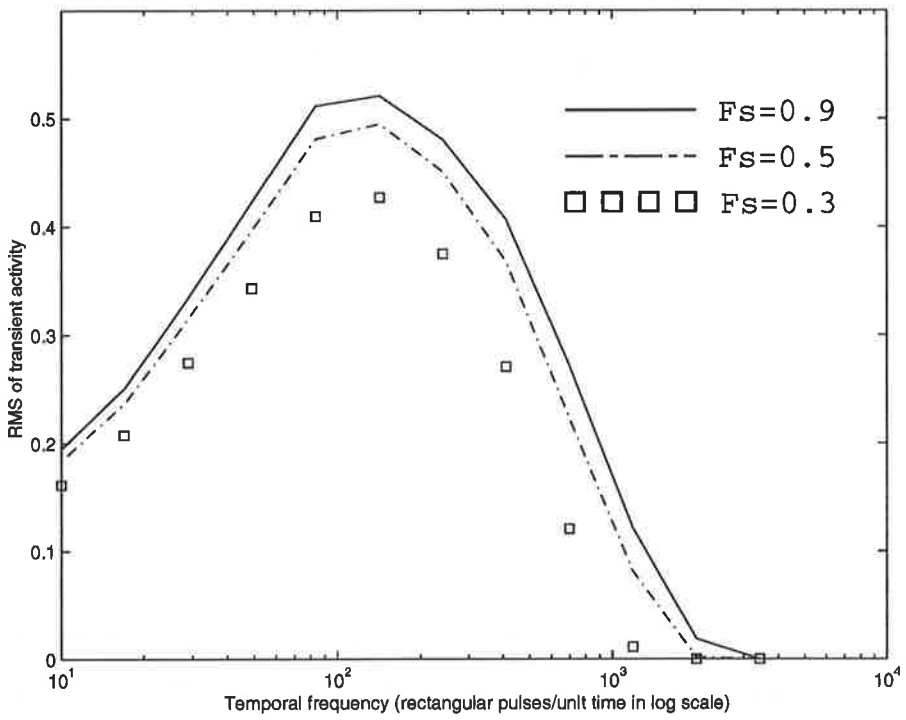
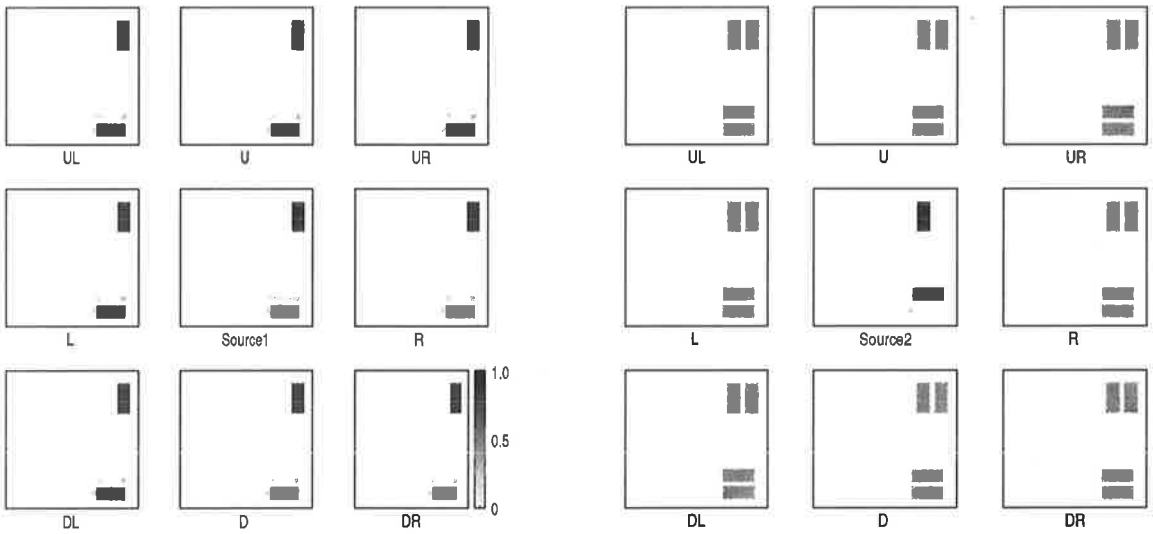


Figure 6.27: Root-mean-square value of transient cell activity as a function of log speed and facilitation.

left. The simplistic visual stimuli were designed to clearly display the effects of the top-down directional bias.

The simulation results are summarized in Figures 6.28-6.31. Each figure consists of nine squares of size $N_i \times N_j$, with the centre one being the source and the surrounding eight the directional fields. The labels beneath the boxes indicate the preferred directions of the directional fields. Figures 6.28 and 6.29 are the results of elementary movements without directional bias. The results can be interpreted as the effects of the preattentive process - that it detects any unusual events in the visual scene, in this case the detection of stimulus movements. Starting from Figure 6.28 we can see that when the stimuli first appear they excite all directional fields equally, and as the frames change over time only the directional fields in the movement directions receive facilitation and gradually become stronger. The inter-competition amongst directional fields is evident in the fading of the activities in opposing directional fields. In each active directional field there appears to be two images of the moving bars, this is caused by the residual activities remained from previous movements, and can be regarded as the short-term memory. Figures 6.30 and 6.31 are the results of a control selection process initiated by the top-down attentional system having a directional bias towards movements to the left.



U-up, D-down, L-left, R-right

Figure 6.28: Moving bars frames 1 and 2.

Notice how at the same stage 'source4', shown in Figures 6.29 and 6.31, the downwards 'D' directional fields are different, with one inactive and the other active. Because the directional cells are excited as long as there are moving stimuli, the strength of these activities depends on how well the stimulus properties match the cells' receptive fields. That is how well the movements match the preferred directions. Since none of the movements matches the downwards directional fields' receptive field, the resultant activities are weak. However, the directional fields are

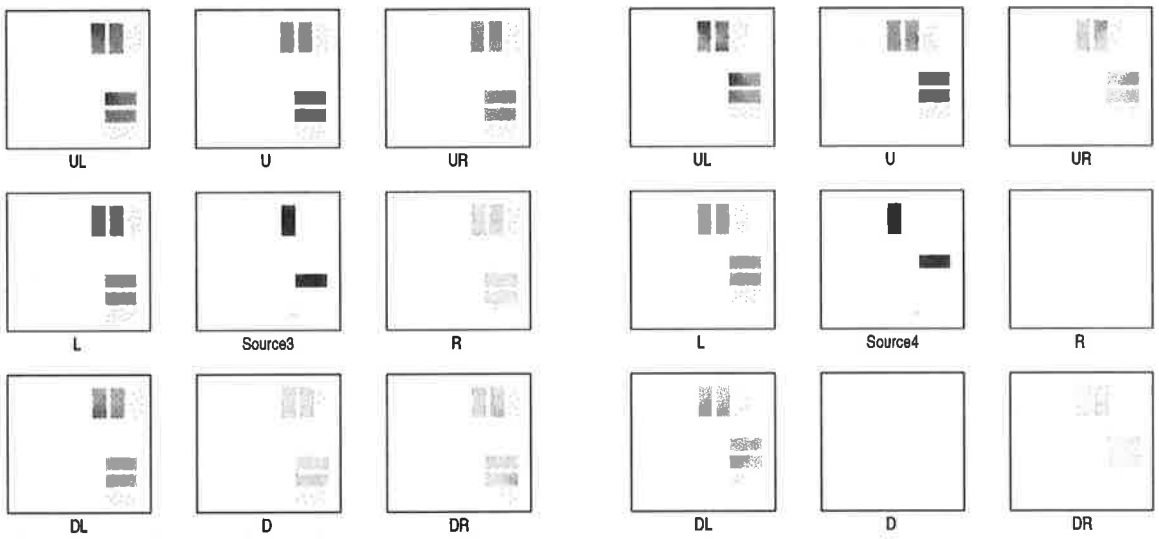


Figure 6.29: Moving bars frames 3 and 4.

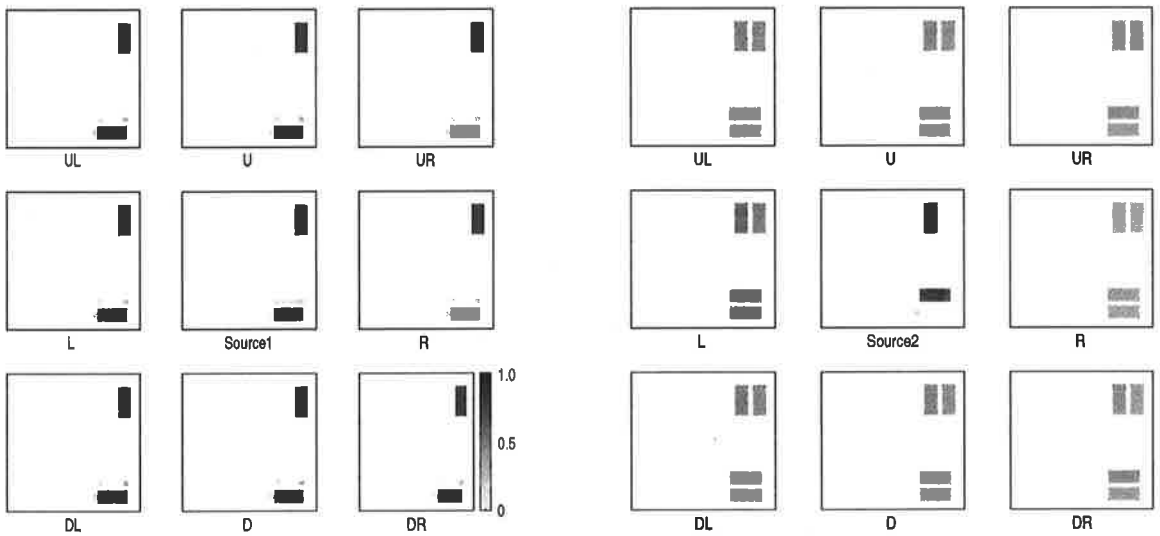


Figure 6.30: Moving bars frames 1 and 2 with directional bias.

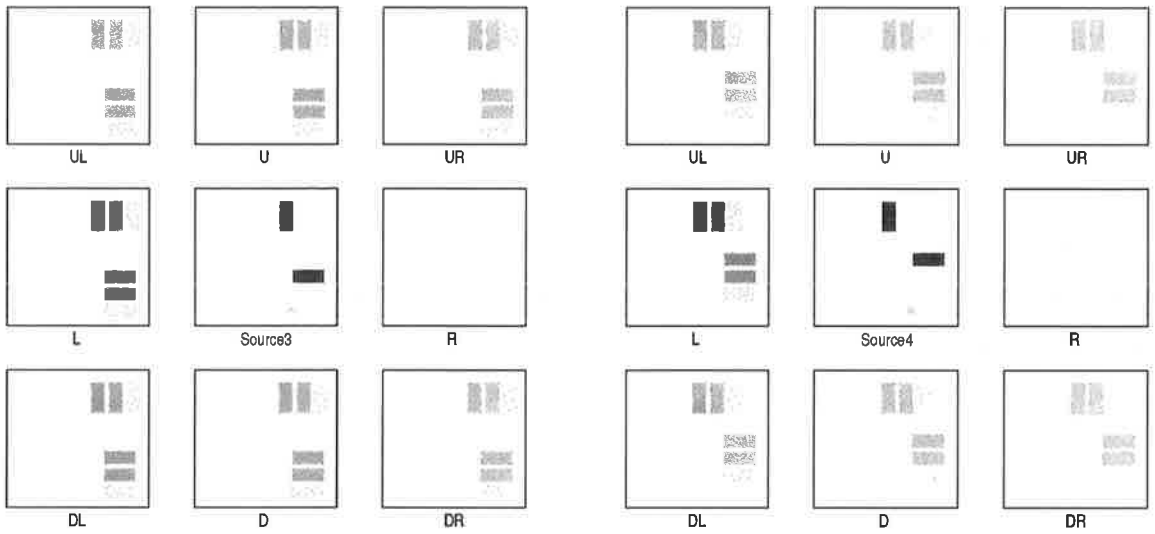


Figure 6.31: Moving bars frames 3 and 4 with directional bias.

engaged in an inter-competition for directional supremacy between opposing directional fields. With strong upwards activities in Figure 6.29 the downwards field is being suppressed by a stronger force, and hence the absence of activity. While in Figure 6.31 the upwards activities are similar to that of the downwards, either one is not strong enough to suppress the other.

6.5 A Visual Motion Cue for Recognition of Moving Objects

Motion has long been recognised as one of the basic perceptual features that can be used as a visual cue to capture attention [19]. As mentioned in Chapters 3 and 4, attentional capture is performed by the preattentive mode using bottom-up features. So far, the proposed recognition model uses luminance contrast as its sole input feature, however it is important to include other features. In this section, the motion module presented in this chapter is incorporated into the framework for the detection, location and recognition of moving objects.

The main objective of the proposed incorporation is to use it as a pilot study to test the feasibility of integrating the static recognition system with the dynamic motion module. For that reason the integrated system is relatively primitive, and as will be seen, the system is tested on a simple synthetic image sequence, featuring line-drawing objects. Nevertheless, it is an important step towards the development of a dynamic visual scene analysis system.

A diagram for the framework is shown in Figure 6.32. It can be seen that the middle to top parts are the same as in Figure 4.14. The motion module appears early in the processing hierarchy. It takes input image frames continuously, from which motion information is inferred. In the meantime, the same image frames are processed by a reverberating memory loop, consisting

of two STM fields. The purpose of this memory loop is to stabilise the input pattern sent to the input field as shown in Figure 6.32, because the static recognition is considered to be much slower than the motion module, thus a buffering mechanism is required to regulate the rate of change of the input signals.

The transient signals generated by the motion module are used to gate the bottom-up stimuli from the input field to the high activity field such that regions with movement are allowed through and the rest are filtered out. The system proceeds to analyse those regions as in static object recognition.

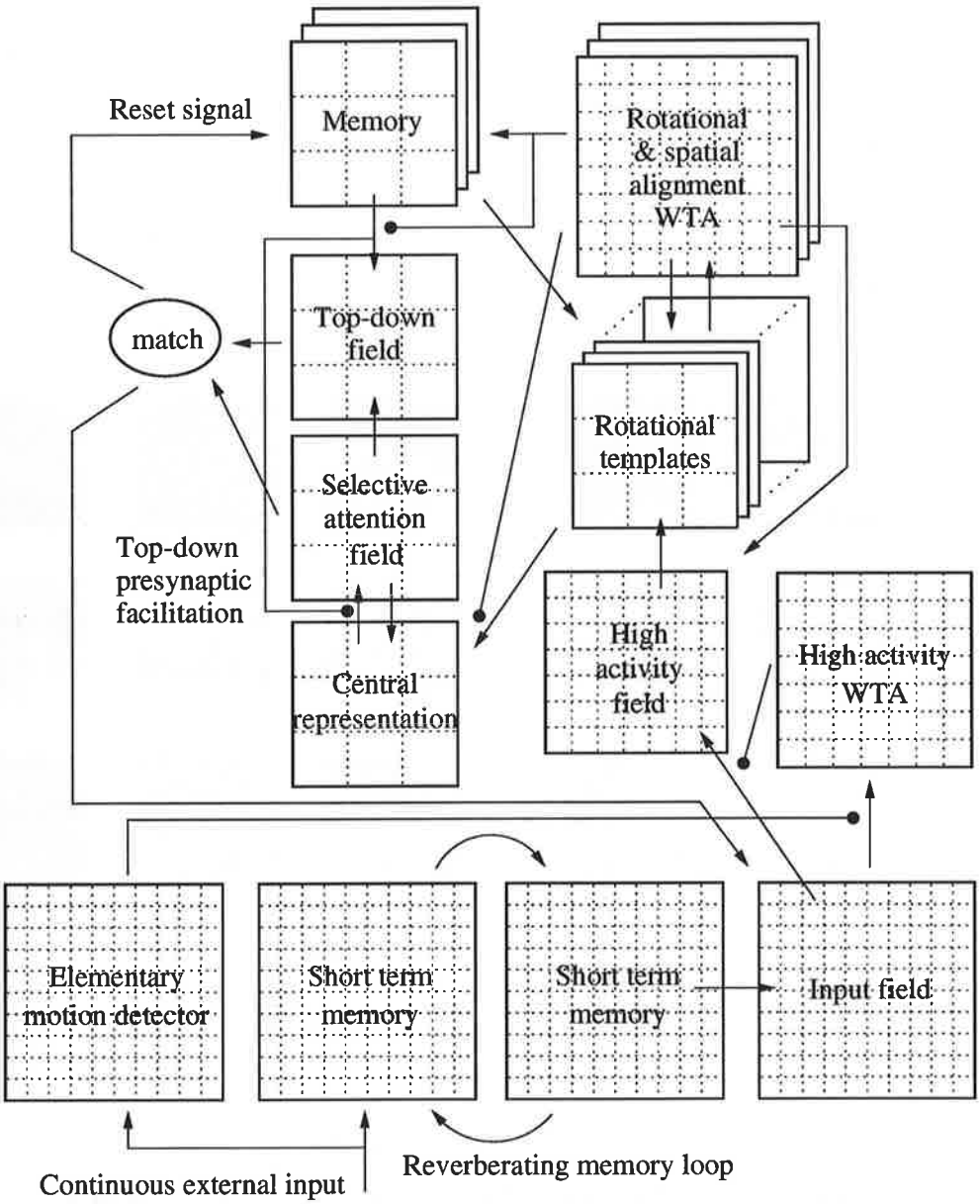


Figure 6.32: Visual object recognition and selective attention framework for moving objects.

6.5.1 A Test Case

A simple sequence with line-drawing objects has been chosen for simulating the integrated system. The input image sequence consists of two static objects, a rectangle and a triangle, and a moving object, a circle, in apparent motion sense. The objects can be seen in the middle patterns in Figures 6.33-6.34. There are five frames in the sequence. In which the rectangle and the triangle remain stationary while the circle moves from left to right. Each frame is 32×32 pixels in size and the objects are 7×7 pixels, with a high activity region of 15×15 pixels.

Results of the motion-direction detection and computation are shown in Figures 6.33-6.34. As before, when an input is first registered all directional fields are activated equally, thus no overall direction of movement. As the number of frames processed increases, the directional field corresponding to the movement direction continues to gain in strength as shown in Figure 6.34. Since the recognition system is slow relative to the motion module, the system has been set such that it analyses the input field for every three frames processed by the motion module.

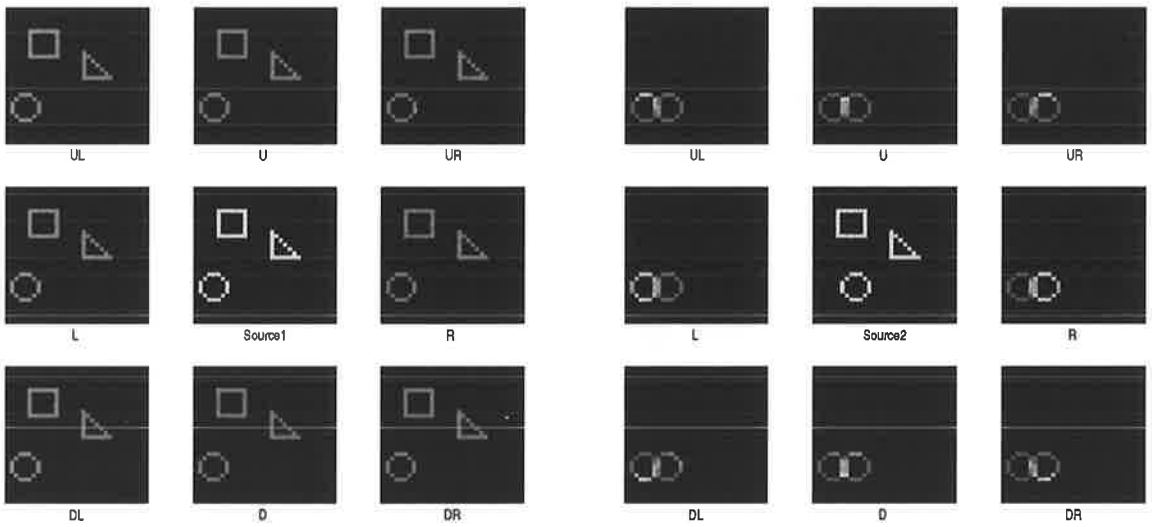


Figure 6.33: Motion cue frames 1 and 2.

Figure 6.35 shows the results of the recognition process. The input field shows the pattern generated by gating the input frame with the transient signals from the motion module. The system proceeds to select a region of interest from the input field. The region is analysed in the usual manner, and as a result of this analysis the moving circle is recognised as one of the stored models in memory.

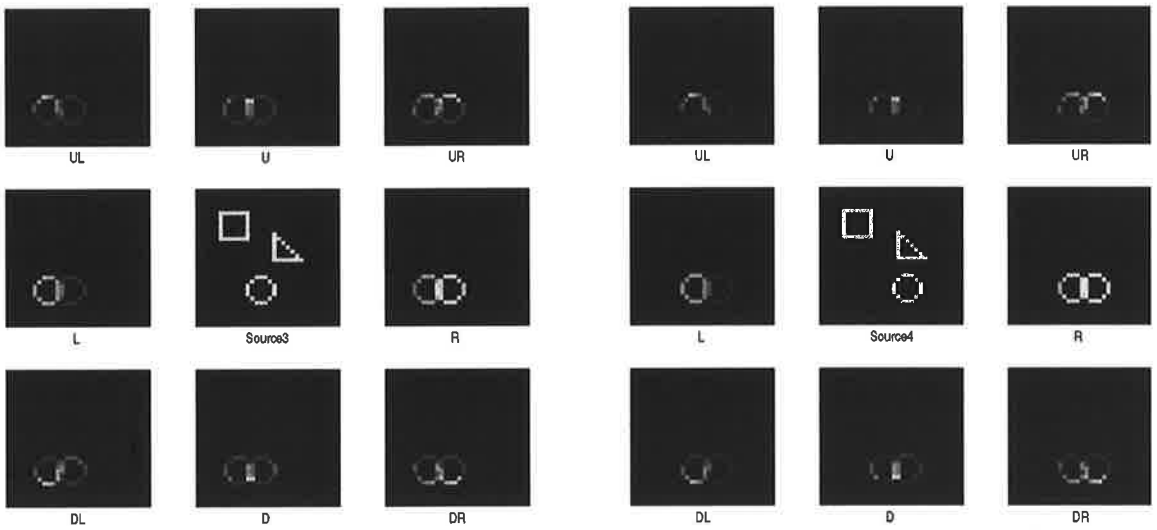


Figure 6.34: Motion cue frames 3 and 4.

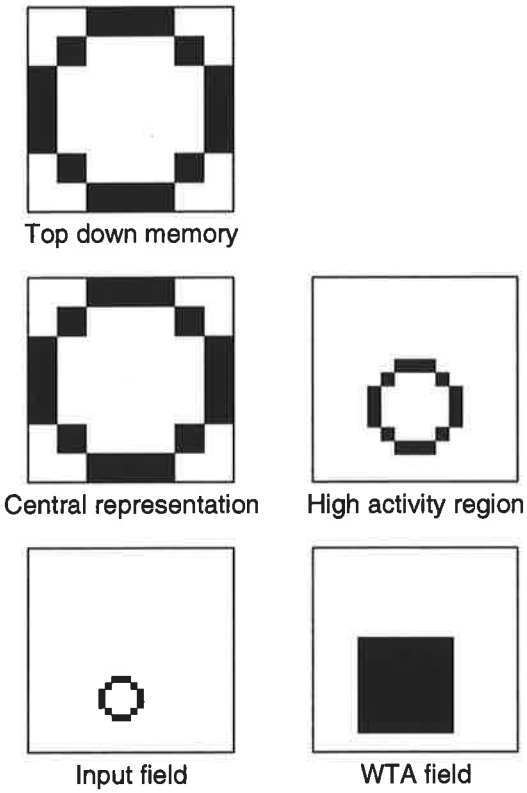


Figure 6.35: Recognition of a moving object.

6.6 Conclusions

A brief overview of the motion pathway was provided in this chapter, which led to the presentation of a biologically inspired neural architecture for motion-direction detection and computation. The entire model is based on the same fundamental building blocks as for the visual scene analysis system presented in Sections 4.3 and 3.4.

The motion module respects key neuro-psychological and physiological findings. The main features of the model are:

- sensitive to direction-of-motion but insensitive to direction-of-contrast;
- cooperative and competitive nature of directional cells;
- possession of facilitatory terminals allowing top-down attentional modulation, the implementation of directional bias is an example of the usefulness of presynaptic facilitation;
- an implementation of the transient ganglion cell based on [76, 145] for detecting transient onset and offset; and
- a complex interconnected synaptic network for modelling the direction-selective layer.

Simulation results have demonstrated the motion module to be effective in detecting and computing motion direction. The effects of stimulus contrast and temporal frequency were also investigated. Top-down attentional modulation has been successfully applied to achieve directional bias.

More importantly, a preliminary study on the integration of the static recognition system with the motion detection architecture has been successful. This demonstrates the flexibility and extendibility of the framework architecture, and it provides insights for extending the model for motion perception.

Chapter 7

Advanced Framework Features

7.1 Introduction

This chapter presents several advanced features that can significantly improve the performance and capability of the neural architecture proposed in Chapter 4. These new features are introduced to overcome some of the deficiencies of the model in the areas of partial object recognition, robust automatic attention and size invariance.

7.2 Complementary Selective Attention Adaptive Resonance Theory

Objects are commonly seen only partially. A recognition system should be able to tell a part of an object is not a new object but merely a subset of an established object category. To solve this problem, the system must be able to activate a stored representation based on its parts, and reconcile the differences between the two.

Anatomical studies have provided evidence that massively parallel feedforward and feedback connections exist along the visual pathway [194]. Throughout the visual areas, all neural connections are matched by reciprocal feedback connections, hence there are reasons to believe the feedforward and feedback connections are complementary in nature [124, 23]. Although the precise uses of the feedback pathways are unclear, some have suggested these are directly involved in priming and attentional modulation of the bottom-up pathway, and in the direct activation of a lower area [131, 191], several experimental studies [50, 172] have produced evidence supporting the theories. Regardless of their exact nature, feedforward-feedback interactions are

essential in neural processing. In this section, we propose an extension to ART and SAART that employs feedforward and feedback pathways for modulating top-down and bottom-up patterns. As a result, the patterns can adapt to each other in a closed-loop manner, allowing incomplete objects to be recognised from cluttered images in the context of 2D shape-based object recognition. Since the proposed architecture is based on ART [28, 29, 30], it is called *Complementary Selective Attention Adaptive Resonance Theory* (CSAART) neural network. Significantly, the model can be used to explain the phenomenon of visual completion and why perceptual grouping of fragmented object parts under occlusion is more effective when the occlusion cues are strong.

7.2.1 Complementary Feedforward-Feedback Interactions

Conceptually, the bottom-up feedforward and top-down feedback interactions may be modelled in an architecture as shown in Figure 7.1. The architecture consists of four fields of network nodes, labelled as $Input$, $F0$, $F1$, and $F2$, and a pair of adaptive filters between $F1$ and $F2$. All four fields are for storing short-term-memory (STM) patterns, in particular, $F2$ is a winner-take-all (WTA) network where at most one node can be active at a time. The active node in $F2$ represents the currently activated category in memory. An incoming signal $\mathbf{x}$ is represented as a neural pattern $\mathbf{x}^{F0}$ across $F0$. The pattern $\mathbf{x}^{F1}$ across $F1$ is the result of the combined neural activities between the bottom-up input and the top-down activated memory from the kth node in $F2$.

The main features are the complementary feedforward and feedback modulatory pathways, which serve to provide synaptic gain control to the top-down and bottom-up neural patterns. When triggered, the top-down feedback modulatory pathway facilitates $F0$ cells that form the desired neural pattern by increasing their synaptic gain. At the same time, the bottom-up feedforward modulatory pathway gates the top-down neural pattern such that only common parts of the patterns can reach $F1$. With the introduction of the modulatory connections, the system is able to resolve any minor differences between the top-down and bottom-up patterns by adapting to each other. Therefore, an incomplete or occluded familiar object can be recognised as a part or parts of its stored model in memory represented by a category node in $F2$.

The complementary selective attention adaptive resonance theory architecture is an extension of ART and SAART. By removing the bottom-up feedforward modulatory pathway, the architecture reduces SAART (see Section 3.4), which has the ability to perform top-down perceptual grouping by filtering out irrelevant input stimuli. Further removal of the top-down feedback modulatory pathway, results in the original ART (see Section 3.3).

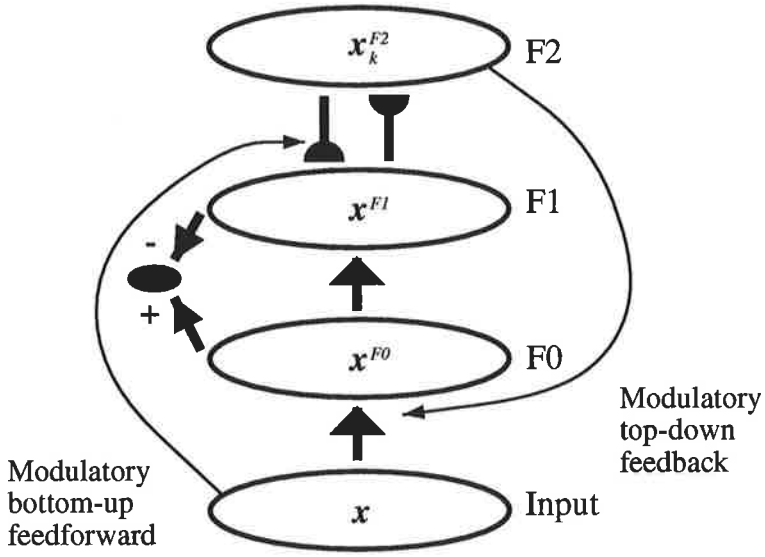


Figure 7.1: Complementary feedforward-feedback interactions.

To illustrate the feedforward-feedback interactions, two situations are considered in Figure 7.2. In Part(a), an input pattern that resembles a part of a learned object is partially occluded, and in Part(b), a familiar object is occluded by some irrelevant stimuli. In both situations the depicted networks are in resonant state.

Figure 7.2(a) shows that an input pattern is incomplete with parts missing. When the input $\mathbf{x}_1$ is presented to the architecture, it propagates through the fields $F0$ and $F1$ to activate a category node in $F2$. Consequently, a stored category is sent to $F1$ through the top-down adaptive filter, together with the bottom-up pattern a STM pattern $\mathbf{x}_1^{F1}$ is formed. Matching between STMs $\mathbf{x}_1^{F0}$ and $\mathbf{x}_1^{F1}$ is considered a success if the degree of match satisfies a matching criterion determined by a dimensionless *vigilance parameter*. Since the missing parts (due to occlusion and incompleteness) can cause matching to fail, an additional matching criterion, based on a *secondary vigilance parameter*, is required to initiate feedforward-feedback modulation in order for the patterns to adapt to each other. In this case, the recalled memory pattern $\mathbf{x}_k^{F2}$ is gated by $\mathbf{x}_1$, so that only parts of the memory pattern that match $\mathbf{x}_1$ are allowed through. Concurrently, the feedback modulatory signal gradually filters out the irrelevant parts from the input as in $\mathbf{x}_1^{F0}$. The feedforward and feedback processes continue in a closed loop fashion until the two patterns have adapted to each other, and resonance is attained.

Consider Figure 7.2(b), where the missing parts are caused by occlusion alone. This situation can be dealt with using the feedback modulatory pathway. In this case, the top-down recalled memory is used to facilitate the bottom-up pattern and non-common elements are suppressed under lateral inhibitory competition, resulting in a filtered version of the input to appear in $F0$ that matches the top-down pattern in $F1$. Hence resonance is established. It should be noted

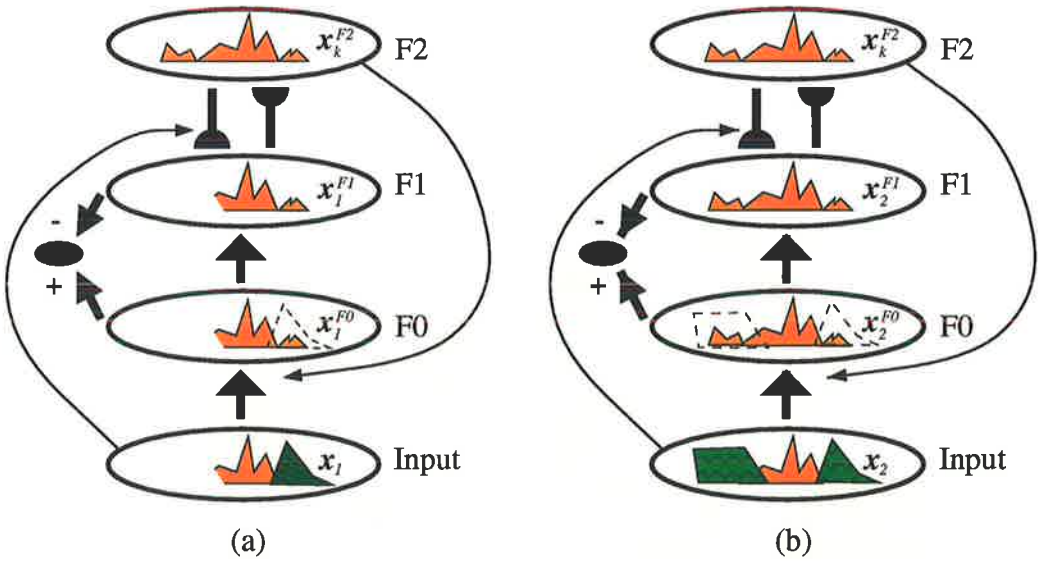


Figure 7.2: Visual recognition with feedforward-feedback interactions. (a) Bottom-up seeing, and (b) top-down imagining.

that the feedforward modulatory pathway is also triggered at the same instance as the feedback, however as shown in Figure 7.2(b) the gating pattern ($\mathbf{x}_2$) is a superset of the gated pattern ($\mathbf{x}_k^{F2}$), thus the whole of $\mathbf{x}_k^{F2}$ is passed to $F1$ unaffected.

7.2.2 Network Implementation

To illustrate the concept of feedforward-feedback interactions, the CSAART network is implemented for self-organising 2D object recognition. The implementation is based on an ART2 architecture [29], and with the feedforward and feedback pathways incorporated as shown in Figure 7.3.

Five major components are required to be modelled for the implementation of CSAART. These are short-term-memory (STM) loops, a winner-take-all (WTA) neural layer, feedforward-feedback modulatory pathways, long-term-memory (LTM) adaptation, and a matching mechanism. All the stages have been described in parts in Chapters 3 and 4, except the feedforward-feedback modulatory pathways. For self-containment, they are briefly summarised below.

STM Equations

Variables w_i, x_i, v_i, u_i, q_i, p_i, I_i, w_{o_i}, x_{o_i} and v_{o_i} in Figure 7.3 represent STM activities. Among them x_{o_i}, I_i, x_i, u_i and q_i are normalised activities of their respective input activities. It follows

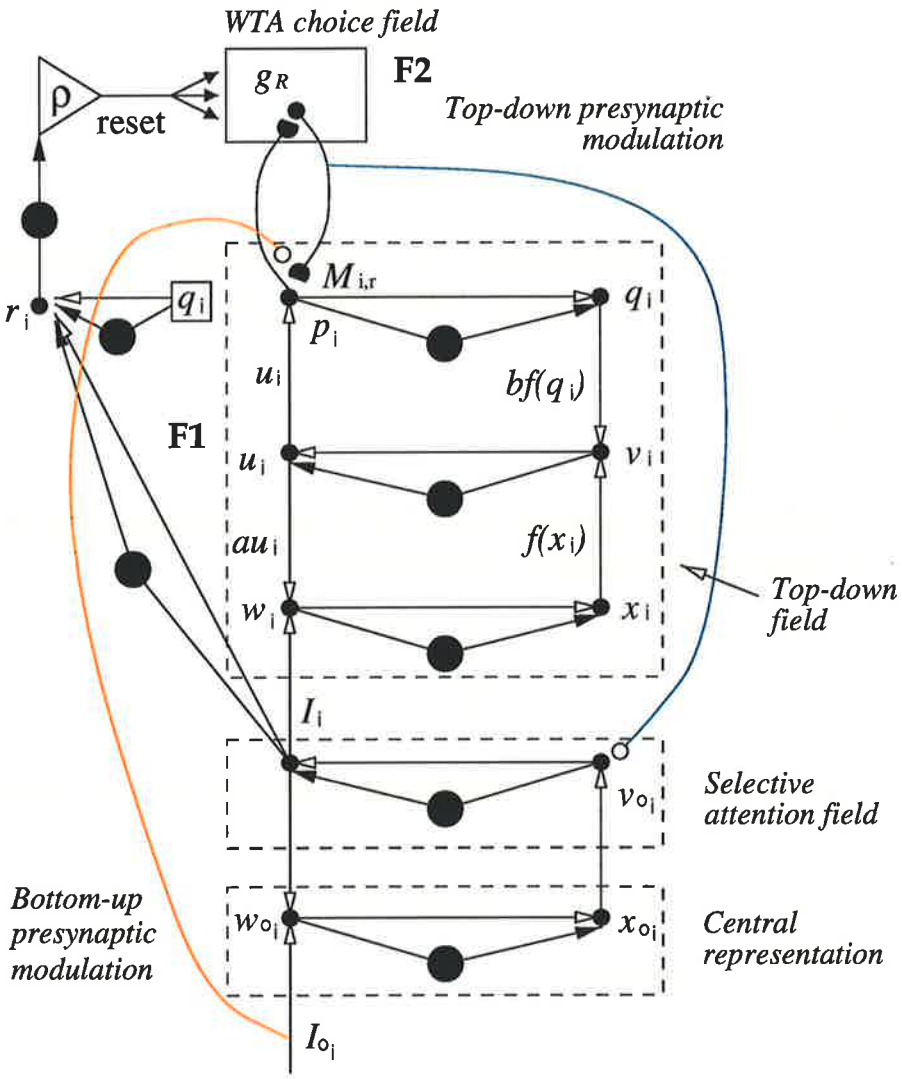


Figure 7.3: Complementary selective attention adaptive resonance theory (CSAART). Large filled circles are gain control nuclei for normalisation, small filled circles are STM nodes, and open circles are gating terminals.

that they all have the form of

$$n_i = \frac{m_i}{\epsilon + \|\mathbf{m}\|} \tag{7.1}$$

where $\|\mathbf{V}\|$ is the L_2-norm of a vector $\mathbf{V}$ and ϵ is a non-zero constant.

The remaining STM activities are

$$\begin{aligned} v_i &= f(x_i) + bf(q_i), \\ w_i &= I_i + au_i, \\ p_i &= u_i + \textstyle\sum_r M_{i,r} g_r \\ w_{o_i} &= I_{o_i} + I_i \\ v_{o_i} &= f(x_{o_i}) \end{aligned} \tag{7.2}$$

where $f(x) = \max(x - \theta, 0)$ is a nonlinear function with threshold θ, a and b are constants, g_r is the rth cell of the WTA choice field, and $M_{i,r}$ is the LTM weight.

WTA Neural Layer

Field $F2$ is a WTA network. In its simplest form, a WTA network is equivalent to a maxima finding operator. A winner y_k out of nodes y_j is expressed as

$$y_k = \begin{cases} 0 & \text{if } y_k < \max\limits_j \{y_j\} \\ 1 & \text{if } y_k = \max\limits_j \{y_j\}. \end{cases} \tag{7.3}$$

More specifically, the winner node R in $F2$ is determined by the maximum rth summed filtered input from $F1 \rightarrow F2$:

$$g_R = \begin{cases} 1 & \text{if } \sum\limits_i p_i M_{i,R} = \max\{\sum\limits_i M_{i,r} p_i : r = 1, \ldots, N\} \\ 0 & \text{otherwise} \end{cases} \tag{7.4}$$

where N is the total number of nodes in $F2$.

A WTA network can be implemented in parallel using the shunting competitive neural layer [31], or other parallel implementations [58, 98].

In general, a shunting competitive layer with cellular activity x_i, fluctuating within the finite interval $[-C_i, B_i]$, stimulated by excitatory and inhibitory inputs I_i and J_i, and nonlinear feedback signals $f_i(x_i)$ and $g_j(x_j)$, can be expressed:

$$\frac{dx_i}{dt} = -A_i x_i + (B_i - x_i)[I_i + f_i(x_i)] - (C_i + x_i)[J_i + \sum_{j=1}^{n} D_{ij} g_j(x_j)] \tag{7.5}$$

where A_i is the passive decay rate, D_{ij} is the gain between nodes i and j, and n is the number of cells in the layer.

Feedforward-Feedback Modulatory Pathways

Under feedforward-feedback modulation, node v_{o_i} is modelled dynamically using a single shunting competitive equation. The equation is a simplification of the full presynaptically modulated shunting competitive neural layer used in Section 4.3.3. The facilitatory terminal at the bottom-up feedforward pathway is modelled using a simple gating mechanism. As a result, the gated synaptic signals are either blocked or passed.

For the top-down presynaptic facilitation pathway, node v_{o_i} is converted to a dynamic cellular activity using an equation similar to (7.5):

$$\frac{dv_{o_i}}{dt} = -Av_{o_i} + (B - v_{o_i})Gx_{o_i}(1 + s_i) - (C + v_{o_i})\frac{1}{N}D\sum_{j\neq i} f(v_{o_j}) \tag{7.6}$$

where A is the passive decay rate, B and C are the saturation limits for the upper and lower bounds respectively, G is a gain factor, s_i is the top-down facilitatory signal, D is the lateral inhibition gain for providing intra-field competition, and N is the number of neurons in a layer.

Node p_i in the top-down field can be expressed as:

$$p_i = u_i + \phi(I_{o_i})\sum_r M_{i,r}g_r \tag{7.7}$$

where $\phi(I_{o_i}) = \max(I_{o_i}, 0)$ is a threshold function.

Weights Adaptation

The adaptive filters between $F1$ and $F2$ are characterised by their long-term-memory (LTM) weights. In ART the weights are adapted during resonance. If the Rth node in $F2$ is activated and the degree of match between the top-down expectation and the bottom-up input satisfies a predefined matching criterion, then the LTM weights ($M_{i,R}$) are updated according to

$$\frac{dM_{i,R}}{dt} = g_R(p_i - M_{i,R}). \tag{7.8}$$

However, under feedforward-feedback modulation learning must be restricted, so that existing object categories are not recoded by their parts. This may be achieved with a very slow learning rate, thus existing LTM weights are not eroded significantly but merely enhanced in the matched object parts.

The Matching Equation

The degree of match between the bottom-up and top-down patterns is determined by the vector r [29], such that each individual element

$$r_i = \frac{q_i + I_i}{\|\mathbf{q}\| + \|\mathbf{I}\| + \epsilon}, \quad \epsilon > 0. \tag{7.9}$$

The matching is considered a success with resonance established, if the degree of match is greater the vigilance parameter ρ by satisfying the constraint

$$\frac{\rho}{\epsilon + \|\mathbf{r}\|} < 1. \tag{7.10}$$

7.2.3 Parts Recognition and Occlusion Simulations

Computer simulations have been performed, four selected simulations of the neural architecture are presented in this section. The simulations are intended to demonstrate the feedforward-feedback interactions for parts recognition in the presence of clutter and occlusion.

The results of the simulations are shown in Figure 7.5. Each column corresponds to a separate simulation. The results are arranged in the same format as in Figure 7.1, showing four neural activity patterns which have been labelled as *activated memory*, *top-down field*, *selective attention field*, and *input*, corresponding to $F2$, $F1$, $F0$, and $Input$ in Figure 7.1. Prior to the four simulated inputs, the objects in Figure 7.4 were learned.

Figure 7.4: Learned object patterns

In Figure 7.5(a) we have an input object partially occluded by white stripes, and the results show that resonance is established after gating out the missing stimuli from the top-down pattern by the feedforward pathway. Part (b) shows recognition can still be achieved if a significant portion of a familiar object is present. The mechanisms achieving this parts recognition are the same as those in Part (a). The input in Part (c) is similar to that in Part (a) except the stripes are black. Although the two inputs are very similar, the mechanisms that enable matching in Part (c) are predominantly feedback in nature. The results show that a mental percept of the occluded object

may be formed by "imagining" what is hidden behind the occluding stripes. In contrast the case in Part (a), where the stripes are clear, such top-down imagination is not possible because the object is not "hidden". Finally in Part (d), we have an incomplete object that is in occlusion, and the results demonstrate the effects of both feedforward and feedback modulations.

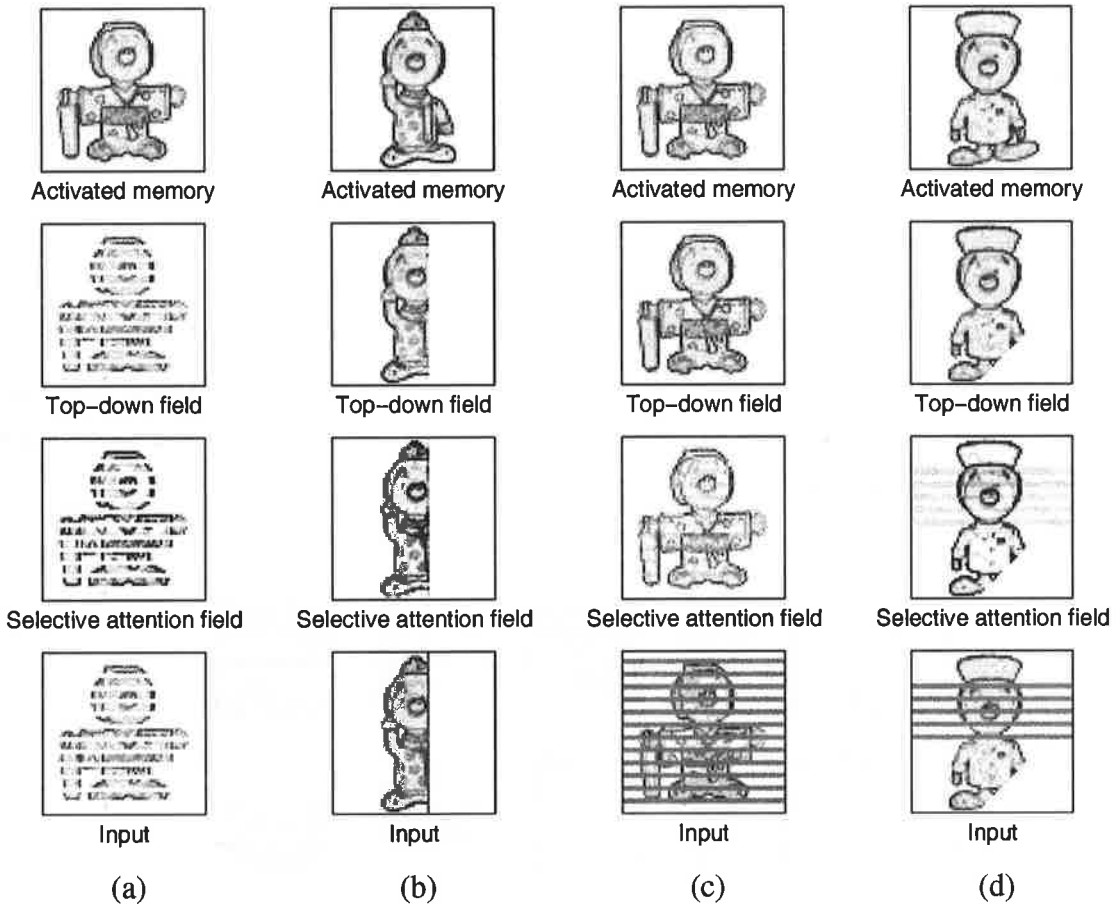


Figure 7.5: Parts recognition and occlusion simulations.

The results show that our model can be used to explain why grouping of fragments caused by occlusion is more effective when the occlusion cues are strong as shown in Figure 7.6(a), in contrast, the recognition of objects fragmented by weak occlusion cues are much more difficult [134]. From our neural architecture point of view, strong occlusion cues allow top-down visual imagination to occur as in Figure 7.5(c), while the case in Figure 7.6(b) is similar to Figure 7.5(a) where object fragments are classified as belonging to a particular learned category but not linked together, thus failing to form a stable mental percept (the top-down field) of the bottom-up pattern.

(a) (b)

Figure 7.6: Perceptual grouping of object fragments. (a) More effective with strong occlusion cues. (b) Weak occlusion cues increase object recognition difficulty. Adapted from [134].

7.2.4 Framework Simulations with CSAART

In this section, four simulations on the neural framework with CSAART incorporated are presented. The simulations are designed to demonstrate the framework's ability to perform partial object recognition in the presence of occlusion and clutter. The simulated scenes feature objects from Columbia Object Image Library (COIL-20) [133]. The set of objects from COIL-20 is shown in Figure 7.7. By using the COIL-20 objects, we show that the system is capable of handling a reasonably large number of objects. These objects can also be used as a benchmark for visual scene analysis.

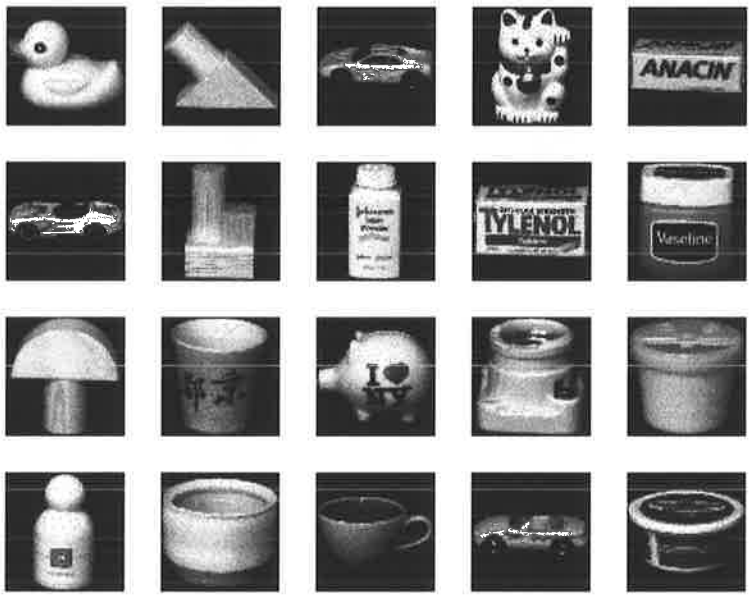


Figure 7.7: The Columbia Object Image Library (COIL-20).

As usual, the COIL-20 objects are learned by the framework using the ART2 learning algorithm. Image scenes featuring these objects are then simulated. The simulation results are shown in Figures 7.8-7.11.

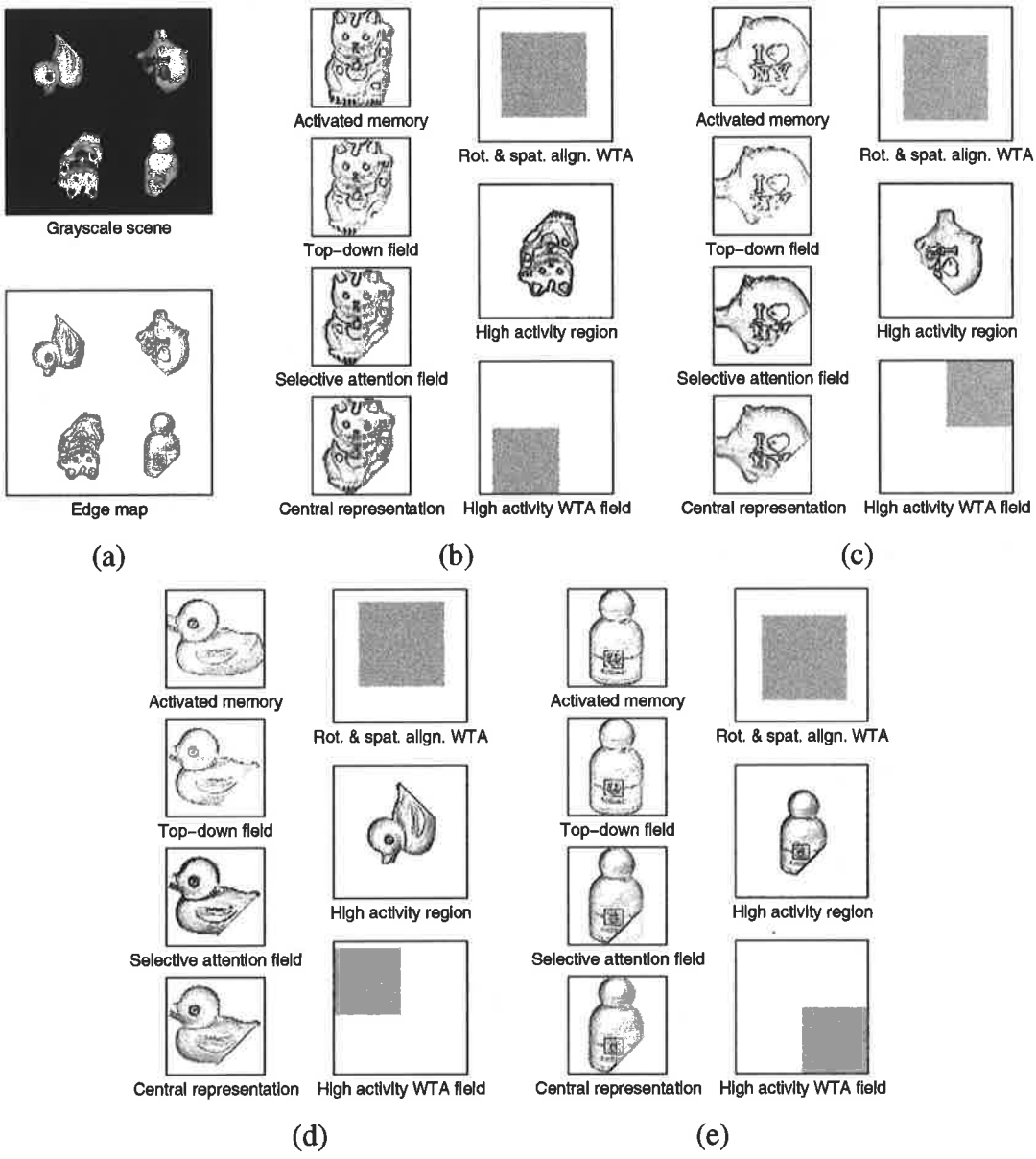


Figure 7.8: CSAART simulation I. (a) Input test scene. Parts (b), (c), (d) and (e) are the simulation results of the recognised objects.

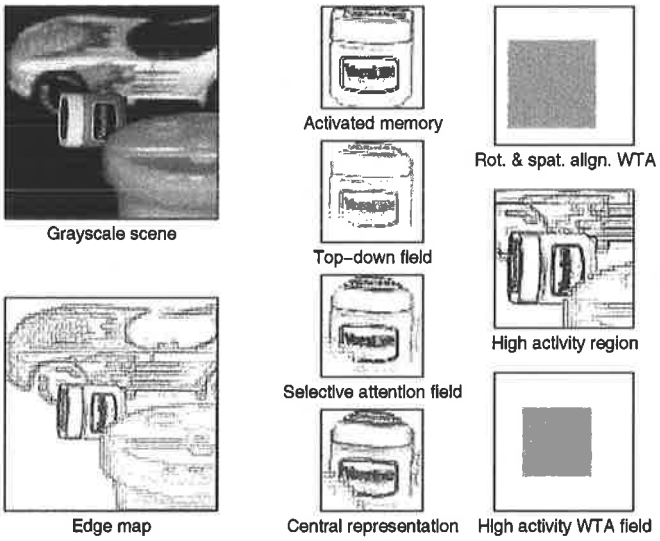


Figure 7.9: CSAART simulation II. (a) Input test scene. Part (b) is the simulation results of the recognised object.

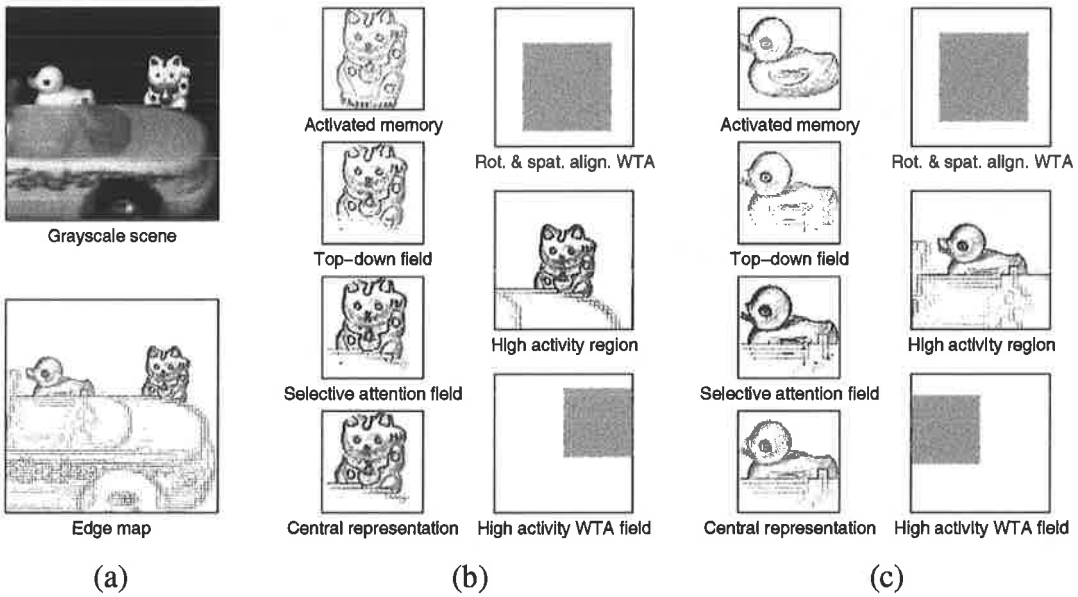


Figure 7.10: CSAART simulation III. (a) Input test scene. Parts (b) and (c) are the simulation results of the recognised objects.

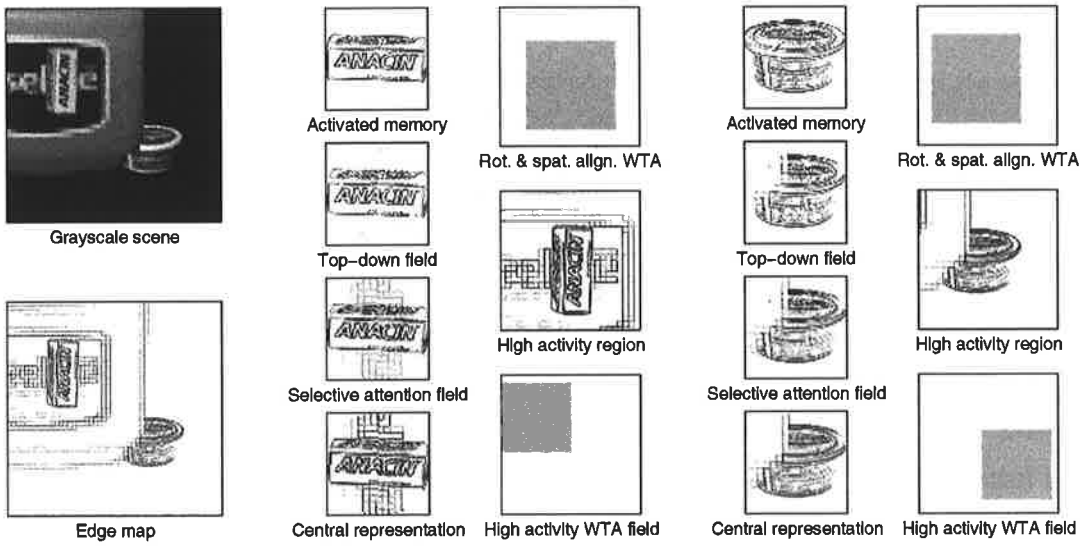


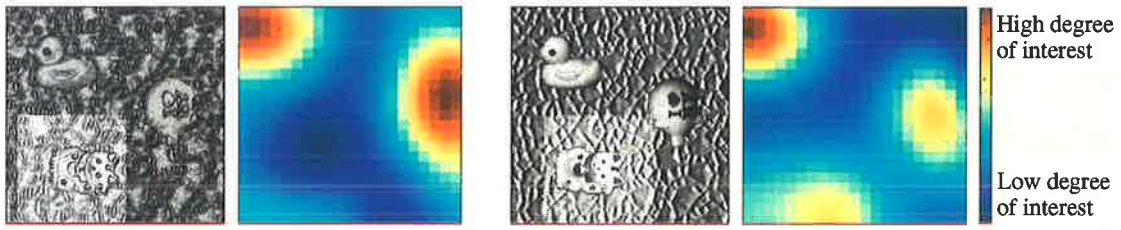
Figure 7.11: CSAART simulation IV. (a) Input test scene. Parts (b) and (c) are the simulation results of the recognised objects.

7.3 Robust Automatic Attentional Capture

Automatic attentional capture was introduced to improve the efficiency of the proposed system. The principle idea of automatic attention is coarse-to-fine sampling, whereby the visual scene is analysed in a coarse manner to locate a region of interest for further detailed analysis. Simulation results in Chapter 5 have shown that implementing automatic attention is more difficult than anticipated. The major problem for robust automatic attention is the presence of strong background clutter, which may offset the window of attention (the high activity field), causing familiar objects to be partially captured. Further investigations have shown that capturing attention based on the region with the maximum amount of edge activity is ineffective in a textured image. The reason is that the background texture generates far more edges than the objects. In such cases, intensity information is more suitable because of its compactness and homogeneous nature. Nonetheless, edges are still an important visual representation for matching, because edges are generally not affected by illumination conditions and surface reflectance. Hence both edge and intensity information may be used to improve the robustness of the framework.

An example of a highly textured image is shown in Figure 7.12. In each part of the figure, there is an input scene in either edge or intensity representation and a *region of interest map* generated from the inputs to the high activity WTA field. Both region of interest maps show three distinct areas, but the one from intensity is more compact and well defined. In Part (a), there is more edge activity in the background than the object regions, selection is based on the region with

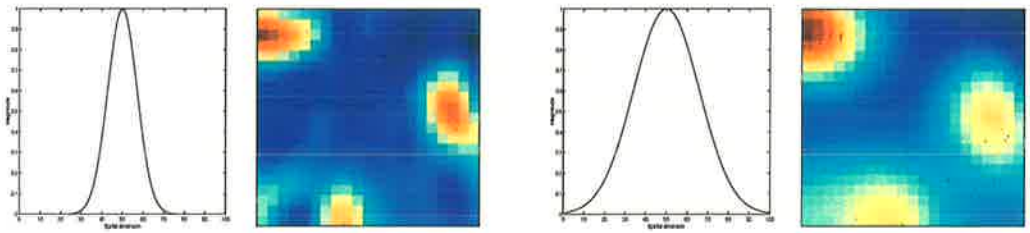
the minimum amount of edge activity.



(a) (b)

Figure 7.12: Attentional capture in a highly textured background image using edge and intensity maps. (a) Edge map and an activity map showing regions of interest. (b) Intensity map and its corresponding regions of interest. The highlighted area is one of the captured regions.

A narrow Gaussian receptive field produces sharper and more compact regions of interest than a wider receptive field as evident in Figure 7.13. It is not always suitable to use a narrow Gaussian receptive field. An appropriately chosen receptive field can provide continuity, so that an object consists of several different intensity regions may be treated as a whole. The highlighted object in Figure 7.12(b) is detected as having two separate regions in Figure 7.13(a), while it is correctly shown as one in Figure 7.13(b).



(a) (b)

Figure 7.13: Effect of Width of Gaussian receptive field on region of interest map. The input scene is the same as in Figure 7.12(b). (a) $\sigma^2 = 100$. (b) $\sigma^2 = 500$.

In order to improve the robustness of automatic attention, several issues must be resolved:

- the choice of automatic selection threshold;
- automatic resizing of the window of attention; and
- recognition of partially captured objects.

7.3.1 Automatic Selection Threshold

The threshold was previously obtained by the worst case scenario, in which the region of interest containing a familiar object with the lowest level of activity is used to set the threshold. This guarantees no region with a familiar object would go undetected. However this approach requires prior knowledge of the worst case scenario for each scene. A more general approach is to use a certain percentage of the peak region of interest activity as a guide for setting the threshold. In most of the simulations, 50% of the maximum region of interest activity was adequate to locate all regions containing familiar objects. We have simulated 30 randomly created test scenes, and in 83% of them we have been able to detect all regions of interest.

7.3.2 Automatic Resizing of Window of Attention

The robustness and efficiency of automatic attention may be improved by allowing automatic resizing of the window of attention. Isolated objects on a clear background may be detected easily, thus a tightly fit window of attention is sufficient. On the other hand, a window of attention that is much larger than the object might be required if the object is closely located to other objects. Such an approach ensures the window of attention is able to capture most of the significant information.

An example of closely located objects is shown in Figure 7.14. The region of interest map indicates that there are two regions of interest: a large region in the top left-hand corner of the scene and a smaller one in the bottom right-hand corner. Their relative sizes may be used to resize the window of attention, hence giving one large and one small window of attention. The large window is expected to capture both objects, from which one of the objects is expected to be recognised. Instead of applying a spatial inhibition as in Section 4.4.5 after the recognition of the first object, an object-based inhibition is used. This reduces the risk of suppressing the yet to be recognised object.

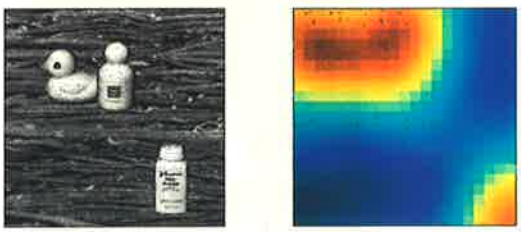


Figure 7.14: Automatic resizing of window of attention. The region of interest map provides information on the selection of the window size.

7.3.3 Partially Captured Objects

With the aid of automatic resizing of the window of attention, an added degree of flexibility and robustness has been introduced to the framework. As a result, partial capturing in the high activity field is not expected to occur as frequently as for a fixed window size. However should it happen, there are two strategies which may be employed to deal with it:

- *Zero padding* - problems arise when a familiar object is located near the border of the window, where it lacks a full set of neighbours for convolution. This is a common technique for edge detection [38]. Although not sophisticated, it is the most direct approach.
- *Partial sampling* - each LTM is restricted to its central region, such that bottom-up memory activation is based on the restricted region only. The assumption is that any partially captured object would contain its central region, and sampling by partial LTMs can determine the position and orientation of the object. Several examples of LTMs reduced to the central regions are shown in Figure 7.15. If the captured region is as shown in Figure 7.16, then convoluting with partial memories can correctly activate the corresponding memory and determine the captured object's position and orientation. The window of attention is duly shifted to the new location, and the rest of the framework operates as normally.

Figure 7.15: Examples of partial long-term-memory.

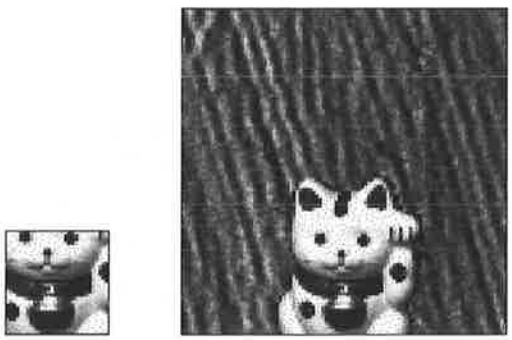


Figure 7.16: Memory activation using partial sampling.

7.4 Size Invariance

The size of the region of interest map has strong implication to the actual size of the captured object. From the size of the region of interest, we can estimate the size of the window of attention, which gives the maximum size of any potentially recognisable object within the region. This has only provided the upper limit to the object size. To cater for a range of other possible sizes we extend the idea of parallel reference frame introduced in Section 4.5 by using multi-resolution LTMs as shown in Figure 7.17. Instead of using fixed-size LTMs for bottom-up memory activation, multiple copies of different resolution of each LTM are used to determine the probable size, location and orientation of the captured object, as well as its potential match in memory. Since all object patterns are normalised, the bottom-up and top-down pair with the most common features will have the largest activity regardless of pattern size. Although only discrete sizes are used for the multi-resolution LTMs, minor size variations are dealt with using band transformation and shape attraction as in Section 4.6. Because there is no hint as to what the lower size limit should be, it must be set manually.

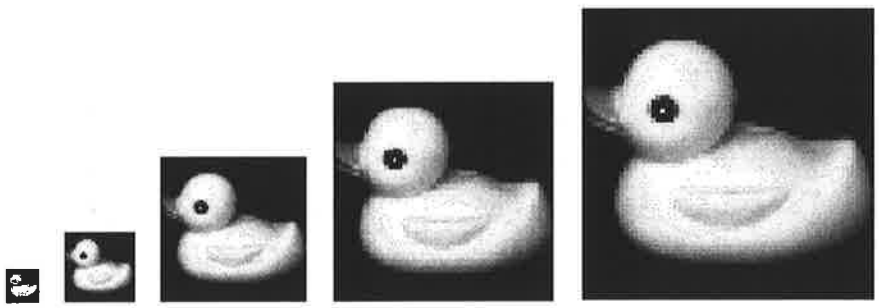
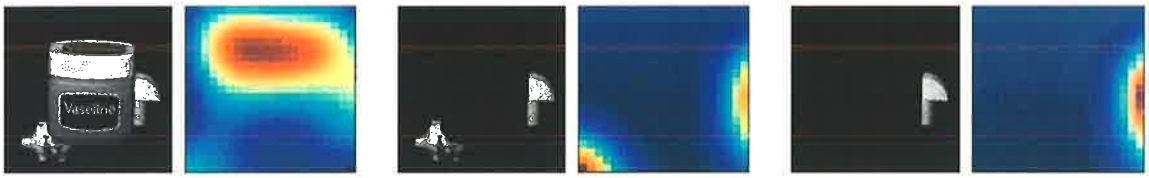


Figure 7.17: Multi-resolution LTMs. The sizes of the depicted objects are 16×16, 32×32, 64×64, 96×96 and 128×128 pixels.

An example of a visual scene featuring different size objects is shown in Figure 7.18. Part (a) shows that a large region of interest is detected from the scene. A window of attention based on the detected region is generated, from which multi-resolution LTM patterns are used to identify the bottom-up object. After recognising the object, it is suppressed from the scene using object-based inhibition. As a result, two smaller regions of interest are detected as in Part (b). With the framework's ability to handle incomplete objects, the two remaining objects are readily recognised in order shown in Parts (b) and (c).



(a) (b) (c)

Figure 7.18: Object size from region of interest map. In each part, there are the input scene and its region of interest map. (a) A large region of interest requires a large window of attention. (b) After object-based inhibition, two regions of interest remain. (c) The last of the three detected regions.

7.5 Conclusion

In this chapter we proposed three advanced features for the framework: i) feedforward-feedback modulatory pathways for parts recognition, ii) robust automatic attention, and iii) size invariance.

In the first part, we demonstrated the importance and usefulness of feedforward-feedback interactions in neural network modelling. In particular, we found that both ART and SAART may be enhanced to recognise incomplete objects by having the complementary feedforward and feedback modulatory pathways. The simulation results illustrated how occluded and incomplete objects are handled explicitly using top-down imagination (forming an expectation for the occluded parts) and implicitly using bottom-up seeing (limiting the mental percept to the visible object parts only). The results also showed why strong occlusion cues allow better perceptual grouping.

The feedforward-feedback interactions enable the top-down and bottom-up patterns to adapt to each other by forming a closed-loop in the neural architecture, thus resonance is allowed to occur.

In the second part, we described several ways in which the robustness of the automatic attention stage could be improved. These improvements were specifically designed to overcome some of the deficiencies experienced in the simulations in Chapter 5.

Lastly, we proposed a method for incorporating size invariance into the framework. Although the framework is capable of dealing with minor size changes as shown in Section 5.9, it lacks the mechanisms to handle arbitrary size change.

Chapter 8

Conclusions and Recommendations

8.1 Recapitulation of the Thesis

This study was motivated by the need to incorporate attentional mechanisms into models of visual functions. Attention is essential to visual perception if one is to remain vigilant to changes in one's immediate surroundings. The ability to detect and locate changes visually is crucial to the smooth operation of our everyday lives. Amongst all sensations, vision is unparalleled in its richness of information content and remoteness from the source. It allows us to examine and understand our environment without having to come into physical contact with the source, thus minimising any potential danger that might exist.

In acknowledging the importance of attention in visual perception, this study set out to investigate the role of attention in performing visual functions, in particular object recognition and visual scene analysis. Specifically, this study aimed to model the computational properties and dynamics of attentional processes in spatial attention and top-down memory-guided attention using feedforward-feedback interactions and neural mechanisms. A neural architecture with attentional mechanisms for the analysis of visual scenes under a variety of visual conditions was the main expected outcome of this study.

A top-down modelling approach consisting of three levels was employed to achieve the objectives of this study. The three levels, formed according to their degree of abstraction, are psychological, neurophysiological and implementational. In the psychological level, a framework for the proposed neural architecture was developed using psychological models and theories of visual perception and attention. The level of neurophysiology enabled us to establish connections and computational relationships among various neural substructures within the framework. Finally, the implementation level provided details on the modelling and construction of those

neural substructures.

In order to lay the psychological and neurophysiological foundation for the proposed system, a review of cognitive findings on visual perception, object recognition, visual attention, and the visual system was provided in Chapter 2. In addition, the chapter addressed computational issues in object recognition such as variability in sensory information and object representation. The main conclusions of this chapter were: attention is a limiting process whereby attended information is allowed through unchanged or amplified, while unattended information is attenuated; spatial attention occurs via the modulation of sensory processes in the visual cortex, and top-down feedback pathways have been found to boost neuronal activities in LGN cells during attention.

Common approaches to the modelling of the high-level visual function of object recognition were considered in Chapter 3, not only to provide a survey of this area, but also to highlight some of the advantages and limitations of the existing approaches. In particular, we focussed on the biological approach using artificial neural networks. It was our intention to equip the proposed vision system with some of the computational benefits of neural networks derived from their massively parallel networking architecture. Of the several neural architectures and their associated learning paradigms discussed, the Adaptive Resonance Theory (ART) neural network architecture was the most suitable choice to form part of our proposed neural architecture. ART is a theory of self-organising network that includes the cognitive concept of attention, vigilance, top-down priming and bidirectional learning in real-time neural systems. An extension of ART, called Selective Attention Adaptive Resonance Theory (SAART), was also considered. SAART is an important class of neural networks that complements ART by overcoming some of the weaknesses of ART. This theory models the computational role of top-down feedback pathways and chemical synapses during selective attention. It suggests that top-down feedback signals may be used to selectively process stimuli from a complex scene, and attentional modulation is achieved by regulating the amount of chemical transmitter flow from synaptic terminals to postsynaptic cells, thereby controlling the net excitation or inhibition available to the postsynaptic cells. Under SAART selective processing was achieved by a process called top-down presynaptic facilitation, where top-down recalled memory was used to selectively facilitate individual synaptic signals by modulating their synaptic gain.

However, ART and SAART alone were insufficient for visual scene analysis as the two classes of neural networks did not address problems such as shifts in position and orientation, attentional capture and shift, distortion, occlusion, and recognising multiple objects from the visual scene. This led us to propose a neural architecture that is based on ART and SAART with computational properties and dynamics of attentional processes for visual scene analysis.

Chapter 4 represents the main body of the thesis, in which a biologically inspired neural frame-

work for distorted and cluttered visual scene analysis was proposed. The proposed model unifies some of the cognitive findings in Chapter 2 with computational models described in Chapter 3 by using the methodology adopted in Chapter 1. Psychologically, the proposed model encompasses a two-stage theory of biological vision, namely the parallel preattentive stage and the serial attentive stage. Architecturally, the model consists of massively parallel feedback and feedforward connections, and is based on a bi-directional structure with both bottom-up and top-down pathways. Bottom-up signals were converted to elementary features and used as visual cues for capturing attention, from which a region of interest was located for further processing with the eventual goal of memory activation. Top-down signals were used in memory guided search and recognition. In particular, memory was used as a feedback to achieve attentional modulation of the bottom-up pathways. Specifically, the model was capable of detecting, locating, localising and recognising any familiar objects from a visual scene, in the presence of occlusion, distortion and background clutter, in an autonomous fashion. Moreover, it formed a translation, rotation and distortion invariant object representation of the recognised object in an object-based reference frame. The framework was presented in a modular fashion with visual functions considered individually. In addition, we considered the effect of proximity and similarity on attentional capture and other psychological processes such as mental rotation.

The proposed neural architecture was implemented and simulated in Chapter 5 using a number of digital images, of both synthetic and real-world scenes. As mentioned in Chapter 5, synthetic images are useful during the development and testing stages as they exclude many practical considerations such as camera angles and distance, illumination conditions, non-uniform object size, and variations in 2D projections. Whereas real-world images are required to prove the effectiveness and robustness of the system in practical applications. Besides verifying the model, the simulations also served to illustrate the concepts behind the model and expose any weaknesses that it might have. The simulation results demonstrated the two modes of attention modelled in Chapter 4, showing a fast, coarse sampling process that operated in parallel over the entire visual scene, and a slower process that analysed a limited spatial region in great details. On the whole, the proposed system was found to be effective in detecting, locating and recognising familiar objects from the test images despite shifts in position and orientation, distortion, and the presence of severe background clutter and occlusion. Furthermore, the objects captured in the real-world images were specifically chosen to have a very similar body shape in order to increase the difficulty of the problem. Because of its ability to deal with distortion, the system is robust against minor changes in size and 2D projections of 3D objects. Discussion on the design of the system parameters was also provided in Chapter 5.

Chapter 6 provided an illustrative example on how the framework model may be extended to incorporate other visual functions. Extendibility was one of the aims of this study, as vision is a vast and diverse field, comprising of many visual functions, the proposed model must be

flexible enough to allow for additional functions. This chapter showed elementary motion as a visual cue for attentional capture in the framework, thus allowing moving objects to be recognised. We proposed a motion detection neural architecture that was modelled according to the motion pathway, hence it respects key neuro-psychological and physiological findings. Significantly, this extension demonstrated the flexibility and extendibility of the framework, and it also provided important insights for modelling motion perception.

Chapter 7 extended the idea of presynaptic modulation to the top-down pathway by incorporating a feedforward modulatory pathway. The feedforward connection could modulate any top-down expectation, so that incomplete objects may be recognised as subsets of learned categories. The resultant feedforward-feedback interactions enabled the top-down and bottom-up patterns to adapt to each other by forming a closed-loop in the neural architecture, thus resonance was allowed to occur. Ways to improve the robustness of the automatic attention stage were proposed. An extension to achieve size invariance was also presented.

8.2 Concluding Statement

This thesis addressed the development of a biologically inspired neural framework with attentional mechanisms for visual scene analysis. The original contribution of the research was outlined in the body of the thesis. The main objectives of the work that have been fulfilled are:

- A detailed study of attentional phenomena in visual perception with a view to understanding the computational role of attention in relation to object recognition and visual focussing.
- A thorough review of the current status of artificial vision systems in the context of object recognition.
- The development of a neural framework with computational properties and dynamics of attentional processes for visual scene analysis using feedforward-feedback interactions and various biological mechanisms.
- The development of a neural architecture for elementary motion detection modelled according to the motion pathway and its incorporation into the framework.
- The investigation, via simulation studies, of the effectiveness of the proposed system under a variety of non-favourable visual conditions.
- The application of the proposed system to practical real-world images under realistic conditions.

- The development of a self-organising neural architecture called *Complementary Selective Attention Adaptive Resonance Theory* (CSAART) network for parts recognition and its incorporation into the framework.

We conclude that attention plays a very important role in many high-level visual functions. By modelling computational properties and dynamics of attentional processes we can enhance artificial vision systems to cope with difficult visual conditions. We show that feedforward-feedback interactions with synaptic modulation are a versatile and powerful mechanism for performing many useful functions such as transformations, filtering, gain control, and selective processing in neural network based vision systems. As demonstrated in the thesis, these interactions have been used in modelling selective neural pattern transfer, top-down memory-guided selective processing, synaptic signal facilitation, attentional capture and shift, shape attraction and band transformation, and directional selectivity.

Although the theory and results presented in the thesis were primarily concerned with vision, the concepts and basic computational mechanisms that were proposed do not place any restrictions on the type of input signals being estimated. The theory can thus be readily applied to other modalities such as audition.

8.3 Recommendations for Future Work

Visual scene analysis is a much more complex process than what has been considered in this thesis. Although in our investigation we restricted ourselves to visual selective attention and object recognition, visual scene analysis also relies heavily on other higher-level cognitive and mental processes such as interpretation, memory, inference, and intelligence. There is a great degree of flexibility in what one can perceive from a complex visual scene, thus it is common for two people to have a very different percept from each other upon encountering a visual scene. Even within our scope of visual attention and object recognition, many improvements can be made to enhance the proposed model. Specifically, we have identified several areas where future work may be carried out.

1. Elementary features - In this thesis we considered only luminance contrast and elementary motion as the features extracted from the visual environment for analysis. There are, however, many other features which we could have used to represent objects such as colour, texture, and orientation. In doing so, we generate a separate feature map for each feature from the input scene. These feature maps are subsequently combined to form the

saliency map from which a region of interest based on all those features may be selected for further processing [98].

2. Attentional processes - With the use of multiple features to represent objects we can implement the so-called feature search mode (instead of the singleton detection mode implemented in Section 4.4) in which attention is directed to locations that match some desired visual feature [54]. It is clear that singleton detection is achieved in a bottom-up manner, while feature search is a top-down process where the synaptic signals representing the desired feature are biased in premeditation. From an implementation point of view a presynaptic facilitatory signal may be used to facilitate signals from the corresponding feature map, so that the desired feature may be selected from the winner-take-all competition in the saliency map. Another form of top-down expectation that can be implemented is during a visual search in which the target of the search is known by its shape alone. Therefore we can apply a "body shape" receptive filter instead of a simple Gaussian receptive field to measure the bottom-up activation from the neural input pattern. For example, the real-world objects captured for simulations in Section 5.9 all shared a similar body shape, if we were to search for all those objects, it would be a lot more efficient if their body shape receptive field was applied during attentional capture. It should be noted that both the preattentive and attentive modes interact with top-down and bottom-up processes, so there are many attentional processes that may be implemented.

3. Visual functions - As mentioned in the limitations of the proposed model in Section 5.10 additional visual functions are required to deal with 3D objects and changes in size, as well as a number of other visual conditions. For 3D object recognition, we may employ compressed object representations as in appearance-based object recognition in Section 3.1.2, via principal components analysis. It has been successfully employed in various neural architectures [80].

4. Applications - Since the proposed theory is not restricted to the visual domain, we can reformulate our model to suit the audio domain, so that the neural architecture could detect and recognise audio patterns such as words or sentences from a noisy environment, or model the well known cocktail party effect.

Appendix A

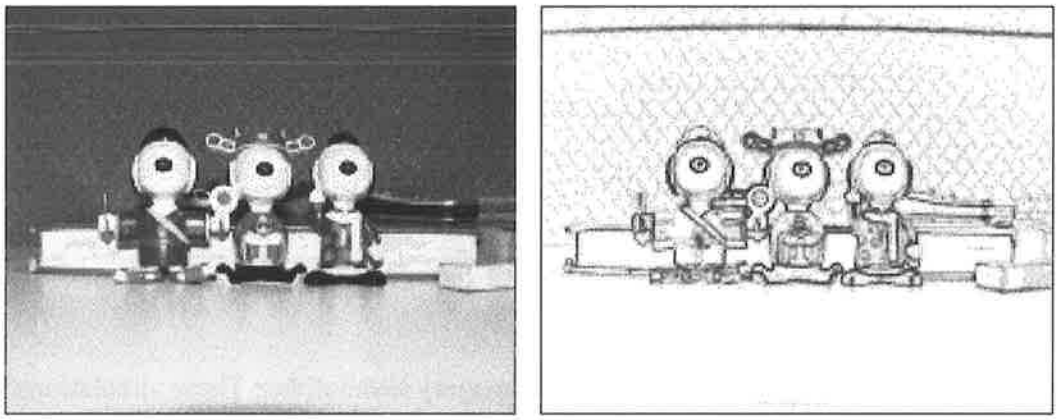
Additional Simulations

This appendix provides four additional real-world imagery simulations. These simulations were performed prior to the incorporation of the feedforward modulatory pathway proposed in Chapter 7, therefore minor occlusions were dealt with using top-down presynaptic facilitation alone. Satisfactory results were obtained in each of the four simulations; of the ten recognisable objects in the input scenes, nine were correctly detected and recognised. The one that failed to be recognised is of different size to its learned projection.

The four simulations are briefly described below:

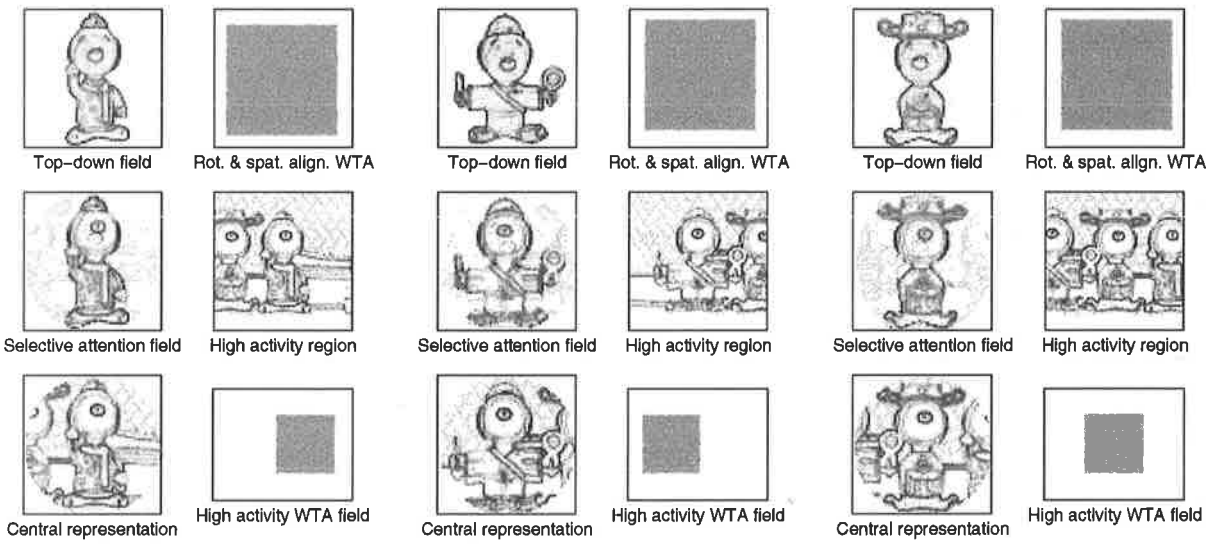
- Simulation I, shown in Figures A.1 and A.2, features three recognisable objects standing side-by-side in a cluttered scene.
- Simulation II, shown in Figures A.3 and A.4, also features three recognisable objects but they are closely positioned such that one of the objects occludes the the other two.
- Simulation III, shown in Figures A.5 and A.6, contains one recognisable object that is placed behind a toy figurine.
- Simulation IV, shown in Figures A.7 and A.8, has three recognisable objects. This simulation was specifically designed to show the system's ability to distinguish and recognise two highly similar objects in occlusion (similar in shape and both have outstretched arms).

Simulation I



(a) (b)

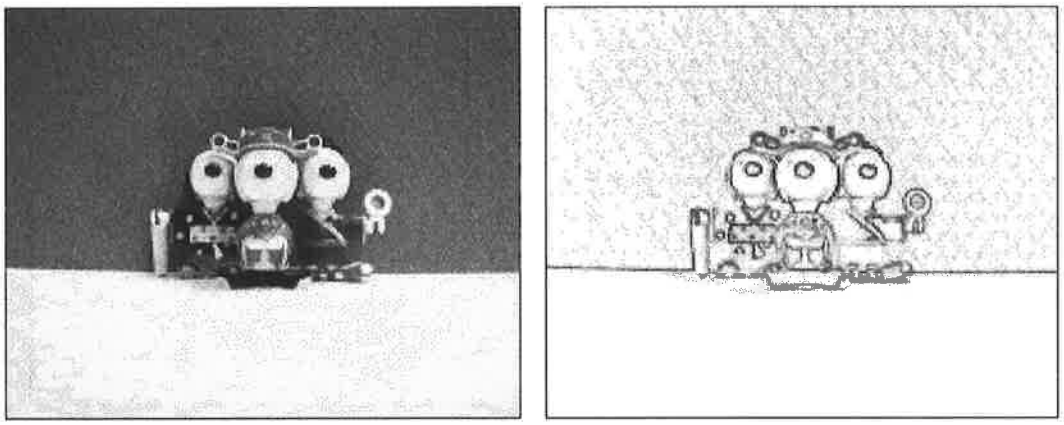
Figure A.1: Additional real-world imagery: Simulation I input scene. (a) Intensity map. (b) Edge map.



(a) (b) (c)

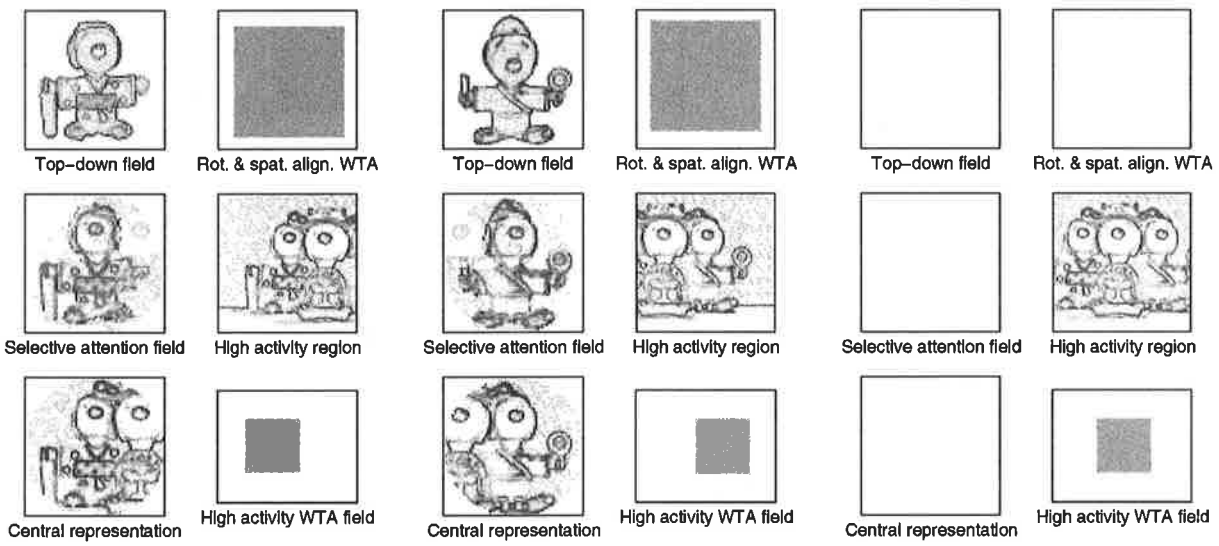
Figure A.2: Additional real-world imagery: Simulation I - Parts 1, 2 and 3.

Simulation II



(a) (b)

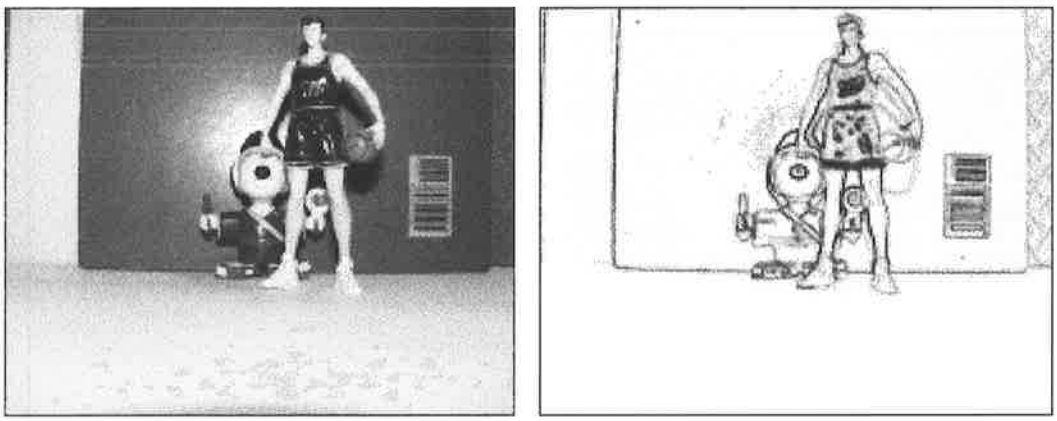
Figure A.3: Additional real-world imagery: Simulation II input scene. (a) Intensity map. (b) Edge map.



(a) (b) (c)

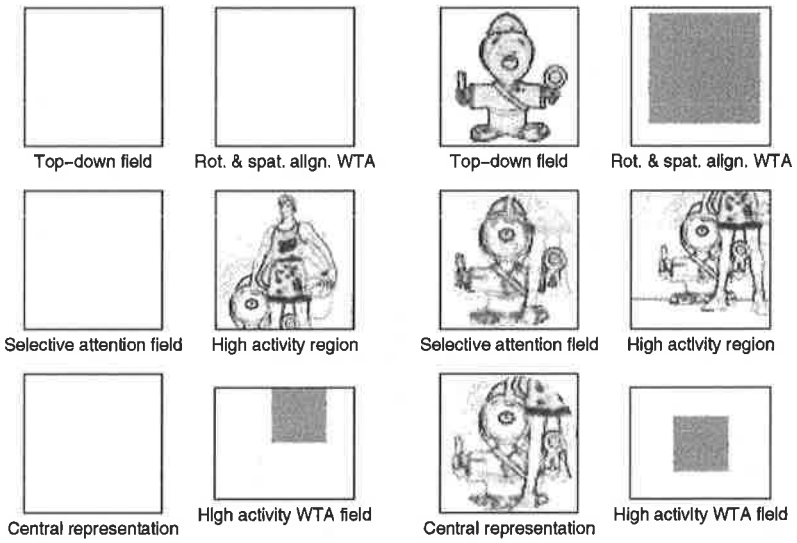
Figure A.4: Additional real-world imagery: Simulation II - Parts 1, 2 and 3.

Simulation III



(a) (b)

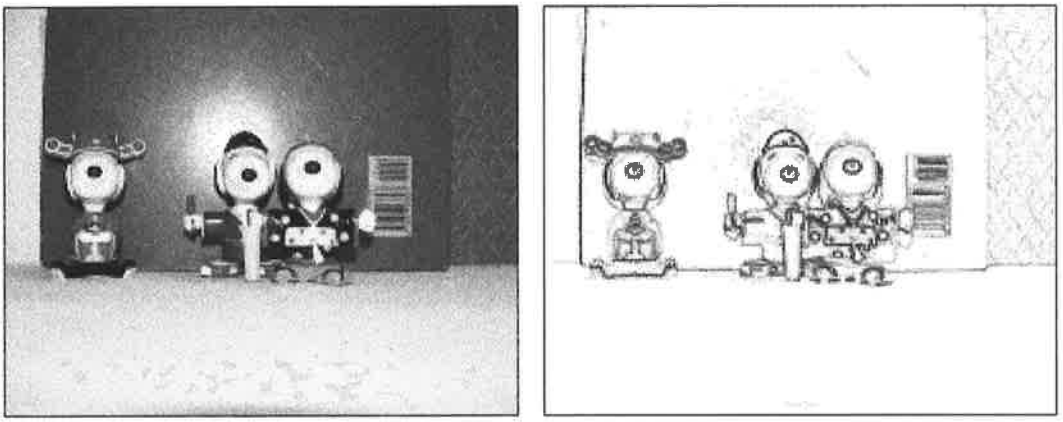
Figure A.5: Additional real-world imagery: Simulation III input scene. (a) Intensity map. (b) Edge map.



(a) (b)

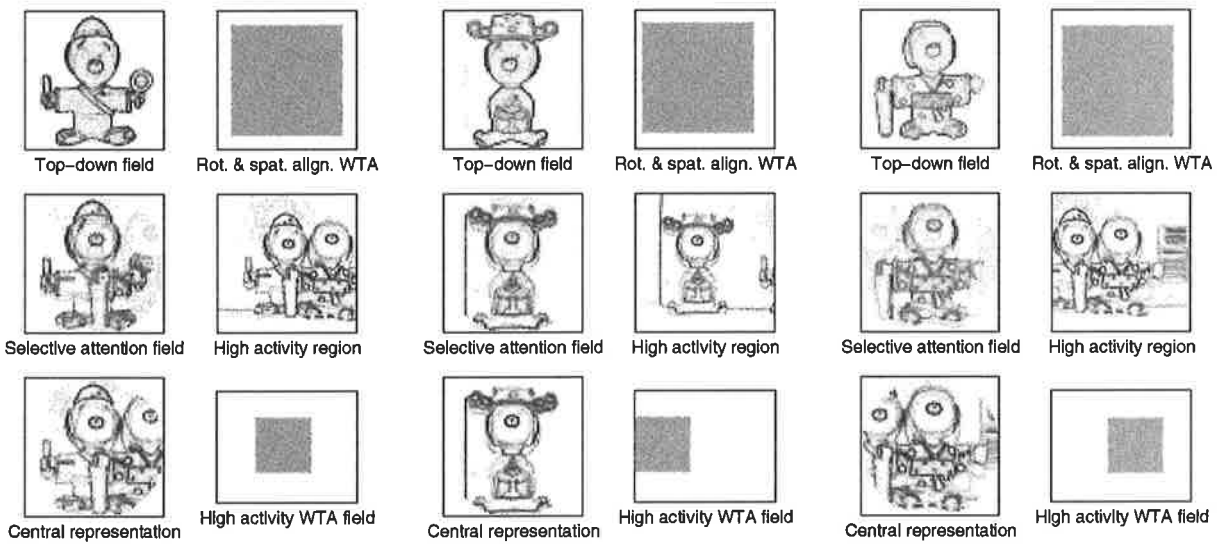
Figure A.6: Additional real-world imagery: Simulation III - Parts 1 and 2.

Simulation IV



(a) (b)

Figure A.7: Additional real-world imagery: Simulation IV input scene. (a) Intensity map. (b) Edge map.



(a) (b) (c)

Figure A.8: Additional real-world imagery: Simulation IV - Parts 1, 2 and 3.

Bibliography

[1] F.W. Adams, H.T. Nguyen, R. Raghavan, and J. Slawny. A parallel network for visual cognition. *IEEE Transactions on Neural Networks*, 3(6):906–922, 1992.

[2] A. Allport. Visual attention. In M.I. Posner, editor, *Foundations of Cognitive Science*, pages 631–682. MIT Press, 1989.

[3] A. Allport. Attention and control: Have we been asking the wrong questions? A critical review of twenty-five years. In D.E. Meyer and S. Kornblum, editors, *Attention and Performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience*, pages 183–218. Cambridge, MA: MIT Press, 1993.

[4] J. Andrade-Cetto and A.C. Kak. Object recognition. In J.G. Webster, editor, *Wiley encyclopedia of electrical engineering*, pages 449–470. John Wiley & Sons, 2000.

[5] A.M. Annaswamy and Ssu-Hsin Yu. θ - Adaptive neural networks: A new approach to parameter estimation. *IEEE Transactions on Neural Networks*, 7(4), 1996.

[6] M.A. Arbib, editor. *The handbook of brain theory and neural networks*. The MIT Press, 1995.

[7] W.F. Bacon and H.E. Egeth. Overriding stimulus-driven attentional capture. *Perception & Psychophysics*, 55:485–496, 1994.

[8] D.H. Ballard and C.M. Brown. *Computer Vision*. Englewood Cliffs, N.J: Prentice-Hall, 1982.

[9] E. Barnard and D. Casasent. Shift invariance and the neocognitron. *Neural Networks*, 3:403–410, 1990.

[10] E. Barnard and D. Casasent. Invariance and neural nets. *IEEE Transactions on Neural Networks*, 2(5):498–508, 1991.

[11] S. Barro, M. Fernández-Delgado, J.A. Vila-Sobrino, C.V. Regueiro, and E. Sánchez. Classifying multichannel ECG patterns with an adaptive neural network. *IEEE Engineering in Medicine and Biology Magazine*, 17:45–55, 1998.

[12] G. Bartfai and S. Grossberg. Fast learning VIEWNET architectures for recognizing 3-D objects from multiple 2-D views. *Neural Networks*, 8:1053–1080, 1995.

[13] R. Basri. Viewer-centered representations in object recognition: a computational approach. In C.H. Chen, L.F. Pau, and P.S.P. Wang, editors, *Handbook of Pattern Recognition and Computer Vision*, pages 863–882. World Scientific Publishing Company, Singapore, 1993.

[14] P.J. Besl and R.C. Jain. Three-dimensional object recognition. *ACM Computing Survey*, 17:75–144, 1985.

[15] C.E. Bethell-Fox and R.N. Shepard. Mental rotation: effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, 14(1):12–23, 1988.

[16] I. Biederman. Recognition by components: A theory of human image understanding. *Psychological Review*, 94(115-147), 1987.

[17] I. Biederman and E.E. Cooper. Evidence for complete translational and reflectional invariance in visual object priming. *Perception*, 20:585–593, 1991.

[18] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[19] P.O. Bishop. Processing of visual information with retinostriate system. In I. Darian-Smith, editor, *Handbook of Physiology*. American Physiological Society, 1984.

[20] D.E. Broadbent. *Perception and Communication*. London: Pergamon, 1958.

[21] D.E. Broadbent. Stimulus set and response set: two kinds of selective attention. In D.I. Motofsky, editor, *Attention: Contemporary Theory and Analysis*, pages 51–60. New York: Appleton-Century-Crofts, 1970.

[22] D.E. Broadbent. *Decision and stress*. London: Academic Press, 1971.

[23] V. Bruce, P.R. Green, and M.A. Georgeson. *Visual perception: physiology, psychology, and ecology*. Psychology Press, 3rd edition, 1996.

[24] D.C. Burr and J. Ross. Contrast sensitivity at high velocities. *Vision Research*, 22:479–484, 1982.

[25] D.J. Burr. Elastic matching of line drawings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3:708–713, 1981.

[26] F.W. Campbell and L. Maffei. The influence of spatial frequency and contrast on the perception of moving patterns. *Vision Research*, 21:713–721, 1981.

[27] G.A. Carpenter. Neural network models for pattern recognition and associative memory. *Neural Networks*, 2:243–257, 1989.

[28] G.A. Carpenter and S. Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, 37:54–115, 1987.

[29] G.A. Carpenter and S. Grossberg. ART 2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, 26:4919–4930, 1987.

[30] G.A. Carpenter and S. Grossberg. ART 3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. *Neural Networks*, 3:129–152, 1990.

[31] G.A. Carpenter and S. Grossberg. *Pattern Recognition by Self-Organizing Neural Networks*. MIT Press, 1991.

[32] G.A. Carpenter, S. Grossberg, N. Markuzon, J. Reynolds, and D.B. Rosen. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*, 3(5), 1992.

[33] G.A. Carpenter, S. Grossberg, and J.H. Reynolds. ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-oganizing neural network. *Neural Networks*, 4:565–588, 1991.

[34] G.A. Carpenter, S. Grossberg, and J.H. Reynolds. A fuzzy ARTMAP nonparametric probability estimator for nonstationary pattern recognition problems. *IEEE Transactions on Neural Networks*, 6(6), 1995.

[35] G.A. Carpenter, S. Grossberg, and D.B. Rosen. ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition. *Neural Networks*, 4:493–504, 1991.

[36] G.A. Carpenter, S. Grossberg, and D.B. Rosen. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, 4:759–771, 1991.

[37] D.P. Casasent and L.M. Neiberg. Classifier and shift-invariant automatic target recognition neural networks. *Neural Networks*, 8(7/8):1117–1129, 1995.

[38] K.R. Castleman. *Digital image processing*. Englewood Cliffs, N.J. : Prentice Hall, 1996.

[39] C.H. Chen, L.F. Pau, and P.S.P. Wang. *Handbook of Pattern Recognition & Computer Vision*. World Scientific, 1993.

[40] R.T. Chin and C.R. Dyer. Model-based recognition in robot vision. *ACM Computing Survey*, 18:67–108, 1986.

[41] E. Chong and C.C. Lim. Elementary motion detection with selective attention. In *Proceedings of the Third International Conference on Knowledge-Based Intelligent Information Engineering Systems, KES'99*, pages 365–368. IEEE Press, 1999.

[42] E. Chong, C.C. Lim, N. Atsikbasis, and P. Lozo. Design of a 2-D neural motion detection filter. In *IEEE Region 10 Annual Conference*, pages 667–670, Brisbane, 1997.

[43] E. Chong, C.C. Lim, and P. Lozo. Modelling of a neural motion detection filter for attentional modulation. In *International Workshop on Image Analysis and Information Fusion*, pages 311–321, Adelaide, 1997.

[44] E. Chong, C.C. Lim, and P. Lozo. Neural model of visual selective attention for automatic translation invariant object recognition in cluttered images. In *Proceedings of the Third International Conference on Knowledge-Based Intelligent Information Engineering Systems, KES'99*, pages 373–376. IEEE Press, 1999.

[45] E.W. Chong and C.C. Lim. Neural model for distorted and cluttered scene analysis. *Submitted to Image and Vision Computing*, 2000.

[46] E.W. Chong and C.C. Lim. A self-organising neural architecture for parts recognition in occlusion. *Submitted to Neural Computation*, 2000.

[47] S.R. Chu, R. Shoureshi, and M. Tenorio. Neural networks for system identification. *IEEE Control Systems Magazine*, 10(3):31–35, 1990.

[48] J.G. Daugman. Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7):1169–1179, 1988.

[49] A. Delopoulos, A. Tirakis, and S. Kollias. Invariant image classification using triple-correlation-based neural networks. *IEEE Transactions on Neural Networks*, 5(3):392–407, 1994.

[50] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18:193–222, 1995.

[51] R. Desimone, E.K. Miller, and L. Chelazzi. The interaction of neural systems for attention and memory. In C. Koch and J.L. Davis, editors, *Large-Scale Neuronal Theories of the Brain*, pages 75–91. A Bradford Book, 1994.

[52] M. Dill and M. Fahle. Limited translation invariance of human visual pattern recognition. *Perception and Psychophysics*, 60(1):65–81, 1998.

[53] R. Dubner and S. Zeki. Response properties and receptive fields of cells in an anatomically defined region of superior temporal sulcus in the monkey. *Brain Research*, 35, 1971.

[54] H.E. Egeth and S. Yantis. Visual attention: control, representation, and time course. *Annual Review of Psychology*, 48:269–297, 1997.

[55] C. Enroth-Cugell and J.G. Robson. The contrast sensitivity of retinal ganglion cells of the cat. *Journal of Physiology*, 1966.

[56] C.W. Eriksen and T.D. Murphy. Movement of attentional focus across the visual field: a critical look at the evidence. *Perception & Psychophysics*, 42:299–305, 1987.

[57] L. Fausett. *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*. Prentice Hall, 1994.

[58] J.A. Feldman and D.H. Ballard. Connectionist models and their properties. *Cognitive Science*, 6:205–254, 1982.

[59] M. Fukumi, S. Omatu, and Y. Nishikawa. Rotation-invariant neural pattern recognition system estimating a rotation angle. *IEEE Transactions on Neural Networks*, 8(3):568–581, 1997.

[60] M. Fukumi, S. Omatu, F. Takeda, and T. Kosaka. Rotation-invariant neural pattern recognition system with application to coin recognition. *IEEE Transactions on Neural Networks*, 3(2):272–279, 1992.

[61] K. Fukunaga. *Introduction to statistical pattern recognition.* Academic Press, London, 1990.

[62] K. Fukushima. Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics*, pages 121–136, 1975.

[63] K. Fukushima. Neural network model for selective attention in visual pattern recognition and associative recall. *Applied Optics*, 26(23):4985–4992, 1987.

[64] K. Fukushima. Neocognitron: A hierarchial neural network capable of visual pattern recognition. *Neural Networks*, 1:119–130, 1988.

[65] K. Fukushima. A neural network for visual pattern recognition. *Computer*, 21(3):65–75, 1988.

[66] K. Fukushima and S. Miyake. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition*, 15(6):455–469, 1982.

[67] K.-I. Funahashi. On the approximation realization of continuous mappings by neural networks. *Neural Networks*, 2:183–192, 1989.

[68] M.S. Gazzaniga, R.B. Ivry, and G.R. Mangun. *Cognitive Neuroscience: the Biology of the Mind*. W. W. Norton & Company, Inc., 1998.

[69] D.E. Glover. An optical Fourier/electronic neurocomputer automated inspection system. In *Proceedings of the IEEE International Conference on Neural Networks*, pages I–569–I–576, 1988.

[70] S. Grossberg. Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, 52:217–257, 1973.

[71] S. Grossberg. Adaptive pattern classification and universal recoding, I: parallel development and coding of neural feature detectors. *Biological Cybernetics*, 23:121–134, 1976.

[72] S. Grossberg. Adaptive pattern classification and universal recoding, II: feedback, expectation, olfaction, and illustions. *Biological Cybernetics*, 23:187–202, 1976.

[73] S. Grossberg. Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks*, 1:17–61, 1988.

[74] S. Grossberg and G. Bradski. VIEWNET architectures for invariant 3-D object learning and recognition from mulitple 2-D views. In B. Bouchon-meunier, R.R. Yager, and L.A. Zadeh, editors, *Fuzzy logic and soft computing*. Singapore: World Scientific Publishing, 1995.

[75] S. Grossberg, E. Mingolla, and D. Todorovic. A neural network architecture for preattentive vision. In M.M. Gupta and G.K. Knopf, editors, *Neuro-vision systems*. IEEE Press, 1994.

[76] S. Grossberg and M.E. Rudd. A neural architecture for vision motion percetion: Group and element apparent motion. *Neural Networks*, 2:421–450, 1989.

[77] S. Grossberg and L. Wyse. Invariant recognition of cluttered scenes by a self-organizing art architecture: figure-ground separation. In *IJCNN-91-Seattle International Joint Conference on Neural Networks*, pages 633–638, 1991.

[78] M.M. Gupta and G.K. Knopf. A multitask visual information processor with a biologically motivated design. *Journal of Visual Communication and Image Representation*, 3(3):230–246, 1992.

[79] S. Haykin. *Neural networks: a comprehensive foundation*. Macmillan Publishing Company, 1994.

[80] J. Hertz, A. Krogh, and R.G. Palmer. *Introduction to the theory of neural computation*. Addison-Wesley, 1991.

[81] G.E. Hinton and L.M. Parsons. Frames of reference and mental imagery. In A. Baddeley and J. Long, editors, *Attention and performance: IX*, pages 261–277. Hillsdale, NJ: Erlbaum, 1981.

[82] J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the USA*, 79:2554–2558, 1982.

[83] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4:251–257, 1991.

[84] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[85] K. Hornik, M. Stinchcombe, and H. White. Universal approximation of an unknown mapping and its derivatives using multilayer Feedforward Networks. *Neural Networks*, 3:551–560, 1990.

[86] C. Huang, O. Camps, and T. Kanungo. Object recognition using appearance-based parts and relations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 878–884, 1997.

[87] D.H. Hubel and T. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195, 1968.

[88] D.H. Hubel and T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160, 1962.

[89] H.-L. Hung, H.-Y.M. Liao, S.-J. Lin, W.-C. Lin, and K.-C. Fan. Cascade fuzzy ART: a new extensible database for model-based object recognition. *Proceedings of the SPIE - The International Society for Optical Engineering*, 2727:187–198, 1996.

[90] M. Ito, H. Tamura, I. Fujita, and K. Tanaka. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol.*, 73:218–226, 1995.

[91] A.K. Jain, J. Mao, and K.M. Mohiuddin. Artificial neural networks: a tutorial. *Computer*, 29(3):31–44, March 1996.

[92] A.K. Jain, Y. Zhong, and S. Lakshmanan. Object matching using deformable templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(3):267–277, 1996.

[93] S. Jamieson and J.F. Boyce. A fuzzy ARTMAP control network for optimisation of an automatic target recognition system. Technical report, King's College London, 1995.

[94] W.A. Johnston and V.J. Dark. Selective attention. *Annual Review of Psychology*, 37:43–75, 1986.

[95] P. Jolicoeur. Orientation congruency effects on the identification of disoriented shapes. *Journal of Experimental Psychology: Human Perception and Performance*, 16(2):351–364, 1990.

[96] P. Jolicoeur and P. Cavanagh. Mental rotation, physical rotation, and surface media. *Journal of Experimental Psychology*, 18(2):371–384, 1992.

[97] B. Julesz and J.R. Bergen. Textons, the fundamental elements in preattentive vision and perception of textures. *Bell System Technical Journal*, 62:1619–1645, 1983.

[98] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[99] J.J. Koenderink and A.J. van Doorn. The internal representation of solid shape with respect to vision. *Biolog. Cybern.*, 32:211–216, 1979.

[100] T. Kohonen. The self-organizing map. *Proc. IEEE*, 78(9), 1990.

[101] V. Kreinovich. Arbitrary nonlinearity is sufficient to represent all fuctions by neural networks: A theorem. *Neural Networks*, 4:381–383, 1991.

[102] S.S. Kumar and A. Guez. ART based adaptive pole placement for neurocontrollers. *Neural Networks*, 4:319–335, 1991.

[103] H.W. Kwak, D. Dagenbach, and H. Egeth. Further evidence for a time-independent shift of the focus of attention. *Perception & Psychophysics*, 49:473–480, 1991.

[104] M. Lades, J.C. Vorbrüggen, and J. Buhmann. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.

[105] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. In M.A. Arbib, editor, *The handbook of brain theory and neural networks*, pages 255–258. MIT Press, 1995.

[106] W.C. Lin, F.Y. Liao, and C.K. Tsao T. Lingutla. A hierarchical multiple-view approach to three-dimensional object recognition. *IEEE Transactions on Neural Networks*, 2(1):84–92, 1991.

[107] R.P. Lippmann. An introduction to computing with neural nets. *IEEE ASSP Magazine*, 4(2):4–22, 1987.

[108] L. Liu and S. Sclaroff. Deformable shape detection and description via model-based region grouping. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 21–27, 1999.

[109] N.K. Logothetis and J. Pauls. Psychophysical and physiological evidence for viewer-centered object representation in the primate. *Cerebral Cortex*, 3:270–288, 1995.

[110] S. Loncaric. A survey of shape analysis techniques. *Pattern Recognition*, 31(8):983–1001, 1998.

[111] D.R. Lovell, T. Downs, and A.C. Tsoi. An evaluation of the neocognitron. *IEEE Transactions on Neural Networks*, 8:1090–1105, 1997.

[112] P. Lozo. Neural circuit for self-regulated attentional learning in Selective Attention Adaptive Resonance Theory (SAART) neural networks. In *Proceedings of the Fourth International Symposium on Signal Processing and its Applications, ISSPA'96*, 1996.

[113] P. Lozo. *Neural Theory and Model of Selective Visual Attention and 2D Shape Recognition in Visual Clutter*. PhD thesis, The University of Adelaide, 1997.

[114] P. Lozo and C.C. Lim. Neural circuit for object recognition in complex and cluttered visual images. In *Proc. 1996 Australian and New Zealand Conference on Information Systems*, pages 254–257, 1996.

[115] P. Lozo, C.C. Lim, and D. Nandagopal. Translation invariant pattern recognition: A real-time neural network architecture based on biological visual spatial attention. *Australian Journal of Intelligent Information Processing Systems*, 1995.

[116] P. Lozo and Nanda Nandagopal. Selective transfer of spatial patterns by presynaptic facilitation in a shunting competitive neural layer. In *Proc. 1996 Australian and New Zealand Conference on Information Systems*, pages 178–181, 1996.

[117] Z.L. Lu and G. Sperling. Attention-generated apparent motion. *Nature*, 1995.

[118] S.J. Luck, S. Fan, and S.A. Hillyard. Attention-related modulation of sensory-evoked brain activity in a visual search task. *Journal of Cognitive Neuroscience*, 5:188–195, 1993.

[119] S.G. Mallat. Multifrequency channel decompositions of images and wavelet models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(12):2091–2110, 1989.

[120] G.R. Mangun, S. Hillyard, and S. Luck. Electrocortical substrates of visual selective attention. In D.E. Meyer and S. Kornblum, editors, *Attention and Performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience*, pages 219–243. Cambridge, MA: MIT Press, 1993.

[121] G.R. Mangun and S.A. Hillyard. Modulations of sensory-evoked brain potentials indicate changes in perceptual processing during visual-spatial priming. *Journal of Experimental Psychology. Human Perception and Performance*, 17:1057–1074, 1991.

[122] D. Marr. *Vision.* W. H. Freeman and Company, 1982.

[123] J.H.R. Maunsell and V.P. Ferrera. Attentional mechanisms in visual cortex. In M.S. Gazzaniga, editor, *The Cognitive Neurosciences*, pages 451–461. MIT Press, 1994.

[124] J.H.R. Maunsell and W.T. Newsome. Visual processing in monkey extrastriate cortex. *Annual Review of Neuroscience*, 10:363–401, 1987.

[125] P. Mehra and B.W. Wah. *Artificial Neural Networks : Concepts and Theory.* IEEE Computer Society Press, 1992.

[126] J.M. Mendel and R.W. McLaren. Reinforcement-learning control and pattern recognition systems. In *Adaptive, Learning, and Pattern Recognition Systems: Theory and Applications*, pages 287–318. Academic Press, New York, 1970.

[127] M.M. Menon and K.G. Heinemann. Classification of patterns using a self-organising neural network. *Neural Networks*, 1:201–215, 1988.

[128] M.M. Mesulam. Attention, confusional states, and neglect. In *Principles of behavioral neurology*. Philadelphia: F.A. Davis, 1985.

[129] M. Mignard and J.G. Malpeli. Paths of information flow through visual cortex. *Science*, 251:1249–1251, 1991.

[130] A. Moore, J. Allman, and R.M. Goodman. A real-time neural system for color constancy. *IEEE Transactions on Neural Networks*, 2(2):237–247, 1991.

[131] J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:782–784, 1985.

[132] J.A. Movshon, E.H. Adelson, M.S. Gizzi, and W.T. Newsome. The analysis of moving visual patterns. In C. Chagas, R. Gattass, and C. Gross, editors, *Pattern Recognition Mechanisms*. Vatican Press, Rome, 1985.

[133] H. Murase and S.K. Nayar. Visual learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, 1995.

[134] K. Nakayama, S. Shimojo, and G.H. Silverman. Stereoscopic depth: its relation to image segmentation, grouping and the recognition of occluded objects. *Perception*, 18:55–68, 1989.

[135] K.S. Narendra and K. Parthasarathy. Identification and control of dynamical systems using neural networks. *IEEE Transactions on Neural Networks*, 1(1), 1990.

[136] N.M. Nasrabadi and C.Y. Choo. Hopfield network for stereo vision correspondence. *IEEE Transactions on Neural Networks*, 3(1):5–13, 1992.

[137] S.K. Nayar, H. Murase, and S.A. Nene. Parametric appearance representation. In S.K. Nayar and T. Poggio, editors, *Early Visual Learning*. Oxford University Press, 1996.

[138] S.K. Nayar, S.A. Nene, and H. Murase. Real-time 100 object recognition system. In *Proceedings of ARPA Image Understanding Workshop*, 1996.

[139] U. Neisser. *Cognitive psychology*. Appleton-Century-Crofts, New York, 1967.

[140] R. Nelson and A. Selinger. A cubist approach to object recognition. In *Proceedings of International Conference on Computer Vision*, pages 614–621, 1998.

[141] R. Nelson and A. Selinger. Large-scale tests of a keyed, appearance-based 3-D object recognition system. *Vision Research*, 38:2469–88, 1998.

[142] E. Oja. *Subspace methods of pattern recognition*. Research Studies Press, Hertfordshire, 1983.

[143] B.A. Olshausen, C.H. Anderson, and D.C. Van Essen. A neurobiological model of visual attention and pattern recognition based on dynamic routing of information. *J. Neuroscience*, 13(11):4700–4719, 1993.

[144] G.A. Orban, J. De Wolf, and H. Maes. Factors influencing velocity coding in the human visual system. *Vision Research*, 24:33–39, 1984.

[145] H. Öğmen and S. Gagné. Neural network architectures for motion perception and elementary motion detection in the fly visual system. *Neural Networks*, 3:487–505, 1990.

[146] S.E. Palmer. *Vision science: photons to phenomenology*. MIT Press, 1999.

[147] S.J. Perantonis and P.J.G. Lisboa. Translation, rotation, and scale invariant pattern recognition by high-order neural networks and moment classifiers. *IEEE Transactions on Neural Networks*, 3(2):241–251, 1992.

[148] D.I. Perret, P.A.J. Smith, D.D. Potter, A.J. Mistlin, A.S. Head, A.D. Milner, and M.A. Jeeves. Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proceedings of the Royal Society of London B*, 223:293–317, 1985.

[149] E. Persoon and K.S. Fu. Shape discrimination using Fourier descriptors. *IEEE Transactions on Systems, Man, and Cybernetics*, 7(3):170–179, 1977.

[150] D.A. Pollen and S.F. Ronner. Visual cortex neurons as localized spatial frequency filters. *IEEE Transactions on Systems, Man, Cybernetics*, SMC-13(5):907–916, 1983.

[151] M. Porat and Y.Y. Zeevi. Localized texture processing in vision: analysis and synthesis in the Gaborian space. *IEEE Transactions on Biomedical Engineering*, 36(1):115–129, 1989.

[152] M.I. Posner, A.W. Inhoff, F.J. Friedrich, and A. Cohen. Isolating attentional systems: A cognitive-anatomical analysis. *Psychobiology*, 15:107–121, 1987.

[153] S.-Z. Qin, H.-T. Su, and T.J. McAvoy. Comparison of four neural net learning methods for dynamic system identification. *IEEE Transactions on Neural Networks*, 3(1), 1992.

[154] R.P.N. Rao. *Dynamic appearance-based vision*. PhD thesis, Department of Computer Science, University of Rochester, 1997.

[155] M.J. Riddoch and G.W. Humphreys. The effect of cueing on unilateral neglect. *Neuropsychologia*, 21:589–599, 1983.

[156] D.L. Robertson, M.E. Goldberg, and G.B. Stanton. Parietal association cortex in the primate: sensory mechanisms and behavioral modulation. *Journal of Neurophysiology*, 41:910–932, 1978.

[157] D.F. Rogers. *Mathematical elements for computer graphics*. McGraw-Hill, New York, 1990.

[158] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning internal representations by error propagation. In D.E. Rumelhart and J.L. McClelland, editors, *Parallel distributed processing: Explorations in the microstructure of cognition.*, pages 318–362. Cambridge, MA: MIT Press, 1986.

[159] I.A. Rybak, A.V. Golovan, V.I. Gusakova, and N.A. Shevtsova L.N. Podladchikova. A neural network system for active visual perception and recognition. *Neural Networks World*, 4:245–250, 1991.

[160] S. Satoh, J. Kuroiwa, H. Aso, and S. Miyake. Recognition of rotated patterns using neocognitron. In *Proc. Int. Conf. Neural Information Processing*, volume 1, pages 112–116, 1997.

[161] S. Satoh, J. Kuroiwa, H. Aso, and S. Miyake. Pattern recognitionsystem with top-down process of mental rotation. In *Proc. of IWANN'99*, volume 1, pages 816–825, 1999.

[162] M. Seibert and A.M. Waxman. Adaptive 3D-object recognition from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:107–124, 1992.

[163] M. Seibert and A.M. Waxman. Learning and recognizing 3D objects from multiple views in a neural system. In H. Wechsler, editor, *Neural Networks for Perception, vol 1, Human and Machine Perception*, pages 427–444. San Diego, CA: Academic Press, 1992.

[164] R. Sekuler and R. Blake. *Perception*. McGraw-Hill, 1994.

[165] A. Selinger and R.C. Nelson. Improving appearance-based object recognition in cluttered backgrounds. Technical Report TR725, Department of computer science, University of Rochester, 2000.

[166] M. Shah and R. Jain. Visual recognition of activities, gestures, facial expressions and speech: an introduction and a perspective. In M. Shah and R. Jain, editors, *Motion-based recognition*, pages 1–14. Kluwer Academic Publishers, 1997.

[167] R.N. Shepard and L.A. Cooper. *Mental images and their transformations*. Cambridge, MA: MIT Press, 1982.

[168] R.N. Shepard and J. Metzler. Mental rotation of three-dimensional objects. *Science*, 171:701–703, 1971.

[169] S. Shepard and D. Metzler. Mental rotation: effects of dimensionality of objects and type of task. *Journal of Experimental Psychology: Human Perception and Performance*, 14(1):3–11, 1988.

[170] S. Shipp and S. Zeki. Segregation of pathways leading from area V2 to areas V4 and V5 of macaque monkey visual cortex. *Nature*, 315:322–325, 1985.

[171] G.L. Shulman, R.W. Remington, and J.P. McLean. Moving attention through space. *Journal of Experimental Psychology: Human Perception and Performance*, 5:522–526, 1979.

[172] A.M. Sillito, H.E. Jones, G.L. Gerstein, and D.C. West. Feature-linked synchronization of thalamic relay cell firing induced by feedback from the visual cortex. *Nature*, 369:479–482, 1994.

[173] A.T. Smith and G.K. Edgar. The influence of spatial frequency on perceived temporal frequency and perceived speed. *Vision Research*, 30(10):1467–1474, 1990.

[174] P. Soille. *Morphological image analysis: principles and applications*. Springer-Verlag Berlin Heidelberg, 1999.

[175] G. Sperling and E. Weichselgartner. Episodic theory of the dynamics of spatial attention. *Psychological Review*, 102:503–532, 1995.

[176] L. Spirkovska and M.B. Reid. Robust position, scale, and rotation invariant object recognition using higher-order neural networks. *Pattern Recognition*, 25(9):975–985, 1992.

[177] L. Spirkovska and M.B. Reid. Coarse-coded higer-order neural networks for PSRI object recognition. *IEEE Transactions on Neural Networks*, 4(2):276–283, 1993.

[178] N. Srinivasa and M. Jouaneh. An invariant pattern recognition machine using a modified ART architecture. *IEEE Transactions on Systems, Man, and Cybernetics*, 23(5):1432–1437, 1993.

[179] L.S. Stone and P. Thompson. Human speed perception is contrast dependent. *Vision Research*, 32, 1992.

[180] P. Suetens, P. Fua, and A.J. Hanson. Computational strategies for object recognition. *ACM Computing Surveys*, pages 5–56, 1992.

[181] H. Szu and B. Telfer. Automatic target recognition. In M.A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 123–126. MIT Press, 1995.

[182] Y. Takano. Perception of rotated forms: a theory of information types. *Cognitive Psychology*, 21:1–59, 1989.

[183] M.J. Tarr and S. Pinker. When does human object recognition use a viewer-centered reference frame? *Psychological Science*, 1(4):253–256, 1990.

[184] P. Thompson. Perceived rate of movement depends on contrast. *Vision Research*, 22:377–380, 1982.

[185] P. Thompson. The coding of velocity of movement in the human visual system. *Vision Research*, 24(1):41–45, 1984.

[186] A. Treisman. The role of attention in object perception. In O.J. Braddick and A.C. Sleigh, editors, *Physical and biological processing of images*. Springer-Verlag, Berlin, 1983.

[187] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12, 1980.

[188] Y. Tsal. Movements of attention across the visual field. *Journal of Experimental Psychology: Human Perception and Performance*, 9:523–530, 1983.

[189] S. Ullman. Models of image segmentation and object recognition. In T.A. Poggio and D.A. Glaser, editors, *Exploring Brain Functions: Models in Neuroscience*. John Wiley & Sons Ltd., 1993.

[190] S. Ullman. Sequence seeking and counterstreams: a model for bidirectional information flow in the cortex. In C. Koch and J.L. Davis, editors, *Large-scale neuronal theories of the brain*. MIT Press, 1994.

[191] S. Ullman. *High-level vision: object recognition and visual cognition.* MIT Press, 1996.

[192] S. Usui, S. Nakauchi, and M. Nakano. Reconstruction of Munsell color space by a five-layer neural netowrk. *Journal of the Optical Society of America*, 9(4):516–520, 1992.

[193] D.C. Van Essen, C.H. Anderson, and D.J. Felleman. Information processing in the primate visual system: An integrated system persceptive. *Science*, 1992.

[194] D.C. Van Essen and J.H.R. Maunsell. Hierarchical organization and functional streams in the visual cortex. *Trends in Neurosciences*, 6:370–375, 1983.

[195] B.A. Wandell. *Foundations of Vision*. Sinauer Associates, Inc., 1995.

[196] Z. Wang and J. Ben-Arie. Generic object detection using model based segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 428–433, 1999.

[197] A.M. Waxman, M. Seibert, and I.A. Bachelder. Visual processing of object form and environment layout. In M.A. Arbib, editor, *The handbook of brain theory and and neural networks*, pages 1021–1024. MIT Press, 1995.

[198] E. Weichselgartner and G. Sperling. Dynamics of automatic and controlled visual attention. *Science*, 238:778–780, 1987.

[199] J. Westmacott, P. Lozo, and L. Jain. Distortion invariant selective attention adaptive resonance theory neural network. In *Proceedings of the Third International Conference on Knowledge-Based Intelligent Information Engineering Systems, KES'99*, pages 13–16, 1999.

[200] H. White. Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, 3:535–549, 1990.

[201] B. Widrow, R.G. Winter, and R.A. Baxter. Layered neural nets for pattern reocgnition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7):1109–1118, 1988.

[202] J.M. Wolfe. Extending guided search: Why guided search needs a preattentive "item map". In A.F. Kramer, M.G.H. Coles, and G.D. Logan, editors, *Converging operations in the study of visual selective attention*, pages 247–270. American Psychological Association, 1996.

[203] J.M. Wolfe and K.R. Cave. Deploying visual attention: The guided search model. In A. Blake and T. Troscianko, editors, *AI and the Eye*. John Wiley & Sons Ltd., 1990.

[204] R.P. Würtz. Object recognition robust under translation, deformations, and changes in background. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):769–775, 1997.

[205] J. Yang and S.B. Stevenson. Effects of spatial frequency, duration, and contrast on discriminating motion directions. *Journal of the Optical Society of America A : Optics and Image Science*, 14(9):2041–2048, 1997.

[206] S. Yantis. On analog movements of visual attention. *Perception & Psychophysics*, 43:203–206, 1988.

[207] S. Zeki. The representation of colours in the cerebral cortical of the monkey. *Nature*, 284, 1980.

[208] S. Zeki. *A vision of the brain.* Blackwell Scientific Publications, 1993.

[209] J. Zhang, Y. Yan, and M. Lades. Face recognition: Eigenfaces, elastic matching, and neural nets. *Proceedings of IEEE*, 85:1422–1435, 1997.

[210] Q. Zhu. Quantitative object motion prediction by an adaptive resonance theory (ART) neural network. In *ACC/WA2*, 1992.

[211] A. Zisserman, D. Forsyth, J. Mundy, C. Rothwell, J. Liu, and N. Pillow. 3D object recognition using invariance. *Artificial Intelligence*, 78:239–288, 1995.