1994

A Model Of Visual Recognition Implemented Using Neural Networks

Vincent C. Phillips
Edith Cowan University

Edith Cowan University

Copyright Warning

USE OF THESIS

The Use of Thesis statement is not included in this version of the thesis.

A Model of Visual Recognition Implemented Using Neural Networks

By
Vincent Clive Phillips

A Thesis
Submitted to the Faculty of Science and Technology
Edith Cowan University
Perth, Western Australia

In Partial Fulfilment of the Requirements for the Degree
of
Master of Applied Science (Computer Studies)

May, 1994

ABSTRACT

The ability to recognise and classify objects in the environment is an important property of biological vision. It is highly desirable that artificial vision systems also have this ability. This thesis documents research into the use of artificial neural networks to implement a prototype model of visual object recognition. The prototype model, describing a computational architecture, is derived from relevant physiological and psychological data, and attempts to resolve the use of structural decomposition and invariant feature detection.

To validate the research a partial implementation of the model has been constructed using multiple neural networks. A linear feed-forward network performs pre-processing after being trained to approximate a conventional statistical data compression algorithm. The output of this pre-processing forms a feature vector that is categorised using an Adaptive Resonance Theory network capable of recognising arbitrary analog patterns.

The implementation has been applied to the task of recognising static images of human faces. Experimental results show that the implementation is able to achieve a 100% successful recognition rate with performance that degrades gracefully. The implementation is robust against facial changes, minor occlusions, and it is flexible enough to categorise data from any domain.

DECLARATION

I certify that this thesis does not incorporate, without acknowledgment, any material previously submitted for a degree or diploma in any institution of higher education and that, to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where due reference is made in the text.

28/11/94

ACKNOWLEDGMENTS

I wish to acknowledge the supervision, guidance, and trust of Dr J W L Millar and Dr C L Smith. I would also like to thank Edith Cowan University for providing the means to complete this study, and the Department of Computer Science for providing much needed financial assistance. Finally, I would like to thank all the P'grads of GP1.07, especially Stephan Bettermann and Andrew Mehnart, for sharing the insanity.

This work, like everything I do, is dedicated to the four people who share my life, my family. This one is especially for Rachel, Anthony, and Jason, the best children a parent could love.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

Section 1: Introduction

The ability to recognise objects is an important property of biological vision. It allows us to process, categorise, and interact with the patterns of our environment. It is highly desirable for artificial vision systems to have this ability. Although artificial ***neural networks*** are recognised as powerful pattern recognition devices, little effort appears to have been made to establish their part in ***systems*** for object recognition. This section introduces these issues, containing:

- A brief discussion of models of perception, especially the human visual system.
- The motivation for modelling the computational behaviour of the visual system and justification for using neural networks to implement the model.
- The objectives that governed the direction of this research.
- An overview of the contents of the thesis.

1.1 Background

Computational models are still a long way from capturing the flexibility and generality of biological visual processes. This is despite the wealth of physiological and psychological data, and the considerable research that has been invested over the last century. The impact of providing our machines with the means to directly perceive their environment is significant, and research efforts to this end can easily be justified. This thesis documents research into the use of artificial ***neural networks*** to implement a prototype model of visual ***object recognition.***

Discussion of more general problems in computational vision is provided by many authors. The influential monograph by Marr (1982) is probably required reading for research in this field. Other recommended surveys are Barrow and Tenenbaum (1986), Ballard and Brown (1982), and Hildreth and Ullman (1989). Inspiration for the approach taken during the review and the theoretical formulation is provided by Overington (1992).

1.1.1 Models of Perception

Two principles have, historically, guided models of the way we process visual patterns. The first, known as the *principle of prägnanz*, is based upon *Gestalt* laws of organisation and grouping (Pomerantz and Kubovy, 1986). The second, the *principle of likelihood*, is derived from the work of Helmholtz, Boring, and others, and describes perception as hypothesis testing and matching (Pomerantz and Kubovy, 1986; Bruce and Green, 1990).

The principle of prägnanz states that the visual organisation we perceive will be the one with the least complexity, the organisation with the most stable geometrical arrangement. That is, perceptions are *globally* organised "to simplify maximally the representation of holistic stimulus configurations" (Pomerantz and Kubovy, 1986, p.14). The organisational principles most often considered are *proximity, similarity, common fate, continuation, closure, relative size, surroundedness, orientation,* and *symmetry* (Bruce and Green, 1990). These properties were considered to mirror innate structures in the brain (Pomerantz and Kubovy, 1986).

The principle of likelihood states that perceptions are organised "into the most *probable* [italics added] object or event . . . in the environment consistent with the sensory data" (Pomerantz and Kubovy, 1986, p.9). The likelihood principle describes perception as a process of hypothesis testing, rejection, and verification, i.e. *reasoning*. What is, perhaps, unique is the idea that this reasoning process is separate from conscious intelligence using what Helmholtz terms *unconscious inference*.

The surveys by Pomerantz and Kubovy (1986) and Chase (1986) are recommended for a more thorough treatment than that given above. These works also describe more current approaches and provide reference to the work of Helmholtz, the Gestalt psychologists, and the structuralists. The *ecological* approach of Gibson provides a different view of perception, considering perception to be the passive detection of the invariant information in the environment (Gibson, 1966).

Modern theories of perception generally lie between these two extremes, although it might be hypothesised that differences in current models are similar to the differences between prägnanz and likelihood, see, for example, the argument of Grossberg (1984/1987). Unfortunately, the data regarding perception is often contradictory and paradoxical (Grossberg, 1987), leading to an abundance of disparate models. This makes it difficult to determine an appropriate theoretical framework. Hildreth and Ullman conclude "that in object recognition, which is one of the most fundamental aspects of human vision, theories (as well as experimental work) still have a long way to go" (Hildreth and Ullman, 1989, p.40).

1.1.2 A Suitable Design Approach

Research into neural networks has centred upon *low-level* problems, for example, sufficiently powerful learning algorithms, which has sometimes been at the expense of higher level design issues (Mrsic-Flogel, 1991). Perhaps a more suitable approach is to consider perception as a process producing a description of the external world that is useful and not cluttered with irrelevant information (Marr, 1982). Thus vision can be described as a mapping from one *representational* domain to another - beginning with a two-dimensional array of sensor information that is transformed, through a hierarchy of processes, into a concise description of the objects in the image (Marr, 1982; Barrow and Tenenbaum, 1986). According to Marr (1982) vision must be described and understood at three distinct levels:

1. As a *computational* model, defining the goals and the function of the activity.
2. As an *algorithmic* process, describing the representations appropriate to a computational model.
3. As a description of the specific *implementation* of a computational model.

This strategy, typical of traditional artificial intelligence techniques, is concerned with explaining the abilities in question, thus postponing detailed empirical validation of the algorithms used in the model (Pylyshyn, 1989).

1.1.3 Artificial Neural Networks

A more detailed discussion of the networks used to implement the model can be found in later sections. Here it will suffice to justify their deliberate use to implement the model. Neural networks are recognised as powerful pattern recognition devices, capable of considerable functional adaption and generalisation (Pao, 1989). Perhaps their most attractive features are massive parallelism, simple computational units that are inherently *brain-like*, and algorithmic flexibility - properties that are highly desirable for artificial vision systems (Feldman, 1985; Uhr, 1987; Barrow and Tenenbaum, 1986). Although it is possible for other techniques to exhibit these properties (Uhr, 1987; Overington, 1992) they are available, almost by definition, with neural networks.

Some familiarity with general concepts of neural networks will be henceforth assumed, including knowledge of *back-propagation* (Rumelhart, Hinton, and Williams, 1986), *competitive learning* (Kohonen, 1982; Rumelhart and Zipser, 1985), *Hebbian* learning (Hebb, 1949), and *Adaptive Resonance Theory* (Carpenter and Grossberg, 1987/1991a).

1.2 Research Objectives

The objective of this research was to determine how a general model of object recognition could be implemented using neural networks. This objective was divided into four separate aims:

1. The design of a computational model of object recognition.
2. The selection of neural networks appropriate to the task of object recognition.
3. The development of a prototype implementation using a number of simplifying assumptions and limitations.
4. Validation of the model, and the implementation, by testing classification ability.

To validate the model it was decided to use a single domain representing data similar to that found in *real-world* applications - in this case, images of human faces.

The first aim, to design a computational model, was pursued with the intent of not explicitly contradicting known facts about biological systems. As biological vision is the best example we have that vision is possible it is wise not to ignore how these mechanisms perform the task: they may show how the process should be modelled.

The fourth aim, validation and testing, was designed to quantify the robustness, i.e. predicability, of the techniques used. The third aim obviously interacted with this goal, especially the limitations that were required. To minimise the impact of this interaction both the model and the implementation were designed to be portable, and as domain independent as possible.

1.3 Hypothesis

The hypothesis of the research can now be stated as follows:

> **A general model of object recognition will, when implemented using artificial neural networks, reliably classify images with performance that degrades predictably as the recognition task becomes more difficult.**

The hypothesis has been verified by the prototype implementation and the validating application. Although testing was limited, the results show that the hypothesis is tenable and should be subjected to further examination.

1.4 Thesis Structure

The thesis is developed conventionally. Section two reviews the literature from a number of disciplines, describing relevant physiological, psychological, and computational data. Section three derives an appropriate computational model of object recognition, and describes the prototype instantiation used to test the model. Section four details the validating application, the experimental hypotheses, and the results of three simple experiments. Section five generalises from the results from section four, discussing their immediate implications, and some general observations derived from the experiments. The thesis concludes with section six, where the significance of the results, and the research generally, are discussed. At the end of the thesis can be found the list of references consulted during this research.

1.5 Summary

This section has described some preliminaries needed to place the thesis in an appropriate context, and refers to the ideas that have led to the research hypothesis. Objectives for the research were derived, implicitly and explicitly, from the background literature and they provide a series of goals that will be validated in the conclusion. Most importantly, a suitable approach has been identified and this has been used to structure both the research and the thesis, including the next section - a review of the relevant literature and data.

Section 2: Visual Recognition

Biological vision has been the subject of research for over a century and it is the best known perceptual system. Vision can be described as a number of functionally distinct areas, with a definite hierarchy of processes and activities. In fact, it has been found that perceptual systems have a shallowly serial, massively parallel architecture and there are many computational descriptions of these processes. After a brief review of physiological constraints this section examines these issues further.

2.1 Background

Ramón y Cajal, born in the 1850's, stands out in the field of neuroanatomy. His major contributions being to establish that neurons act independently and to show that, by using the Golgi staining method, they form extremely complicated, but orderly, networks of cells (Hubel, 1988). From these beginnings understanding of the physiology of the visual system has increased dramatically, and the process can be now described reasonably well (Hubel, 1988).

Studies of cortical activity, especially in cases of brain ***lesions***, have shown that there are two visual processing mechanisms (McCarthy and Warrington, 1990; Desimone and Ungerleider, 1989; Carpenter, 1984):

- Spatial vision - associated with the posterior parietal cortex
- Pattern vision - associated with the inferior temporal cortex

Both mechanisms originate in striate cortex (primary visual cortex) and are made up of a number of distinct functional areas.

Computational models of biological vision have improved, and there are now a number of strategies. The more popular approaches are based upon the detection of ***invariant features***, or a representation derived from ***structural decomposition*** (Hildreth and Ullman, 1989).

Modelling visual recognition with artificial neural networks has a long history, beginning with the work of Pitts and McCulloch (1947/1988) through to more recent studies by Grossberg

(1983/1987) and Kohonen (1988). However, it has been recognised only recently that more global models of perception are required, resulting in a move towards perceptual *systems* rather than isolated networks of neurons (Mrsic-Flogel, 1991).

2.2 Physiological Vision

2.2.1 The Retina

Light data enters the retina and is converted into a pattern of neural activity by the rods and cones (Carpenter, 1984). Retinal response is transmitted to the ganglion cells, which make up the optic fibres, through the activities of *bipolar*, *horizontal*, and *amacrine* cells. Bipolar cells have been divided into *midget* and *diffuse* types and it is hypothesised that high resolution visual information from the fovea is directed through the midget bipolars to the ganglions, with approximately one to one correspondence between foveal cones, midget bipolars, and ganglions (Hood and Finkelstein, 1986; Overington, 1992; Hubel, 1988). Horizontal and amacrine cells are usually associated with lower resolution, achromatic vision and their function is species dependant (Carpenter, 1984).

Receptors instantaneously adapt to the current level of illumination, filtering out irrelevant illumination changes (Carpenter, 1984; Abramov and Gordon, 1974). This *field adaption* occurs through the bleaching of photopigments and can be modelled, for a large range of luminance levels, by a *Weber* ratio (Falmagne, 1986; Hood and Finkelstein, 1986; Carpenter, 1984; Gonzalez and Wintz, 1987).

Bipolar and ganglion cells are of two types - *on-centre, off-surround* and *off-centre, on-surround* - that occur in approximately equal numbers in the retina. This mechanism allows the retina to respond to both the presence, and the absence, of light.

2.2.2 Lateral Geniculate Nuclei (LGN)

The retinas send their output to the lateral geniculate nuclei through the optic fibres. The nuclei contain six layers, or laminae, each of which receives a retinotopic map of half the visual field (Abramov and Gordon, 1974; Hubel, 1988; Bruce and Green, 1990). In the macaque monkey three layers receive input from one eye, and three from the other eye (Hubel, 1988; Bruce and Green, 1990) - and there are good reasons to suppose that the early stages of human vision are similar (Hubel, 1988).

The LGN contain two distinct areas, the ventral, or lower, regions, and the dorsal, or upper, regions. In the two ventral layers are found *magnocellular* cells, with large receptive fields

(topographic regions that evoke a response), high acuity, and a broad-band centre-surround response to light. They are often thought to be colour blind and are assumed to be involved in form, depth, and movement perception (Hubel, 1988).

In the four dorsal layers are found ***parvocellular*** cells, with smaller receptive fields than the ventral cells, mixed acuity, and a complicated response to colour. Most cells have red-green opponent centre-surround responses, i.e. Red+Green-, Red-Green+, Green+Red-, Green-Red+ (Hubel, 1988). These cells seem to capture the information required for colour processing.

Many cells in both the dorsal and ventral regions have small receptive field centres, of approximately two minutes of arc in diameter. This is about the same acuity as the fovea and it seems that a single cone provides information for these centres (Hubel, 1988).

2.2.3 Striate Cortex (V1)

Axons from the LGN, see **Figure 2.1**, form the optic radiations that ascend and terminate in layer 4C of striate cortex (V1 or Area 17) - which contains a hierarchy of six separate layers (Hubel, 1988). Those axons originating in the two ventral layers of the LGN synapse in layer 4Cα, while those from the parvocellular layers end in layers 4Cβ and 4A (Hubel, 1988; Desimone and Ungerleider, 1989).

Layer 4Cα projects into layer 4B and, possibly, layers 2,3, and 6 (Desimone and Ungerleider, 1989). Like cells in the ventral regions of the LGN the cells in 4Cα display broad-band spectral properties with sensitivity to low contrasts (Desimone and Ungerleider, 1989). The cells in layer 4B selectively respond to the direction of stimulus motion and are probably important for motion analysis (Desimone and Ungerleider, 1989).

Projections from layer 4Cβ terminate in layers 2 and 3 (Hubel, 1988). Cells in these areas show spectral sensitivity similar to that found in the dorsal regions of the LGN, i.e. colour opponency. However, these cells have more varied and complex receptive field properties and often form homogenous fields (Desimone and Ungerleider, 1989). A class of cells in the striate cortex also have a ***double colour opponent*** (DCO) response (Desimone and Ungerleider, 1989). DCO cells, which occur in regions called ***blobs*** (Hubel, 1988), respond to small spots of light with their preferred stimuli - red-green and yellow-blue. It is now assumed that these cells are devoted to colour processing and achieving colour constancy (providing colour information despite instantaneous illumination) (Hubel, 1988; Desimone and Ungerleider, 1988).

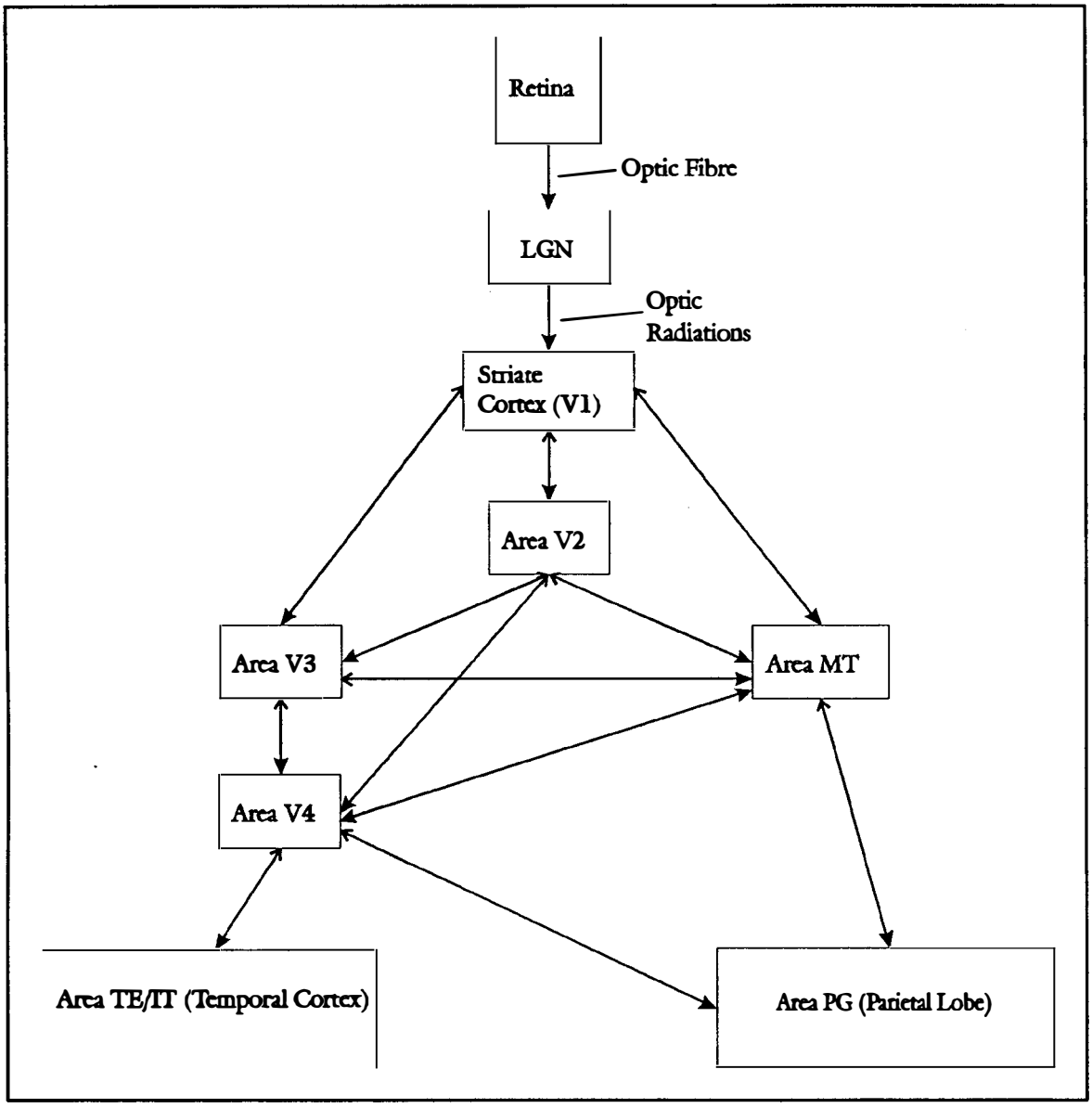

Figure 2.1. Diagram of visual areas showing major functional regions and connections. Heavy arrowheads indicate forward neural projections, light arrowheads show that some information feeds back.

Cells in striate cortex, except blobs and centre-surround cells in layer 4, selectively respond to visual stimuli at particular orientations (Hubel, 1988). Cells with similar orientation preference are grouped into columns that overlap with the ocular dominance columns. These *complex* cells all respond to properly oriented lines that occur within their receptive field (Hubel, 1988).

It is generally accepted that cells in striate cortex function as "general-purpose spatial filters that transform the visual image in a number of useful ways", including colour, orientation, movement, size, and spatial frequency (Desimone and Ungerleider, 1989). *Simple* cells, which occur in layer 4 and respond to properly oriented lines in *particular* positions, carry shading and three-dimensional

contour information, while complex cells carry information on fine surface textures (Desimone and Ungerleider, 1989).

2.2.4 Area V2

Area V2 in the visual cortex has not been studied as extensively as V1 but it appears that cells have many of the same properties (Desimone and Ungerleider, 1989). Area V2 is composed of three regions that, using staining techniques, appear as alternating thin and thick stripes separated by interstripe regions (Desimone and Ungerleider, 1989).

1. Thin stripes receive synapses from the blobs of striate cortex and project to area V4. Cells in this region may be selective for non-oriented colour information.
2. Thick stripes receive synapses from layer 4B of striate cortex, project to area MT, and are selective for orientation, disparity, and direction of motion.
3. Interstripe regions receive synapses from the interblob areas of striate cortex, project to area V4, and are selective to the length and orientation of stimuli.

Although there are some doubts about the functional classification it is generally agreed that the three regions perform different types of processing.

2.2.5 The Occipitotemporal Pathway

The impression that there are two separate physiological pathways, beginning with the magnocellular and parvocellular regions of the LGN, increases after area V2. Synapses from layer 4B of area V1 and the thick stripe region of V2 terminate in areas V3 and MT. Synapses from area MT project into the parietal lobe, which has been associated with motion analysis (Desimone and Ungerleider, 1989), localisation (Carpenter, 1984), geographical and spatial orientation, and coordination of multiple sensory modalities (Howard, 1974). This has led to this pathway being identified with the term *spatial* vision (Desimone and Ungerleider, 1989) and the neurophysiology is discussed in Goldberg and Colby (1989).

Area V4 receives synapses from the interstripe and thin stripe regions of V2, and synapses from area V3 and area MT. Projections from V4 terminate in areas MT and the parietal lobe, and, more importantly for this discussion, in inferior temporal cortex (IT). It is known that lesions in the IT area effect the ability to perform visual discrimination, visual recognition, and form and colour analysis (Desimone and Ungerleider, 1989; Carpenter, 1984; Bruce and Green, 1990; McCarthy and

Warrington, 1990). The term *pattern* vision has been used to describe the function of the occipitotemporal pathway (Desimone and Ungerleider, 1989).

Cells in IT have much larger receptive fields than those found in lower layers, and these have been associated with translational invariance (Desimone and Ungerleider, 1989). Most cells in IT have a selective response to stimuli based upon the *features* found in the stimuli, for example, shape, colour, and texture (Desimone and Ungerleider, 1989). Physiological evidence suggests "that the neural code for objects in IT must be a population code [response of multiple cells] based on object features" (Desimone and Ungerleider, 1989).

Face Recognition. Cells that respond selectively to faces have been found in IT regions of primates and, initially, were thought to be examples of cells that coded individual objects (i.e. *grandmother* cells) (Bruce and Green, 1990; McCarthy and Warrington, 1990; Desimone and Ungerleider, 1989). It is now known that these cells respond to different faces, expressions, and orientations, and are probably parts of a larger population code (Desimone and Ungerleider, 1989).

2.3 *Models of Recognition*

It has been suggested that object recognition be achieved by the formation of *perceptual*, or *semantic*, categories (Desimone and Ungerleider, 1989; McCarthy and Warrington, 1990). This view is supported by lesion studies and various forms of visual agnosia. It has been shown that while localised brain damage can lead to a failure to discriminate *specific* attributes, the ability to identify objects within *broad* categories may be retained (McCarthy and Warrington, 1990; Desimone and Ungerleider, 1989).

2.3.1 Memory Organisation

In a general sense, the model described above can be considered as a continuously evolving associative memory. Such a memory would be distributed across the cortex with areas specifically associated with different sensory modalities (McCarthy and Warrington, 1990). Such a model can be used to explain how recognition can occur when partial, or imperfect, cues are provided (Damasio, Tranel, and Damasio, 1989, p.10);

> The overall mapping of any entity and by extension, any event is the potential sum total of sub-representations available in sensory and motor cortices. It follows . . . that the recognition of an object is determined by which feature, dimension and context is offered perceptually . . . to the available *representation* [italics added].

This view of memory and recognition emphasises the importance of both internal and perceived representation. Although the issue of representation does not describe a theory of recognition it is important because it influences the nature of such a theory.

2.3.2 Stimulus Equivalence

Before examining representation it is instructive to consider the problem of ***stimulus equivalence***, as described by Bruce and Green (1990, p.177);

> If the stimulus controlling behaviour is a pattern of light, or image, on the retina, then an infinite number [possible distributions of light] are equivalent in the effects, and different from other sets of images.

This description can be extended to any sensory process where a sample of the continuous environment is made. The problem of describing the recognition of an infinite number of sensory patterns has led to different views of the internal representation constructed by biological systems.

2.3.3 Structural Representation

Sutherland (1973) describes object recognition as involving the "formation, storage and retrieval of ***structural*** [italics added] descriptions. A description comprises a list of entities, the properties of those entities and the relationships obtaining between them" (Sutherland, 1973, p.2). Marr (1982) supports and expands this view, arguing that "object recognition demands a stable shape description that depends little, if at all, on the viewpoint" (Marr, 1982, p.295). Marr describes such a description, calling it the *3D representation*, a representation describing objects hierarchically using both volumetric and surface primitives (Marr, 1982).

The approach described by Sutherland (1973), and expanded by Marr (1982), Ballard and Brown (1982), and Barrow and Tenenbaum (1986), can be broadly described as structural decomposition, or decomposition into parts (Treisman, 1986). This is an analytical approach: objects are represented as conjunctions of symbolic primitives. The approach has widespread appeal and there is a wealth of psychological evidence for such a process (Treisman, 1986). Structural descriptions can be categorised using the techniques of ***syntactic*** pattern recognition, where the representation is parsed in much the same way that a sentence in a formal grammar is processed (Fu, 1980; Fu, 1986; Rosenfeld, 1986).

2.3.4 Invariant Features

A problem with structural decomposition is that it is difficult to determine the structural primitives that should be used with a large range of naturally occurring stimuli, i.e. those that are not composed of straight lines and simple geometric shapes (Treisman, 1986). An alternate representation is based upon the detection, and recognition, of *invariant* features, as described by Gibson (Gibson, 1966, p.278);

> The judgement of "same" reflects the tuning of a perceptual system to the invariants of stimulus information that specify the same real place, the same real object, or the same real person. The judgement of "different" reflects the absence of invariants, or sometimes the failure of the system to pick up those that exist.

Recognition is achieved by the perceptual system *resonating* to the invariant features detected (Gibson, 1966). This is a *holistic* (Treisman, 1986), or ecological (Bruce and Green, 1990), approach that is based upon the assumption that an object is recognised as a whole, rather than a collection of parts.

Unfortunately, it is difficult to decide (1) what invariant features are actually used, and (2) how the idea of invariant features can operate in view of the stimulus equivalence problem (Hildreth and Ullman, 1989; Bruce and Green, 1990; Marr, 1982). A very simple solution to the second problem, which is the most serious, is to store many different descriptions of an object, each description representing a different view of the object. Obviously such an approach will, eventually, require an infinite memory (Hildreth and Ullman, 1989; Treisman, 1986).

A more flexible approach, with psychological validity (Treisman, 1986; Grossberg, 1983/1987; Rock, 1986), is described as *perceptual learning* (Treisman, 1986, p.47);

> Perceptual learning is seen as a process of (1) abstracting the central tendency, the most typical instances, rather than defining the boundary conditions, and (2) learning which transformations or dimensions of variation are acceptable without changing the identity of the object.

This view fits very well with the idea of perceptual categories, or *prototypes*, and is particularly applicable when verbal descriptions are not available or required (Treisman, 1986).

Unlike hierarchical structural descriptions, invariant features are usually captured within the form of a *feature vector*. Matching a feature vector with an internal description occurs in a high dimensional feature space (Pao, 1989), and can be viewed as a process of *hypothesis* testing (Grossberg, 1984/1987), or top-down matching (Treisman, 1986). The techniques of *statistical* pattern

recognition can be used to match feature vectors (Habibi, 1986; Fukunaga, 1990; Therrien, 1989). Category formation can be seen as a process of clustering similar vectors in feature space.

2.3.5 Parallel Processes

Physiological and psychological evidence suggests that two separate processes occur in the visual cortex, one concerned with spatial processing and movement, the other with semantic processing and object recognition. Both pathways originate from striate cortex, they appear to operate in parallel (Treisman, 1986), and it is possible that they communicate through cortical areas MT, V3, and, probably more importantly, V4 (see 2.2.5; Desimone and Ungerleider, 1989).

There is compelling evidence that visual stimuli are described analytically, i.e. decomposed into structural components. The resulting description appears to use a canonical coordinate frame that is updated as objects move and change. The evidence for structural decomposition seems to conflict with evidence suggesting a more holistic approach, where the total configuration of features in an object is important. Treisman (1986) suggests that, after initial sensory processing, a holistic analysis may take precedence over decomposition. The existence of two methods of recognition is evidenced by the ability of some face agnosic patients to perform *fragmentary* recognition, i.e. recognise parts of an object rather than the whole object (Damasio et al., 1989).

From the above, it seems reasonable to assert that spatial vision requires the construction of a representation similar to the one described by Marr (1982), while pattern vision may operate on a set of invariant features. Although little is known about the mechanisms used for pattern vision, it is clear that spatial vision occurs in the parietal lobe, which is involved in spatial processing. Computational approaches usually assume that structural decomposition and feature analysis are mutually exclusive. This is not necessarily so, the evidence suggests that structural decomposition and feature analysis should both be used and performed in parallel.

2.4 Recognition Using Neural Networks

Neural networks are usually viewed as massively parallel machines composed of simple computational elements, called neurons. Neurons communicate state information through weighted uni-directional (or bi-directional) links called synapses. Thus a neural network can be considered to form a *graph*, and many architectures, learning algorithms, and neuron models have been described.

Neural networks have been applied to many problem domains but since their first practical description (McCulloch and Pitts, 1943/1988; Pitts and McCulloch, 1947/1988) they have been used

to model low level sensory processes (Kohonen, 1982), particularly vision (Linsker, 1988; Azencott, Doutriauz, and Younes, 1990; Rybak, Shevtsovam and Sandler, 1992; Kollias, Tirakis, and Milios, 1991). However, perhaps the most comprehensive description of neural networks as sensory processors is by Stephen Grossberg and Gail Carpenter (Carpenter, 1989/1991; Carpenter and Grossberg, 1987/1991a, 1987/1991b; Grossberg. 1983/1987, 1987/1988a, 1989, 1988/1990, 1976/1991a, 1976, 1991a, 1976, 1991b, 1988/1981c).

2.4.1 Dimensionality Reduction

The association process can be viewed as a mapping from a high dimensional input space to an output space of lower dimensions, or dimensionality reduction. Specialised architectures and learning algorithms (Oja, 1982; Sanger, 1989; Baldi and Hornik, 1989; Linsker, 1988; Rumelhart, Hinton, and Williams, 1986) have been described and investigated, the general result being that the network learns to perform an eigenvector transformation and hence extract the principal components of the input data (Oja, 1982; Gonzalez and Wintz, 1989).

2.4.2 Associative Memory

Neural networks can be considered to operate as associative memories, i.e. associating similar groups of input patterns to desired output patterns. Appropriate architectures and algorithms are presented by Hopfield (1982), Hinton, Sejnowski, and Ackley (1984), Peterson and Anderson (1987), and Kohonen (1988). In effect, neural networks perform a clustering operation using significant features in the input vectors, i.e. they are performing pattern recognition.

Kohonen (1988) and Weiss and Kapouleas (1989/1992) have both shown that neural networks can perform as well as, or better than, conventional statistical and syntactic techniques. However, although they have been used to classify structural descriptions (Sabbah, 1988; Feldman, 1985) they are not particularly suited to processing hierarchical symbolic information.

2.4.3 System Models

Mrsic-Flogel argues that there is an increased need for research into "high-level systems design of neural architectures" (Mrsic-Flogel, 1991, p.1). He describes an architecture suitable for perceptual processing, consisting of three layers;

1. A sensory level dedicated to input/output processing.

2. A cognitive layer performing the transformation and normalisation required for cognitive functions.

3. A layer monitoring and controlling the other layers.

As previously identified, biological sensory processes are shallowly serial yet massively parallel, a model not matched by single neural network architectures. It seems obvious that a neural network *embedded* within a serial system architecture is required. It is interesting that conventional processing models, for example, Marr (1982), follow this systematic approach.

2.5 Summary

A review of the physiological processes reveals that vision can be viewed as a series of distinct processes. Importantly, two parallel biological processes can be identified in spatial and pattern vision - which may be hypothesised to use different representations and mechanisms. This hypothesis would account for conflicting psychological data, and provides a framework for the use for existing computational theory.

Neural networks have been found to be efficient statistical classifiers and can be used to perform pattern recognition and, hence, pattern vision. However, a single neural network cannot perform the functions attributed to biological vision. It is suggested that the use of a serial system model with multiple processing stages is more suitable.

Section 3: Computational Model

The physiological model identified in the last section has been used to derive a computational architecture. This section details this model, containing:

- A description of the functional architecture for the model.
- A description of the implementation constructed for this research.
- A discussion of the neural networks used for the implementation of the model.

3.1 A Prototype Model

Input to the model is a two-dimensional array of spatially sampled attributes. Output is an appropriate semantic classification and a structural description, allowing further syntactic processing of an object.

It is known that the retina samples a scene several times per second (Hubel, 1988), implying that the visual system operates in discrete time steps with an object undergoing a number of distinct transformations before recognition occurs. Finally, it is assumed that only high-resolution foveal information is important for object recognition.

3.1.1 Design Criteria

The following criteria were used during design of the model:

- The model must be clearly delineated from any implementation, i.e. processing stages must be specified in terms of inputs, outputs, and general function only.
- The model must be suitable for parallel implementation, although a shallow hierarchy of processing steps is desirable.

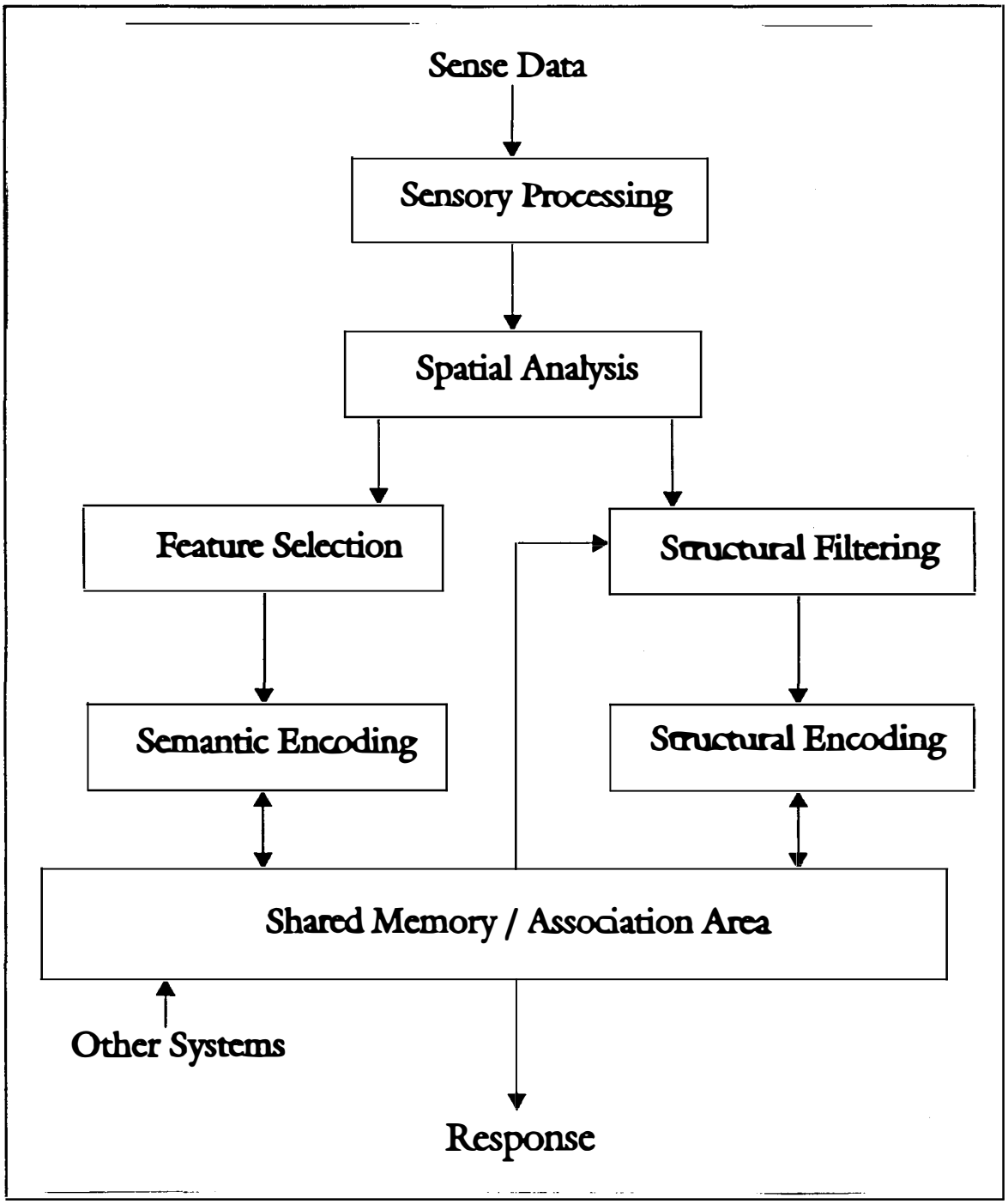

Figure 3.1. Prototype computational model of object recognition showing processing hierarchy.

- It must be possible to enhance the architecture of the model without disturbing existing processes.

Due to the limited nature of this research it is also assumed that a partial implementation of the model is valid. It is hoped that such an implementation can be used for future research.

3.1.2 Sensory Processing

Sensory processing is intended to capture some of the functionality of the retina and the LGN. Input to the process is a two dimensional array of ***tri-chromatic coefficients***. The primary purpose of sensory processing is to capture ***absolute*** variations in the input data. Secondary purposes are to prepare the data for later processing, to preserve essential features in the image, and to reduce the dimensionality of the image without degrading the ***quality*** of the information. Output from this processing retains the form of the original input.

3.1.3 Spatial Analysis

Spatial analysis requires several distinct processes that, like the first levels of striate cortex, characterise the image along spatial dimensions - for example, contour and textural information, colour, and intensity change. It is possible that illusory and obscured information is ***filled-in*** at this stage. Output retains the form of the input data; textural and colour information is highlighted for the feature selection, while contour, edge, and depth information is highlighted for structural filtering.

3.1.4 Structural Filtering

Structural filtering uses conventional techniques to extract the information required to form a structural description. This can be considered similar to Marr's 2½-D sketch where orientation, surface depth, and contour information is made explicit (Marr, 1982). Generally this stage is associated with describing the properties of the surfaces found in an image. A top-down signal from the shared memory / association area indicates whether more intensive processing needs to be applied to the input data, for example, a rotation transformation.

3.1.5 Feature Selection

Description of feature selection and detection is problematic, requiring definition for the term ***feature***. Simon defines a feature in terms of the detection operation, i.e. "a feature is the operator itself" (Simon, 1989, p.2). Feature detection operations usually operate upon the statistical properties of an image, and this was the approach used here. However, any definition resulting in a set of features, or feature vector, that uniquely describes an image is sufficient.

3.1.6 Structural Encoding

Structural encoding uses the structural elements extracted by structural filtering to derive a hierarchical description of the surfaces and volumes present in an object. Description is in a canonical coordinate frame, i.e. object-centred, rather than viewer-centred, coordinates. As representation is hierarchical the resulting description will be graph-like and it may be matched against previously derived descriptions using conventional graph-matching, or the techniques developed for syntactic pattern recognition. Output of the process will be an identifier that may, or may not be, unique to an object, allowing access to information stored in the association layer.

3.1.7 Semantic Encoding

The purpose of semantic encoding is to derive a unique identifier for an object. As input to the process is a feature vector the encoding process will require some form of clustering operation in feature space. Thus an identifier is a cluster label, and identification will depend upon the distance between a cluster centre and the input vector. It is essential that cluster centres can be moved to accommodate changes in the appearance, or state, of objects (see 2.3.4).

3.1.8 Shared Memory Area

The shared memory area forms a database with indexes generated by structural and semantic encoding, and any other connected sensory systems. The database is local to the object recognition system and output might be directed to a global memory area. The association area matches the descriptions generated by structural encoding with those generated by semantic encoding. In the case of a mismatch it may request further processing - a request for a particular transformation of the input data or refinement of the encoding processes.

3.2 Implementation

3.2.1 Limitations

Because only a partial implementation of the model was possible some limitations and assumptions were required:

- It was assumed that structural encoding is not necessary for recognition in a controlled environment.
- Only achromatic information would be required.

- Objects of interest were centralised and isolated in the data source, i.e. fixation and figure-ground separation operations were not required.
- Only objects from a single domain - perceptual category - would be presented.

Sensory processing, feature selection, semantic encoding, and a simple association area were implemented.

3.2.2 Sensory Processing

Input consisted of two dimensional RGB images from a video camera. The image was converted into a grey scale image by retaining the magnitude of each RGB vector. Image size was standardised to 128x128 pixels using a median filter - there is some plausibility for assuming the fovea has similar spatial resolution (Overington, 1992).

Variation of luminance within each input image was determined using the Weber law shown in (3.1) where ΔI measures the instantaneous intensity increment, I is the background adapting intensity, and C is the constant contrast ratio.

$$C = \frac{\Delta I}{I} \tag{3.1}$$

This law provided the means to find the required ratios, resulting in a *reflectance pattern* invariant to fluctuations in background intensity. The variation used for the implementation is shown in (3.2) (Freeman and Skapura, 1991)[1], where θ_i describes the reflectance value for pixel location i, I_i describes the measured intensity at pixel location i, and N describes the size of the pixel neighbourhood.

$$\theta_i = \frac{I_i}{\sum_{k=1}^{N} I_k} \tag{3.2}$$

Olzak and Thomas (1986) suggest that this model is only suitable for patterns consisting of a single intensity change over a uniform background. However, for the purposes of the implementation it was found sufficiently powerful and it is a well known technique in neural network processing (Freeman and Skapura, 1991). The implementation used a neighbourhood of eight pixels (Overington, 1992), although other values could be justified.

[1] For notational convenience only the one-dimensional case is presented, the two-dimensional notation is obvious.

3.2.3 Feature Selection

The use of the discrete *Karhunen-Loêve transformation* (KLT) is well known in statistical pattern recognition (Gonzalez and Wintz, 1987). The transformation is optimal in the sense that it captures significant statistical features and achieves large reductions in data dimensionality, while minimising the mean-square error (Habibi, 1986).

If it is valid to assume that the input data is stationary with a Gaussian-like distribution then it is possible to use an approximation to conventional KLT methods. This is a *linear*, *feed-forward*, neural network trained using *Generalised Hebbian Learning* (Sanger, 1989), referred to here as the *KLTNet*. The KLTNet can approximate the KLT to a reasonable level of accuracy (Sanger, 1989), and it has shown sufficient generalisation from its training data to serve as an arbitrary encoding mechanism (Phillips, 1993).

Consider the feed-forward network shown in **Figure 3.2.** When a vector $(x_1,...,x_N)$ is presented to the network, the output neurons $(y_1,...,y_M)$ update their state using the linear summation function shown in (3.3). Thus the output vector is the linear transformation $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{A}$ is the matrix of synaptic weights, $\boldsymbol{y}$ is the output vector $[y_1,...,y_M]$, $\boldsymbol{x}$ is the input vector $[x_1,...,x_N]$, and $M<N$ for a reduction transformation.

$$y_i = \sum_{j=1}^{N} x_j w_{ij} \qquad y_i, w_{ij} \in \Re, \quad x_j \in [0,1] \tag{3.3}$$

$$w_{ij}(t+1) = w_{ij}(t) + \gamma(t) y_i(t) \left(x_j(t) - \sum_{k<i} y_k(t) w_{jk}(t) \right) - \gamma(t) y_i^2(t) w_{ij}(t) \tag{3.4}$$

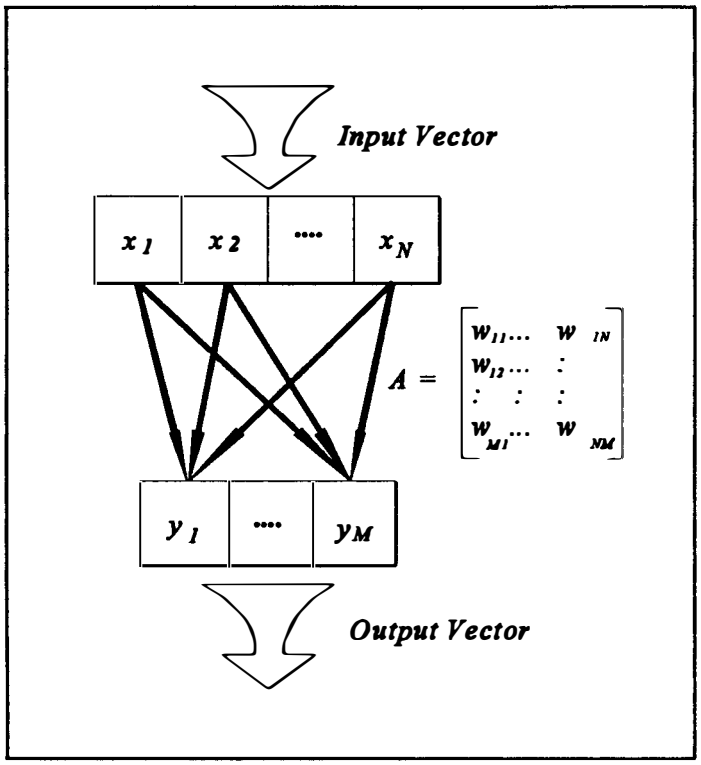

Figure 3.2. A single-layer neural network where each output neuron takes the linear sum of its input as its current level of activity.

The network learns using the algorithm shown in (3.4), where the gain parameter, γ, describes the rate of change in the synaptic weights over discrete time t. Each output neuron updates it state serially, with the available *input energy* reduced in proportion to the activity of previously updated neurons.

The summation term in (3.4) provides a feedback mechanism; changing the efficacy of synaptic transmission and normalising weight changes such that the squared value of the synaptic matrix is bounded with probability equal to one.

Sanger (1989) proves that the synaptic matrix will converge upon the *eigenvectors* of the input data in *eigenvalue* order. When the input data is masked by the synaptic matrix the results are very similar to the KLT, eigenvector transformation, *principal components analysis* (PCA), or the *Hotelling* transformation. As the output vector is ordered by decreasing variance it is a simple matter to specify a suitable level of compression, i.e. only coefficients (output of the transformation) deemed to be significant are retained.

For the implementation a block sixe of 8x8 pixels was used as input to the network with 8 neurons representing the output of the transformation. During training the learning rate, γ, was initialised to a value of 0.01 and decayed towards zero at a rate inversely proportional to the number of presentations made (Sanger, 1989). The 256 input vectors provided by each input image were transformed into a single 2048 element vector containing the block quantised coefficients. This transformation represents a gross compression ratio of 13.5 with a reconstruction error less than 10% (26 shades of grey). The results of training and using this approach are described in Phillips (1993), while Sanger provides more general and more extensive results. The algorithm is derived from the work of Oja (1982) and the approach is discussed in Oja (1991).

3.2.4 Semantic Encoding

Adaptive Resonance Theory (ART) was formally identified by Grossberg (1976/1991b), although Grossberg began deriving computational models for pattern recognition much earlier (Grossberg, 1970). Adaptive Resonance Theory is a mature technology, and deserves a more extensive review than presented here.

Justification for the use of an ART network, in preference to other self-organising neural networks, is provided by Grossberg (1987/1988b) where it is shown that ART is a encompasses other models of competitive learning and that it has properties not available with any other neural network models. Unfortunately, it is difficult to obtain descriptions of the application of ART networks. Carpenter and Grossberg (1987/1991b) indicate that ART has been used for the recognition of range sensor data but no real results are presented. It should also be noted that hardware implementations of ART are patented to Grossberg and, as ART is a complex device best simulated in hardware, may restrict the commercial development of applications using the technology. There are indications that ART has been used extensively in computer music applications, however it was well outside the scope of this research to examine these sources.

ART was described to solve the problems of the *stability-plasticity* dilemma and of *temporally unstable learning*[2] (Grossberg, 1987/1988b). Grossberg describes the problem with a set of questions (Grossberg, 1987/1988b, p.6);

> How can a learning system be designed to remain plastic in response to significant new events, yet also remain stable in response to irrelevant events? How does the system know how to switch between its stable and its plastic modes in order to prevent the relentless degradation of its learned codes by the "blooming buzzing confusion" of irrelevant experience? How can it do so without using a teacher?

Grossberg developed the ART *circuit*, shown of **Figure 3.3**, in response to these questions. The theory describes two recurrent on-centre off-surround networks - two *competitive learning* networks - embedded within a control structure that stabilises the network against recoding by irrelevant patterns. The architecture allows Grossberg to argue for a unifying principle in sensory processing, that ART can be seen "as a general organizational principle *in vivo*" (Grossberg, 1976/1991b, p.7). Of more pragmatic importance is the fact that ART "places no orthogonality or linear predictability constraints" on learnable patterns (Carpenter and Grossberg, 1987/1991a, p.3).

---

[2] Carpenter and Grossberg (1987/1991a) show that presentation of four input patterns can destabilise competitive learning algorithms.

The theory makes no assumption about the environment other than that there *will* be critical features available (Grossberg, 1987/1988b).

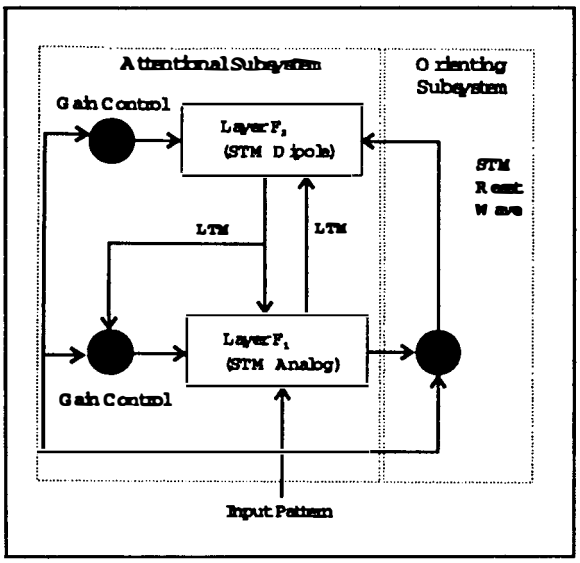

Figure 3.3. General architecture of an ART network showing network layers and control systems. Derived from Carpenter and Grossberg (1987/1991b).

The two layers in an ART network encode short term memory (STM) within their patterns of activation. Long term memory (LTM) traces are held by the synaptic pathways between the two layers, with the two layers fully connected. *Non-specific arousal* from a *gain control nucleus* allows the F1 layer to distinguish between top-down (template) and bottom-up (sensory) patterns, and primes F2 to respond ***supraliminally*** to input signals. The orienting subsystem generates a reset wave to F2 in response to pattern mismatches at F1. A reset wave inhibits active F2 neurons until the current input is removed.

Recognition, see **Figure 3.4**, is comprised of four stages:

A. The input pattern causes an STM pattern in F1 and arouses both the orienting subsystem and the gain control nucleus, called the *attentional gain control*, which in turn sends a non-specific arousal signal to F1. The STM pattern at F1 causes an inhibitory signal to be sent to the orienting subsystem cancelling the effect of the excitatory signal from the input pattern, therefore no reset wave is sent to F2. The STM pattern at F1 is *gated* via the LTM pathway before it arrives at F2.

B. The current STM at F2 is gated via the top-down pathway to F1, with an inhibitory signal to the gain control nucleus.

C. If a mismatch occurs between the top-down signal from F2 and the bottom-up input at F1 then the inhibitory signal to the orienting subsystem will be insufficient to prevent a reset wave being generated. The reset wave inhibits the currently active pattern at F2, therefore removing the inhibitory signal to the gain control nucleus.

D. The original STM pattern at F1 is reinstated and a new processing cycle begins. As the mismatched F2 pattern is enduringly inhibited, it can no longer respond to the F1 signal.

Searching continues until an adequate match is found, or when no further stored patterns can be found. A discussion of the computational properties resulting from the ART architecture, and a critical comparison with other network models, can be found in Grossberg (1987/1988b).

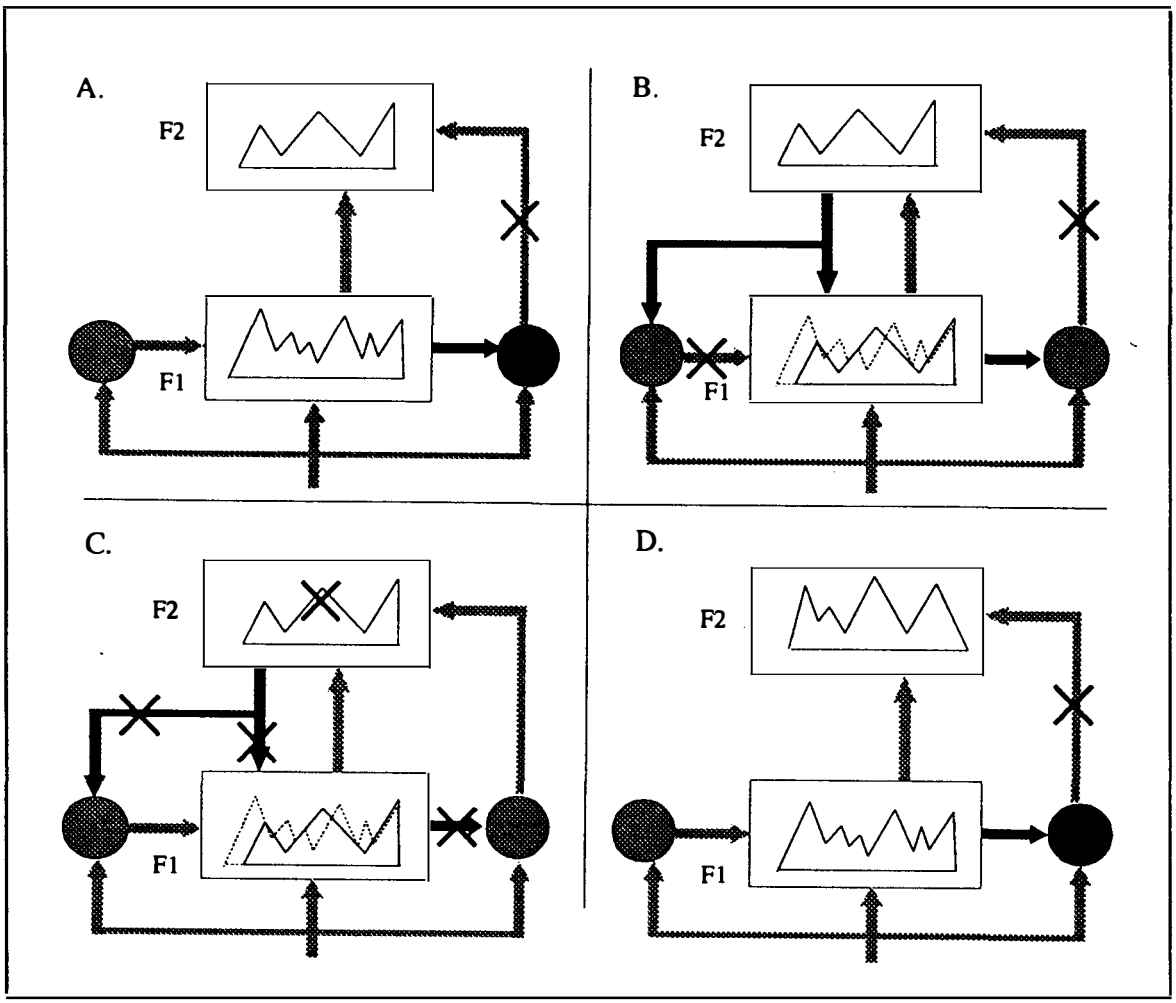

Figure 3.4. The recognition process in an ART network. Inhibited gain control and inhibitory signals are dark, excited gain control and excitatory signals are light. Diagram derived from Carpenter and Grossberg (1987/1991b).

The ART network used for the implementation is called ART2, a variation of the original ART network, designed to encode arbitrary binary and *analog* patterns. To accommodate analog patterns additional processing is required within the F1 layer, as shown in **Figure 3.3**. A complete description of the design goals and performance features of the network architecture can be found in Carpenter and Grossberg (1987/1991b).

STM patterns in the F1 layer. The behaviour of an ART network is defined by a set of self-regulating differential equations. For the ART2 network the *activation state* of any neuron in the F1 layer depends upon the current target node (see **Figure 3.3**). For $i = 1...M$ equations (3.5) - (3.10) describe the STM activities (activation states) computed for F1 (Carpenter and Grossberg, 1987/1991b):

$$w_i = I_i + A u_i \tag{3.5}$$

$$x_i = \frac{w_i}{E + \|w\|} \tag{3.6}$$

$$u_i = \frac{v_i}{E + \|v\|} \tag{3.7}$$

$$v_i = f(x_i) + B f(q_i) \tag{3.8}$$

$$p_i = u_i + \sum_j g(y_j) z_{ji} \tag{3.9}$$

$$q_i = \frac{p_i}{E + \|p\|} \tag{3.10}$$

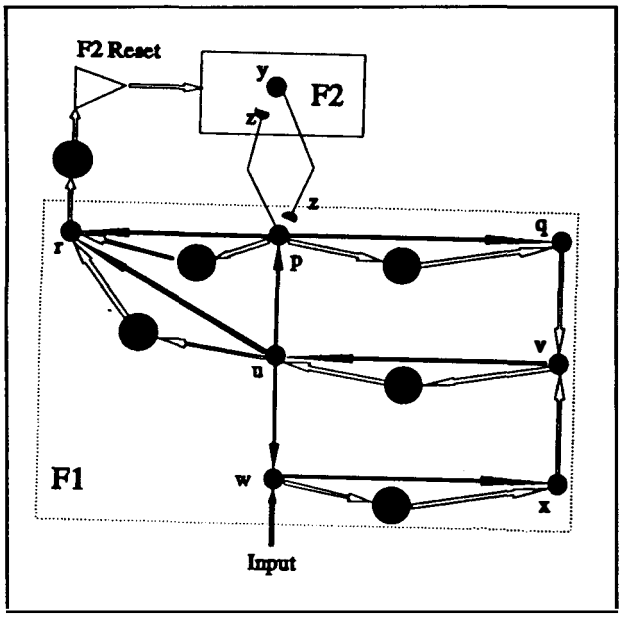

Figure 3.5. ART2 architecture (Carpenter and Grossberg, 1987/1991b). Dark arrows are specific inputs, open arrows are non-specific. Large circles are gain control nuclei, small ones are target nodes. Subscripts are not shown, see text.

where $\|p\|$ is the L_2-norm (magnitude) of vector p, z_{ji} is the synaptic weight between the *jth* neuron in F2 and the *ith* neuron in F1, I_i is the *ith* element in the input vector, and *A,B,C,D,E* are user-specified control parameters. The function f in equation (3.8) is usually chosen from a continuously differentiable function or piecewise linear function. The continuous function shown in (3.11) was used for the implementation.

$$f(x) = \begin{cases} \dfrac{2\theta x^2}{(x^2 + \theta^2)} & \text{if } 0 \le x < \theta \\ x & \text{if } x \ge \theta \end{cases} \tag{3.11}$$

The function g in equation (3.9) is given by (3.12). Since $g(y)$ evaluates to either 0 or D equation (3.12) reduces to the form shown in (3.13). See Carpenter and Grossberg (1987b/1991b) for the computational properties that result from these equations.

$$g(y_j) = \begin{cases} D & \text{if } j^{th} \text{ F2 neuron is active} \\ 0 & \text{otherwise} \end{cases} \tag{3.12}$$

$$p_i = u_i + \sum_j D z_{ji} \tag{3.13}$$

STM patterns in the F2 layer. F2 performs *contrast enhancement* via a competitive process, i.e. F2 "makes a choice when the node [neuron] receiving the largest total input quenches activity in all other nodes" (Carpenter and Grossberg, 1987/1991b, p.15). This process occurs for all neurons, except those inhibited because of a reset event, calculating the *dot product* between the vector arriving from the p target node and the LTM pathway. This is shown in (3.14) and (3.15) (Carpenter and Grossberg, 1987/1991b, p.15), i.e. the *jth* neuron in F2 is selected when it is maximally active.

$$d_j = \sum_i p_i z_{ij} \tag{3.14}$$

$$d_j = \max(d_j : j = M + 1 \ldots N) \tag{3.15}$$

Mismatch and reset conditions. The *r* target node (see **Figure 3.3**) calculates the degree of match between the top-down pattern and the bottom-up input, this is shown in (3.16). Once *r* has been calculated, the orienting subsystem will generate a reset wave if the condition shown in (3.17), where ρ - the vigilance parameter - is set between 0 and 1, is not satisfied.

$$r_i = \frac{u_i + Cp_i}{E + |u| + |Cp|} \tag{3.16}$$

$$\frac{\rho}{E + |r|} > 1 \tag{3.17}$$

Changing the LTM weights. An ART network only learns in a resonant state, i.e. when a top-down pattern adequately represents the current input pattern. Once the appropriate *jth* F2 neuron is selected, the top-down and bottom-up synaptic pathways are updated according to (3.18)

and (3.19), with $0 < D < 1$. This results in the weights incrementally moving towards their asymptotic values, i.e. they will converge to the average of the exemplar patterns. An alternative called ***fast learning*** will move the weights to their asymptotic values when a resonant state is achieved.

$$\frac{d}{dt} z_{ji} = (1 - D)\left[\frac{u_i}{1 - D} - z_{ji}\right] \tag{3.18}$$

$$\frac{d}{dt} z_{ij} = D[p_i - z_{ij}] \tag{3.19}$$

3.2.5 Association

The output of an ART network is a single index, indicating the F2 node that is maximally active to an input pattern. This index can be used to provide access to an external database. The implementation used this index to access a simple linear database containing syntactic information.

3.2.6 Environment

A mixture of special purpose hardware and software simulation was used for the implementation. The median filter and the calculation of the reflectance pattern, used for sensory processing, was implemented in ANSI C and included code to perform conversion between the different image formats used. The KLT network used for feature selection was also constructed in ANSI C using a variation of a feed-forward network simulation. All training of the network was *off-line* due to the nature of the learning algorithm.

An HNC Anza Plus neurocomputer was used to simulate the ART network with ANSI C code providing communication with a flat database of association information. Information on the neurocomputer and the details how ART is implemented can be found in the HNC Techical reference manual (HNC Neurosoftware Manual, 1989).

3.3 Summary

This section describes a model of visual object recognition. It is noted that the three stages implemented, sensory processing, feature extraction, and semantic encoding, form a conventional pattern recognition hierarchy (Simon, 1989). Although only a partial implementation has been described it formed a suitable environment for the application described in the next section - the recognition of frontal images of human faces.

Section 4: Prototype Application

To establish the validity of the model and the implementation described in previous sections it is necessary to examine recognition performance for a specific application. This section:

- Derives hypotheses to validate the thesis of this research.

4.1 Background

It has been identified that certain visual stimuli may be difficult to describe using simple, geometric structural descriptions. Human faces are examples of natural objects that have this property and it is known that they are recognised on the basis of *configurations* of features. It has been identified that "the most powerful means of recognition is by perceiving, storing and retrieving aspects of facial configuration" (Clifford and Bull, 1978, p.71). Recognition performance of individuals using facial information alone is approximately eighty percent (Clifford and Bull, 1978).

The upper areas of the face seem the most useful for identification, although recognition can certainly occur using information from below the eyes. It is known that two areas are important for correct identification (Clifford and Bull, 1978; Bruce and Green, 1990);

1. The eyes and their immediate context.
2. The mouth and its expression.

It has been suggested that the eyes and the mouth convey more information about expression, which is important for social interaction. This suggestion has some correlation with lesion data indicating that face agnosic patients can identify faces when able to observe facial expressions evolving over time (Clifford and Bull, 1978; Bruce and Green, 1990; McCarthy and Warrington, 1990).

4.2 Experimental Hypotheses

4.2.1 Definitions

A *face* represents the unique identification for an individual, and the terms can be considered interchangeable in the context of this research. It is assumed that a face can be represented using a collection of discretely valued features in a high dimensional feature space. The implementation uses a 2048 element feature vector (representing the output of the KLTNet - see 3.2.3).

It is assumed that individual instances, or *images*, of a face form a *cluster* of feature vectors. The central point of this cluster is taken to represent an identity (perceptual category). *Recognition* means selecting the appropriate category and, for an unknown face, the allocation of a new perceptual category.

4.2.2 Experimental Criteria and Methodology

The study was conducted as three separate experiments. Although these experiments are not exhaustive they do demonstrate that the research can be used to develop real applications. Three aims were pursued;

1. To quantify recognition performance for the implementation.
2. To quantify recognition performance when images are partially obscured.
3. To quantify recognition performance when images are transformed due to changes in facial expression.

In all experiments the testing sample was applied for vigilance values (see 3.2.4) of 0.96, 0.97, 0.98, and 0.99; for each vigilance value the rate of correct classification for the entire test sample was recorded. All other parameters were set to the values described in the previous section.

4.2.3 Training

Ten images were sampled from ten different faces to form the training set. All images were centred and frontal. The KLTNet was trained upon a single image representative of the training set. Training for the ART network simply required the presentation of the training set to establish perceptual categories.

4.3 Experiment One

The misclassification rate of the application was determined for novel images of known, and unknown, faces. Ten test images, involving minor facial changes, were sampled. To determine the ability of the network to correctly identify unknown faces only half the training images were presented to the ART network. The rate of misclassification upon presentation of the entire test set was then recorded.

4.3.1 Results

The application recognised all known images, and identified all unknown images as unknown, when the ART network was set to a vigilance level of 0.96. Misclassification of unknown faces occurs until the vigilance is set to 0.96, and continues when the vigilance is set lower than this. Errors for known faces consisted of identifying the images as unknown. The average RMS error between the images in the training set and those in the test set was approximately 27.3 shades of grey, i.e. an average 10.7% difference.

Table 4.1. *Results for Experiment One (Known and Unknown Faces).*

Vigilance	Identify Known / 5	Identify Unknown / 5
0.99	0	4
0.98	3	4
0.97	5	4
0.96	**5**	**5**
0.95	5	4

4.3.2 Discussion

The image from the test set, misclassified when the vigilance was set to 0.95, is shown below, see Figure 4.1, next to the image from the training set that formed the category to which the test image was assigned. Both subjective and objective measures of difference indicate that the images are very different, with an RMS error of approximately 40.3 shades of grey, or an average 15.8% difference. It is interesting that the two images are of females, although this is probably a coincidence.

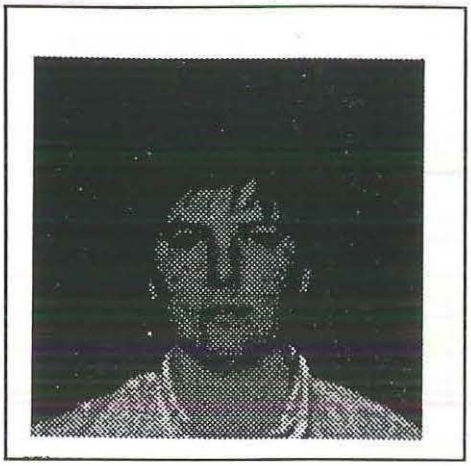

Figure 4.1. Image misclassified in experiment one.

Figure 4.2. Training image that formed category resulting in the mismatch of Figure 4.1 in experiment one.

4.4 Experiment Two

It was hypothesised that the application would make fewer classification errors when the mouth area of an image was masked compared to the number of errors made when the eyes were masked. The ART network was trained by presenting all ten training images (with no masking applied). Classification rates were then collected for when the same images were presented with the mouth area masked, and for when the eye area was masked.

4.4.1 Results

The application recognised all test images with masked eye areas at a vigilance level of 0.95, and at a vigilance level of 0.96 for images with masked mouth areas. Differences in performance between the two data sets are probably insignificant.[3]

[3] The size of the data set used in these experiments is quite small; thus, statistical generalisation is difficult.

Table 4.2. *Results for Experiment Two (masking mouth and eyes).*

Vigilance	Eyes Masked	Mouth Masked
	Identified	Identified
0.99	2	3
0.98	8	8
0.97	8	9
0.96	9	10
0.95	10	10

The average RMS error between images with masked eyes and the original images was approximately 23.9 shades of grey, or an average 9.4% difference. For images with masked mouths the average RMS error was approximately 18.9 shades of grey, or a 7.4% difference. The difference in error levels was probably caused by the predominance of bearded men in the images. Note that the results for images with masked mouths are comparable to the results of experiment one, where subjects were asked to ***smile gently*** for the test images.

4.4.2 Discussion

These results are inconclusive and do not demonstrate that the implementation is making greater use of the eye area. More extensive masking experiments failed to isolate the area being used for recognition. Further experimentation is required to determine whether the implementation uses an area common to all faces for identification, the combination of a number of areas, or an area determined by the features present in an individual face.

4.5 *Experiment Three*

It was hypothesised that the application would be able to recognise a face despite changes in expression between training image and test image. The implementation was trained upon the complete training set and a further ten images were sampled for the test set, with each test image representing a significant change in facial expression. Misclassification of the test set was recorded.

4.5.1 Results

The application achieved 100% recognition performance with a vigilance level of 0.95, comparable to the eye masking results of experiment two. Subjects were asked to *pull a face* for the test images, involving changes to the eyes and the mouth and, possibly, more significant structural change.

Table 4.3. *Results for Experiment Three (change in expression).*

Vigilance	Identified
0.99	0
0.98	1
0.97	5
0.96	9
0.95	10

The average RMS error between the test images and the training images was approximately 41.2 shades of grey, an average 20% difference. The RMS error for the image most difficult to classify, see Figure 4.3, was approximately 76.8 shades of grey, an 30% difference. The RMS error for the image classified at a vigilance level of 0.98, i.e. easiest to classify, see Figure 4.4, was approximately 27.1 shades of grey, a 10.6% difference.

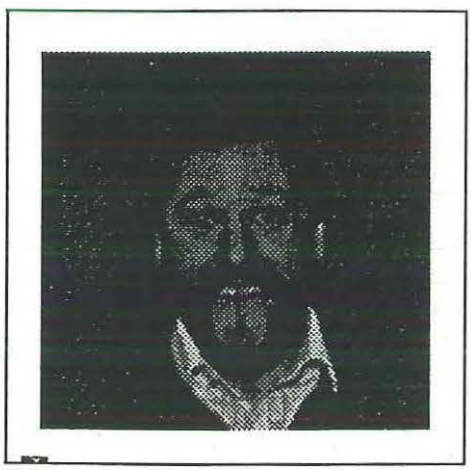

Figure 4.3. Image most difficult to classify in experiment three. Note that there are major changes to mouth and eye areas.

Figure 4.4. Image easiest to classify in experiment three. Note that although there are major changes to both the eye and mouth areas, there is probably little structural change.

4.5.2 Discussion

These results are surprising, with little variation between this task and the previous experiments. Increased difficultly seems to be represented by poorer performance at higher vigilance levels, yet reaching 100% performance at a level only slightly lower than that recorded for experiment one, and at the same level as that found for eye masking in experiment two.

4.6 Summary

To validate the implementation, and hence parts of the model, a number of simple experiments have been performed. The results of these experiments confirm that the implementation can perform as required, yet they do not support some of the assumptions made in regard to the performance of the ART network. These issues are discussed in the next section.

Section 5: Discussion

This section generalises the experimental results presented in the previous section, and discusses some observations of the performance of the implementation. A summary of the results of the experimental hypotheses is also presented.

5.1 Observations and Problems

5.1.1 Vigilance

Performance is sensitive to the vigilance level, with small changes in the vigilance resulting in large changes in performance. The results suggest that vigilance approaches a critical value between 0.96 and 0.95 - indicating that the values used for the experiments may have been too coarse. At these values the ART network is able to correctly assign categories to most, if not all, of the test data. It must be determined whether this value remains constant when the network is trained upon a larger number of images.

5.1.2 Recognition Time

Successful recognition was rapid, approaching real-time performance (search times in the order of one-two seconds). In contrast, images that were not recognised required an extended search by the ART network resulting in significantly slower processing (search times in the order of minutes). ART retains linear growth in search times for known categories, but the poor performance when recognition was not successful has serious scaling implications. Setting the vigilance to an initially high level and gradually reducing it until recognition occurs would seem to be a sensible approach, however, the problem with search times for unknown images would make this a tedious and time-consuming activity.

5.1.3 Scaling

The experiments were performed using a 2048 element feature vector. This size was chosen because of the potential of reconstructing the images using the KLTNet. It is possible that a smaller feature vector would suffice, resulting in a smaller ART network and a number of benefits.

Reducing the size of the feature vector would alleviate problems with the connectivity of the ART network. Because the two layers of the network are fully connected, in both directions, making another category available at the F2 layer currently requires an additional 4096 connections. Unfortunately, this level of connectivity is inherent to ART networks and cannot be totally avoided.

This change is likely to lead to faster searching by ART, possibly alleviating some of the problems associated with extended search times for unknown categories. However, it is also possible that ART may require a very different set of vigilance values to perform successfully.

5.1.4 Measuring Recognition Difficulty

The differences between training and test images are presented above in terms of RMS error, used because it proved useful when measuring the performance of the KLTNet. However, the results presented in the previous section suggest that RMS is a misleading measure of difficulty, especially in the case of experiment three. This may be due to RMS being a measure of *average* difference, while the both the KLTNet and ART network use *local* differences during processing. In this case, peak RMS may have proved to be a more appropriate measure.

5.2 Experimental Hypotheses

Recognition performance for minor transformations is reliable. However, pursuing 100% performance may lead to misclassification errors. The vigilance parameter can be tuned to include or reject misclassification - this is obviously a domain dependant decision. It appears that the range of vigilance values used in the experiments may have been too coarse. Further experimentation is required to a) determine a more suitable range of values, b) whether the range of values remains constant as the number of categories is increased, and c) whether the range of values remains constant when the implementation is applied to a different data set.

The differences in recognition performance for eye and mouth masking, as shown in experiment two, do not seem significant. The immediate conclusion is that the implementation does not rely upon features isolated in either the eye or mouth area. More extensive experimentation (results not presented) failed to determine whether a particular area was used and further experimentation is

required to determine if a particular set of features is required for reliable recognition. Although the experiments failed to locate the features being used in recognition they demonstrate that the implementation is able to perform recognition despite minor occlusion of the images (approximately 20% occlusion).

The results of experiment three show that recognition can occur despite significant changes in the training and the test images. The RMS measures would suggest that this task is more difficult than that performed in experiment two, however, performance is only affected when using higher vigilance values. This implies that the changes in expression found in the test set did not significantly alter the features being used by the ART network. This reinforces the need to determine the set of features being used for recognition, and to determine the amount of transformation that can be tolerated before significant performance degradation occurs.

5.3 Summary

The application can recognise novel images of previously categorised faces, and identify faces that are unknown. Recognition is reliable and robust, with levels of performance directly related to the vigilance setting. The application continues to perform well when images are transformed, although the limits of this performance have yet to be learned. Masking has only a minor influence upon performance and it appears that the application does not depend solely upon the eye or mouth area. The wider implications of these results are discussed in the next section.

Section 6: Conclusions

This section concludes the thesis. It contains a summary of the contents, generalises the results of the experiments and the theoretical development, identifies some limitations of the model, and suggests directions for future research.

6.1 Summary of the Thesis

Section two of this thesis introduces the theory that has influenced this research, including a review of relevant physiological and psychological information. The major computational approaches, structural decomposition and invariant feature detection, are described and an attempt is made to resolve their conceptual differences. Section two also contains a brief discussion of neural networks, and their role in modelling biological perceptual systems.

The review of section two provides the background for the computational model developed in section three. Although this model is incomplete, requiring further study and development, it forms a suitable framework for the partial implementation described. The two neural networks used for the implementation are introduced and their major features are identified.

Section four discusses some aspects of human face recognition, and describes the methodology used to stage three simple experiments. Results of these experiments are presented in section five, as is a discussion of their immediate implications and the results of the experimental hypotheses.

6.2 Generalisation of Results

The implementation reliably classifies frontal images of human faces. Classification performance is related to the parameter settings used to control the implementation. Performance degrades gracefully as the classification task becomes more difficult, although quantifying task difficulty has not been satisfactorily achieved. It is suggested that peak and average RMS error may more suitably measure the task difficulty.

The computational model provides a theoretical foundation for the implementation, and it has been partially validated by the experimental results. Complete validation, which was beyond the scope of this research, would require more extensive implementation and testing. The model also allows for fundamental issues not directly addressed in this thesis, for example, rotation and translation invariance.

6.2.1 Limitations of the Implementation

Rotation invariance can be handled by top-down control signals indicating that a rotation transformation should be performed upon the input data. Although this is not a conceptually satisfying solution, i.e. the process is cumbersome and probably requires global knowledge, it provides the necessary mechanisms to handle rotated input data. A more suitable approach may be provided by a mechanism that detects an object's centre of gravity and performs a rotation until this is in the appropriate position. It is interesting that humans normally perform poorly with complicated objects that have been rotated, although the skill can be learned (Rock, 1986).

Translation invariance is assumed to be handled by a separate mechanism that coordinates fixation and focusing tasks. A related problem, that of figure-ground separation, would be partially solved by such a mechanism.

Size, or scale, invariance has not been addressed and a solution has yet to be found. Hubel (1988) suggests that biological vision systems solve this problem by converging and diverging neural connections in visual cortex, such a process resulting in receptive fields invariant to size. This is an important problem that requires further research.

Object occlusion, which received only a cursory examination in the experiments, is a difficult problem requiring the ability to derive, and retain, geometric structure. Although no such process was performed for the implementation - it is suggested that this process is part of spatial analysis - the ART network can perform recognition despite minor occlusions. Further research is required to find performance limits of the implementation and to investigate more thorough solutions to the problem.

6.2.2 Summary of Hypothesis

Although not completely validated, it does appear that the hypothesis is appropriate. The use of a suitable system model, multiple neural networks with clearly defined tasks, and conventional algorithms where necessary, has resulted in a successful implementation. The results of the experiments are positive and suggest that these techniques deserve further examination.

6.3 Future Research

Any study of object recognition, and vision generally, raises more problems than it solves and the scope for future research is unlimited. Besides the limitations described above and the issues discussed in the previous section, further research is required into the following;

- Structural processing for spatial vision
- Selection of feature detection operations for pattern vision
- Controlling category searches
- Real-time implementation
- Improving accuracy and detail of the model

6.3.1 Implementing Structural Processing

The implementation would be greatly improved by providing the means to extract structural descriptions of the objects in an image. This would allow the performance of object extraction, object description, and provide guidance for object transformation operations, for example, scaling and rotating.

6.3.2 Selection of Feature Extraction Operations

The KLTNet, see 3.2.3, is not sufficiently powerful for a truly comprehensive implementation. The transformation learnt by the network requires that the statistics of the input data do not change over time, i.e. the KLTNet learns a static statistical transformation. Although not examined in this thesis it is likely that the technique would not generalise to images captured under different environmental conditions. Habibi (1986) dismisses the transformation learnt by the KLTNet as obsolete, suggesting that there are more appropriate techniques, such as the *discrete cosine* transformation. As such techniques are adaptive they should generalise to all input data.

Unfortunately, the KLTNet is not the only neural network that suffers the above problem. It is demonstrated by many of the network types trained to perform encoding, including the back-propagation networks of Cottrell and Fleming (1990), the Boltzmann Machine (Hinton, Sejnowski, and Ackley, 1984), and self-organising maps (Ritter, Martinez, and Schulten, 1991).

6.3.3 Controlling Category Searching

As identified in 5.1.2, the search time for the ART network is high when a category cannot be found. A complete implementation would require a self-adjusting vigilance level, beginning with a high value and decreasing automatically as the category search is expanded. This means that, unless the category was known, identification would be a lengthy process with the only information gained being that the category is not known. The only solution that suggests itself is to use something other than an ART network.

6.3.4 Real-Time Implementation

Although not implemented as such, the techniques and the model are designed to operate in real-time (after training of the KLTNet). It is likely that a real-time implementation would suggest many issues that need to be addressed.

6.3.5 Improving the Model

The discussion of biological vision in section two is highly simplified. This simplified information has not been captured by the model, for example, the issue of colour information has been ignored. The model needs refinement and completion, and a more extensive review of physiological and psychological data undertaken. It was assumed for this thesis that the implementation would be required to recognise static images, which has some biological plausibility (see 3.1), however, it is unknown how the model would cope with dynamic images and how the model fits into a machine vision architecture.

6.4 Summary

The research has been successful, although many questions remain to be answered. The central thrust of the thesis, that a model of object recognition can be successfully developed using artificial neural networks, has been satisfactorily verified. The results of this study will be used for future research into object recognition and they demonstrate the advantages of developing a systematic framework for machine interpretation. Although the experimental data is limited, it is suggested that the approach taken in this study has resulted in an implementation superior to what may have been achieved with a single, isolated technique.

Section 7: References

Abramov, I, & Gordon, J. (1973). Vision. In E.C.Carterette and M.P.Friedman (Eds.), *Handbook of perception, Volume III: Biology of perceptual systems* (pp.327-357). New York: Academic Press.

Azencott, R., Doutriaux, A., & Younes, L. (1990). Synchronous Boltzmann machines and outline based image classification. *International Neural Network Conference, July, 1990, Paris, France* (pp.7-10). Dordrecht: Kluwer Academic Publishers.

Baldi, P., & Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks, 2 (1989)*, 53-58.

Ballard, D.H., & Brown, C.M. (1982). *Computer Vision.* Englewood Cliffs, New Jersey: Prentice-Hall Inc.

Barrow, H.G., & Tenenbaum, J.M. (1986). Computational approaches to vision. In K.R.Boff, L.Kaufmanm, & J.P.Thomas (Eds.), *Handbook of perception and human performance, Volume 2: Cognitive processes and performance*, New York: John Wiley and Sons.

Batchelor, B.G. (1978). *Pattern recognition: Ideas in practice.* New York: Plenum Press.

Bruce, V., & Green, P.R. (1990). *Visual perception: Physiology, Psychology, and Ecology* (2nd Ed.). London: Lawrence Erlbaum Associates.

Carpenter, G.A. (1991). Neural network models for pattern recognition and associative memory. In G.A.Carpenter & S.Grossberg (Eds.), *Pattern recognition by self-organizing neural networks* (pp.2-33). Cambridge, Massachusetts: The MIT Press. (Reprinted from *Neural Networks, 1989 (2),* 243-257)

Carpenter, G.A., & Grossberg, S. (1991a). A massively parallel architecture for a self-organizing neural pattern recognition machine. In G.A.Carpenter & S.Grossberg (Eds.), *Pattern recognition by self-organizing neural networks* (pp.316-382). Cambridge, Massachusetts: The MIT Press. (Reprinted from *Computer Vision, Graphics, and Image Processing, 1987 (37),* 54-115)

Carpenter, G.A., & Grossberg, S. (1991b). ART2: Self-organization of stable category recognition codes for analog input patterns. In G.A.Carpenter & S.Grossberg (Eds.), *Pattern recognition by self-organizing neural networks* (pp.398-423). Cambridge, Massachusetts: The MIT Press. (Reprinted from *Applied Optics, 1987 (26),* 4919-4930)

Carpenter, R.H.S. (1984). *Neurophysiology.* London: Aspen Publishers.

Chase, W.G. (1986). Visual information processing. In K.R.Boff, L.Kaufmanm, & J.P.Thomas (Eds.), *Handbook of perception and human performance, Volume 2: Cognitive processes and performance*, New York: John Wiley and Sons.

Cottrell, G.W., & Fleming, M. (1990). Face recognition using unsupervised feature extraction. *International Neural Network Conference, July, 1990, Paris, France* (pp.322-325), Dordrecht: Kluwer Academic Publishers.

Damasio, A.R., Tranel, D., & Damasio, H. (1989). Disorders of visual recognition. In H.Goodglass & A.R.Damasio (Eds.), *Handbook of Neuropsychology, Vol. 2* (pp.317-332). Amsterdam: Elsevier Science Publishers B.V.

Desimone, R., & Ungerleider, L.G. (1989). Neural mechanisms of visual processing in monkeys. In H.Goodglass & A.R.Damasio (Eds.), *Handbook of Neuropsychology, Vol. 2* (pp.267-299). Amsterdam: Elsevier Science Publishers B.V.

Dodwell, P.C. (1970). *Visual pattern recognition.* New York: Holt, Rinehart, and Winston.

Feldman, J.A. (1985). Connectionist models and parallelisn in high-level vision. In A.Rosenfeld (Ed.). *Human and machine vision II* (pp.86-108).

Freeman, J.A., & Skapura, D.M. (1991). *Neural networks: Algorithms, applications, and programming techniques.* Reading, Massachusetts: Addison-Wesley Publishing.

Fu, K.S. (1980). Syntactic (Linguistic) pattern recognition. In K.S.Fu (Ed.), *Digital pattern recognition* (2nd Ed.)(pp.95-134). Berlin: Springer-Verlag.

Fu, K.S. (1986). Syntactic pattern recognition. In T.Y.Young and K.S.Fu (Eds.), *Handbook of pattern recognition and image processing* (pp.85-117). San Diego, California: Academic Press.

Gibson, J.J. (1966). *The senses considered as perceptual systems.* Boston: Houghton Mifflin Co.

Goldberg, M.E., & Colby, C.L. (1989). The neurophysiology of spatial vision. In H.Goodglass & A.R.Damasio (Eds.), *Handbook of Neuropsychology, Vol. 2* (pp.301-315). Amsterdam: Elsevier Science Publishers B.V.

Gonzalez, R.C, & Wintz, P. (1987). *Digital image processing* (2nd Ed.). Reading, Massachusetts: Addison-Wesley Publishing.

Grossberg, S. (1970). Neural pattern discrimination. *Journal of Theoretical Biology, 27, 1970,* 291-337.

Grossberg, S. (1987). The quantized geometry of visual space: The coherent computation of depth, form, and lightnes. In S.Grossberg (Ed.), *The adaptive brain II: Vision, speech, language, and motor control* (pp.3-79). Amsterdam: Elsevier Science Publishers B.V. (Reprinted from *The Behavioural and Brain Sciences, 6, 1983, 625-657*)

Grossberg, S. (1988a). How does a brain build a cognitive code? In J.A.Anderson & E.Rosenfeld (Eds.), *Neurocomputing: Foundations of research* (pp.349-399). Cambridge, Massachusetts: The MIT Press. (Reprinted from *Psychological Review 87 (1980),* 1-51)

Grossberg, S. (1988b). Competitive learning: From interactive activation to Adaptive Resonance. In S.Grossberg (Ed.), *Neural networks and natural intelligence* (pp.213-250). Cambridge, Massachusetts: The MIT Press. (Reprinted from *Cognitive Science, 1987 (11),* 23-63)

Grossberg, S. (1989). *Self-Organizing neural networks: Foundations and applications.* SEARCC-89 Tutorial, December 4, 1989.

Grossberg, S. (1990). The ART of adaptive pattern recognition by a self-organising neural network. In J.Diederich (Ed.), *Artificial neural networks: Concept learning* (pp.69-80). Los Alamitos, California: IEEE Computer Society Press. (Reprinted from *IEEE Computer, 21 (3), March 1988, 77-88*)

Grossberg, S. (1991a). Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors. In G.A.Carpenter & S.Grossberg (Eds.), *Pattern recognition by self-organizing neural networks* (pp.211-236). Cambridge, Massachusetts: The MIT Press. (Reprinted from *Biological Cybernetics, 1976 (23),* 121-134)

Grossberg, S. (1991b). Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, illusions. In G.A.Carpenter & S.Grossberg (Eds.), *Pattern recognition by self-organizing neural networks* (pp.283-311). Cambridge, Massachusetts: The MIT Press. (Reprinted from *Biological Cybernetics, 1976 (23),* 187-202)

Grossberg, S. (1991c). Nonlinear neural networks: Principles, mechanisms, and architectures. In G.A.Carpenter & S.Grossberg (Eds.), *Pattern recognition by self-organizing neural networks* (pp.36-109). Cambridge, Massachusetts: The MIT Press. (Reprinted from *Neural Networks, 1988 (1),* 17-61)

Habibi, A. (1986). Image coding. In T.Young and K.Fu (Eds.), *Handbook of pattern recognition and image processing* (pp.169-189). San Diego, CA: Academic Press.

Habibi, A., & Wintz, P. (1971). Image coding by linear transformations and block quantization. *IEEE Trans. Communication Technology, COM-19 (1)*, 50-62.

Hebb, D. (1949). *The Organization of Behaviour.* New York: John Wiley and Sons.

Hildreth, E.C., & Ullman, S. (1989). The computational study of vision. In M.I.Posner (Ed.), *Foundations of Cognitive Science* (pp.581-630). Cambridge, Massachusetts: The MIT Press.

Hinton, G.E., Sejnowski, T.J., & Ackley, D.H. (1984). *Boltzmann Machines: Constraint satisfaction networks that learn.* Pittsburgh, PA, Technical Report CMU-CS-84-119, Carnegie-Mellon University, Dept. of Computer Science.

HNC Neurosoftware Manual. Release 2.22. June, 1989. HNC Inc.

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science (U.S.), 79 (April),* 2554-2558.

Hood, D.C., & Finkelstein, M.A. (1986). Sensitivity to light. In K.R.Boff, L.Kaufman, & J.P.Thomas (Eds.), *Handbook of perception and human performance, Volume 1: Sensory processes and perception.* New York: John Wiley and Sons.

Huang, J.J.Y, & Schultheiss, P.M. (1963). Block quantization of correlated Gaussian random variables. *IEEE Trans. Communications Systems, CS-11*, 289-296.

Hubel, D.H. (1988). *Eye, brain, and vision.* New York: Scientific American Library.

Keller, J.M, & Qiu, H. (1988). Fuzzy set methods in pattern recognition. In J.Kittler (Ed.), *Pattern recognition, Proceedings of the 4th International Conference, 1988* (pp.173-182). Berlin: Springer-Verlag.

Kohonen, T. (1982). Self-Organized formation of topologically correct feature maps. *Biological Cybernetics, 43, 1982,* 59-69.

Kohonen, T. (1988). An introduction to neural computing. *Neural Networks, 1, 1988,* 3-16.

Kollias, S., Tirakis, A., & Milios, T. (1991). An efficient approach to invariant recognition of images using higher-order neural networks. In T.Kohonen, K.Mäkisara, O.Simula, and J.Kangas (Eds.), *Artificial neural networks* (pp.87-92). Amsterdam: Elsevier Science Publishers B.V.

Linsker, R. (1988). Self-organisation in a perceptual network. *Computer, March (1988)*, 105-117.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York: W.H.Freeman and Co.

McCarthy, R.A., & Warrington, E.K. (1990). *Cognitive Neuropsychology: A clinical introduction.* San Diego, California: Academic Press.

McCulloch, W.S., & Pitts, W. (1988). A logical calculus of the ideas immanent in nervous activity. In J.A.Anderson and E.Rosenfeld (Eds.), *Neurocomputing: Foundations of research* (pp.18-27). Cambridge, Massachusetts: The MIT Press. (Reprinted from *Bulletin of Mathematical Biophysics 5, 1943,* 115-133)

Minsky, M.L., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry* (2nd ed.). Cambridge, Massachusetts: MIT Press.

Mrsic-Flogel (1991). Approaching cognitive system design. In T.Kohonen, K.Mäkisara, O.Simula, and J.Kangas (Eds.). *Artificial neural networks* (pp.879-883). Amsterdam: Elsevier Science Publishers B.V.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology, 15 (1982)*, 267-273.

Oja, E. (1991). Data compression, feature extraction, and autoassociation in feedforward neural networks. In T.Kohonen, K.Mäkisara, O.Simula, and J.Kangas (Eds.), *Artificial neural networks* (pp.737-745). Amsterdam: Elsevier Science Publishers B.V.

Overington, I. (1992). *Computer vision: A unified, biologically-inspired approach.* Amsterdam: Elsevier Science Publishers B.V.

Pao, Y. (1989). *Adaptive pattern recognition and neural networks.* Reading, Massachusetts: Addison-Wesley.

Peterson, C., & Anderson, J.R. (1987). A mean field theory learning algorithm for neural networks. *Complex Systems, 1 (1987)*, 995-1019.

Phillips, V. (1993). *Feature extraction and image compression using a feed-forward neural network.* Perth, WA, Technical Report 1/93, Edith Cowan University, Department of Computer Science.

Pitts, W., & McCulloch, W.S. (1988). How we know universals: The perception of auditory and visual forms. In J.A.Anderson and E.Rosenfeld (Eds.), *Neurocomputing: Foundations of research* (pp.32-41). Cambridge, Massachusetts: The MIT Press. (Reprinted from *Bulletin of Mathematical Biophysics 9, 1947,* 127-147)

Pomerantz, J.R., & Kubovy, M. (1986). Theoretical approaches to perceptual organisation. In K.R.Boff, L.Kaufmanm, & J.P.Thomas (Eds.), *Handbook of perception and human performance, Volume 2: Cognitive processes and performance.* New York: John Wiley and Sons.

Pylyshyn, Z.W. (1989). Computing in Cognitive Science. In M.I.Posner (Ed.), *Foundations of Cognitive Science* (pp.52-91). Cambridge, Massachusetts: The MIT Press.

Ritter, H., Martinez, T., & Schulten, K. (1991). *Neural computation and Self-organising Maps,* Reading, Massachusetts: Addison-Wesley.

Rock, I. (10986). . In K.R.Boff, L.Kaufmanm, & J.P.Thomas (Eds.), *Handbook of perception and human performance, Volume 2: Cognitive processes and performance*, New York: John Wiley and Sons.

Rosenfeld, A. (1986). Computer vision. In T.Y.Young and K.S.Fu (Eds.), *Handbook of pattern recognition and image processing* (pp.355-368). San Diego, California: Academic Press.

Rosenfeld, A., & Weszka, J.S. (1980). Picture recognition. In K.S.Fu (Ed.), *Digital pattern recognition* (2nd Ed.)(pp.135-166). Berlin: Springer-Verlag.

Rueff, M. (1989). Scale space filtering and the scaling regions of fractals. In J.C.Simon (Ed.), *From pixels to features* (pp.49-60). Amsterdam: Elsevier Science Publishers B.V.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. In D.E.Rumelhart, & J.L.McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp.318-362). Cambridge, Massachusetts: The MIT Press.

Rumelhart, D.E., & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science, 1985, 9,* 75-112.

Rybak, I.A., Shevtsova, N.A., & Sandler, V.M. (1992). The model of a neural network visual preprocessor. *Neurocomputing 4 (1992)*, 93-102.

Sabbah, D. (1988). Computing with connections in visual recognition of origami objects. In D.Waltz & J.A.Feldman (Eds.), *Connectionist models and their implications: Readings from cognitive science*. Norwood, NJ: Arlex Publishing.

Sanger, T.D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks, 2 (1989)*, 459-473.

Simon, J.C. (1989). A complemental approach to feature detection. In J.C.Simon (Ed.), *From pixels to features* (pp.229-236). Amsterdam: Elsevier Science Publishers B.V.

Sutherland, N.S. (1973). Object recognition. In E.C.Carterette and M.P.Friedman (Eds.), *Handbook of perception, Vol.III: Biology of perceptual systems* (pp.157-185). New York: Academic Press.

Tasto, M., & Wintz, P. (1971). Image coding by adaptive block quantization. *IEEE Trans. Communication Technology, COM-19 (6)*, 957-972.

Uhr, L.M. (1987). Highly parallel, hierarchical, recognition cone perceptual systems. In L.M.Uhr (Ed.), *Parallel computer vision* (pp.249-292). Orlando, Florida: Academic Press.

Weiss, S.M., & Kapouleas, I. (1992). An empirical comparison of pattern recognition, neural nets, and machine learning classification methods. In P.Mehra & B.W.Wah (Eds.), *Artificial neural networks: Concept and theory* (pp.646-652). Los Alamitos, CA: IEEE Computer Science Press Tutorial. (Reprinted from *Proceedings of the International Joint Conference on Artificial Intelligence, 1989, 781-787*)