Image Categorisation Using Parallel Network Constructs :

An Emulation of Early Human Colour Processing and Context Evaluation

C.Robertson

Department of Computing Science
University of Newcastle–upon–Tyne
Newcastle–upon–Tyne
United Kingdom

May 17, 1998

Abstract

Traditional geometric scene analysis cannot attempt to address the understanding of human vision. Instead it adopts an *algorithmic* approach, concentrating on geometric model fitting. Human vision, however, is both quick and accurate but very little is known about how the recognition of objects is performed with such speed and efficiency. It is thought that there must be some process both for coding and storage which can account for these characteristics. In this thesis a more strict emulation of human vision, based on work derived from medical psychology and other fields, is proposed. Human beings must store perceptual information from which to make comparisons, derive structures and classify objects. It is widely thought by cognitive psychologists that some form of symbolic representation is inherent in this storage. Here a mathematical syntax is defined to perform this kind of symbolic description. The symbolic structures must be capable of manipulation and a set of operators is defined for this purpose. The early visual cortex and geniculate body are both inherently parallel in operation and simple in structure. A broadly connectionist emulation of this kind of structure is described, using independent computing elements, which can perform segmentation, re-colouring and generation of the base elements of the description syntax. Primal colour information is then collected by a second network which forms the visual topology, colouring and position information of areas in the image as well as a full description of the scene in terms of a more complex symbolic set. The idea of different visual contexts is introduced and a model is proposed for the accumulation of context rules. This model is then applied to a database of natural images.

Acknowledgements

I would like to acknowledge the major contribution of my supervisor, Prof. Graham Megson for many useful discussions and guidance throughout this research. He has shown me how to be an altogether better worker in a University context.

This work was supported by a EPSRC CASE award in conjunction with Neural Computer Sciences, Southampton.

People who have also been instrumental in the preparation of this work are:

- Nigel Steele of Coventry University, who set me on the path of research.
- Dr. Said Salhi of Birmingham University, who helped me understand how universities work.
- Dr. Frode Sandnes of Reading University, who supported me on many occasions.
- Dr. Bob Fisher of Edinburgh University and all the staff with whom I have argued and debated over the past months in order to understand the connections from the retina to V1 and the structures of Wittgenstein's propositions.
- All of the management and staff of Neural Computer Sciences in Southampton for showing me how to construct better C++ and who funded my CASE award.

This document was prepared using the LaTeX typesetting system. All of the code was written using GNU C and C++ under the free-ware operating system Linux.

Dedication

This thesis is dedicated to the following people :

To my Dad, who did not live to see its completion.

Altissima quaeque flumina minimo sono labi.

To my partner Karen, whose patience has been greater than I have sometimes deserved and whose support I could not have done without.

And to my family for their constant support.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

" Better give that up, all those who have taken to colour vision have failed to do anything sensible afterwards." – Professor Robert Tigerstedt to R. Granit.

1.1 The Problem of Human Colour Vision

For sighted humans, visual perception is apparently effortless. Glancing across the room, it is seemingly impossible not to identify the first object that one sees, so automatic does the process appear. Undoubtably, the vision process is very important. It is used for navigation, balance, object recognition, the guidance of social interaction and a host of other related activities. It is not entirely surprising then that 60% of the brain cortex of monkeys is devoted entirely to visual processing [112].

There is no easy solution to the problem of vision modeling, only a process of model proposition, evaluation and refinement. Much of modern thinking on this modeling has been heavily influenced by the work of Marr [24,38,85] who has proposed that the task is split into cognitive subtasks providing both basic and high-level processes. These processes are then expressible in terms of information processing.

It is the current view of cognitive psychology that information processing is the appropriate way to study human cognition [110]. This paradigm has several basic characteristics:

- People are viewed as autonomous, intentional beings who interact with the external world.
- The mind through which they interact with the world is a general-purpose symbol processing system. In this context, *symbols* are patterns stored in long term memory which are designated to point to structures outside themselves, [111].
- Symbols are acted upon by various processes that manipulate and transform them into other symbols that ultimately relate to things in the outside world.
- The aim is to specify the symbolic processes and representations that underlie performance on cognitive tasks.

- The mind is a limited capacity processor having both structural and resource limitations.
- The symbol system depends upon a neurological substrate but is not constrained by it.

1.2 The Current State of Scene Analysis

Since the first efforts by NASA to perform image processing, scene analysis has been need driven. To construct cartographic maps of the lunar surface, image enhancement and noise removal had to be performed on received data. As a result a large variety of statistical algorithms have been developed. These statistical algorithms are of several types with well defined aims and applicable contexts :

- *Segmentation algorithms.* These are region finders based upon gradient changes in (very often monochrome) images. These segmentation schemes are often closely adapted to the data in order to find the *expected* regions.
- *Model fitting.* Many different model bases are used depending on the task at hand. This approach is of no real use for understanding the real world since the task becomes one of best fit or optimization in a limited model space. As will be discussed later, a geometric model is a very limiting idea since its generalization is difficult and its expandability quickly becomes very complicated.

1.3 Approach

1.3.1 Philosophical Primer

Kant

Immanuel Kant is thought by many to be the first of the so-called logical- positivists. In Kant's conception, human thought is guided by a priori principles and concepts that are not built on experience. Also, in his view human knowledge constitutes the ultimate reality.

Three of Kant's innovations changed the philosophical landscape and led ultimately to twentieth-century rationalism :

1. Kant's model of a priori concepts is the basis of the logical positivist search for truth in language and the concept of innate structures in knowledge.
2. Kant's rejection of Descartes dichotomy between the instinctive reflex of animal and man's rational thought process.
3. Kant's emphasis on the supremacy of knowledge over other levels of reality is another of the bases of logical-positivism.

Logical Positivism

Wittgenstein further applied the analytic treatment of human thought to the study of human language, our communication and the content of our communication. Over all, he attempted to provide a philosophical definition of knowledge, what we can know, by analyzing the meaning of language.

In his *Tractatus Logico-Philosophicus* [158] he lays out his fundamental belief that all knowledge is language and vice versa :

> **4.0.0.3.1** All philosophy is a "critique of language"
>
> **5.6** *The limits of my language* means the limits of my world.
>
> **5.6.1** We cannot think what we cannot say.

Wittgenstein then goes on to define knowledge in a particular way. He says that there are certain elementary facts, there are propositions about relationships between elementary facts and there are certain allowable transformations on such propositions that yield composite propositions. His conception of human thought is that we receive perceptions which become our elementary facts. We can then transform these elementary facts and derive relationships among them according to logical processes. Any thought outside this scheme is either false or nonsensical.

Alfred Ayer [159,160] has carried on Wittgenstein's work, correcting errors and achieving a rigour that had not been present previously. Logical positivism argues that every statement (all knowledge) is either based upon sense data or is based upon logic. It rejects all metaphysical ideas as meaningless and having only emotive force. Many of the theories of linguistics and computation are derived from the formalisms of logical-positivism and it forms the basis of much of artificial intelligence (AI).

A basic language of atomic truths is required for a processing of human perceptions into symbolic conclusions. Picture languages have tried to restrict their domain to problem specific areas and have made limited progress [19, 20, 21]. This does not mean that scene analysis can do without them. Indeed, if the symbolic approach to vision is to succeed at all then a ground truth must be established.

1.3.2 Context

Context is the issue that complicates all approaches to computer vision, whether from a fundamentally AI perspective (as this is) or from a Computer Vision (CV) perspective. Many CV approaches ignore it altogether and restrict their domain to simplistic (but lucrative) toy-worlds where lighting, models and aims are simple. The mathematics and the truths searched for, however, become more and more complex. For an account of the splitting of vision in AI and CV see [161].

Atomic truths exist in scenes but are convolved into a wide range of contexts that make them difficult to derive. I propose that the assessment of context will provide the de-convolution of these truths and suggest a biologically plausible probabilistic approach for building context models.

1.3.3 Background to the Approach Used

In contrast to previous approaches, the idea behind this work is to establish a general model of scene analysis by outlining a contextual model with wide applicability. It was a requirement from the outset that methods employed should have a strong parallel aspect in order that a hardware implementation could be used at some time in the future when a wider and more cost effective base of parallel hardware was available. With the advent of FPGAs [127] and general hardware implementation of algorithms on the horizon, this is rapidly becoming possible.

1.3.4 The Connectionist Approach

The operation of the human brain is fundamentally different to that of a modern digital computer because the human brain is highly complex, non-linear and massively parallel. Connectionists seek to narrow this difference by designing algorithms that attempt to simulate what is believed to be the operation of the brain in order to exploit its processing abilities.

Connectionism is widely believed to be a new approach to the construction of algorithms. In practice, however, it is nothing more than an extension to the range of approaches already available to the computer-based problem solver. Usefully the implementation of connectionist algorithms also sits easily with the modern trend of "object-oriented" program design. The objects which most easily represent the fundamentals of connectionist algorithms are the computing *neurons* which are small, generally independent, computing elements which perform some simple task in light of the flow of information reaching them. They may be used simply to replicate the action of tasks implemented in parallel or to perform some kind of learning. Learning may take place both in the neurons themselves and in the weighted connections, or *weights*, between them according to some predefined *learning rule*. The trained *neural network* may then be used for classification or whatever predefined behaviour was required.

There are many advantages to using this kind of algorithm implementation:

- *Simplicity of implementation.* Neural algorithms, as has been mentioned, are easily implemented in an object-oriented fashion since they lend themselves to an obvious object hierarchy.
- *Adaptivity or learning.* If no predefined behaviour is available to exploit then the neural network may be trained to some input/output combination in order to provide a conditioned response. This may lead to generalization of the input data to give classification of unseen data.
- *Hardware implementation.* Neural networks are much more straightforward to implement at a hardware level than conventional algorithms. Current breakthroughs in field programmable gate array technology mean that an easy one-to-one mapping between software neural-networks and hardware algorithms seems feasible even in the short term.

1.3.5 The Parallel Approach

The advantages of designing inherently parallel algorithms are :

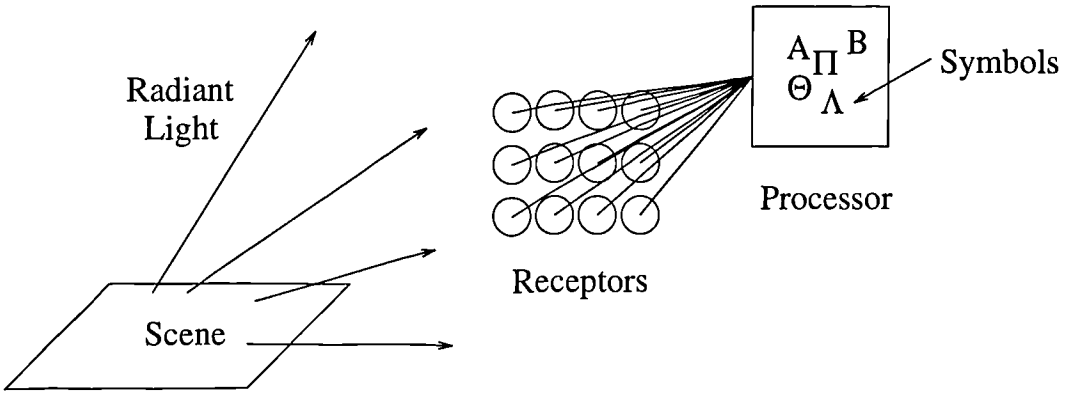


Figure 1.1: Axioms

- *Speed.* Conventional, single processor general architecture (GA) computers are far too slow for many important problems. Perception of objects is a complex problem which must be performed in real-time. Even recognising the moving pictures from a television screen must be performed at around 10-50 milliseconds per frame.
- *Organisation.* It is known that the human perceptual system is massively parallel. Primarily this means that even if damage is sustained then high structural integrity is maintained. Also, basic computing units may be very simple (as neural networks nodes are) and information may be stored in a distributed fashion.

1.4 Statement of Thesis

1.4.1 Assumptions

Assumptions made in this thesis are limited to those required to axiomatically state that if a scene is perceivable by humans then it is perceivable by a computer. That is :

1. Receptors exist that are capable of basic registration of light energy which is radiant from the scene.
2. The scene can be said to be perceivable. That is that incident light falling on the scene has a radiance that is incident on some receptor.
3. Some internal processing of this registration, or *image* is possible.
4. An internal symbolic set is used to represent this image, since without such a set no *meaning* can be derived from such an image.
5. Operations are performable on this symbolic representation.

If all of these axioms are in place then there is no arguable difference between the perception of the human and that of the computer. Once the signals are received then the axioms are simply a restatement of the Church-Turing thesis [81].

1.4.2 Developments

The tacit assumption of previous work in scene analysis is that there is only one way of analyzing a scene in order to derive meaning from it. This method involves knowing everything about the scene to start with and simply applying this model to any given image in order to try and spot the premodelled elements *in situ*. As has been argued above, this is purely because the subject has been need driven and no real interest has been shown in deriving meaning, only in fulfilling a given task. This is the derivation of what Umberto Eco [58] has described as $true_1$, which is truth from pre-modelled or written definition. More subtle truth from perception or context, $true_2$ and $true_3$ respectively are much more difficult to derive from disparate data. Humans and other intelligent agents excel at deriving this kind of information. Indeed from a small set of internal symbols they can construct more and more complex meta-symbols which help them describe and interact with an external environment.

Colour plays an important part in the development of the ideas in this thesis. Colour is a fundamental part of the way humans perceive the world around them and make objective decisions about the nature of objects. So far, colour has been overlooked in scene analysis because the technology for colour sampling by computer has not been available or has been prohibitively expensive. Recently such technology has reached a new level of availability and has finally come of age. In fact, all of the work in this thesis has been performed using cheap, widely available, consumer products with high reliability and quality. No doubt as the technology continues to progress, this technology will appear even more modest and lead to a further more detailed exploration of the concepts presented.

1.4.3 Limitations

Some of the approaches used in this thesis are *entirely* applicable. The context model described in chapters 4 to 6 represents a departure in how to think about the concept of scene analysis and scene analysis techniques for the first hand development of understanding. The syntactic description and its accompanying predicates developed in chapter 3 are also widely applicable for the construction of objects and analysis of images.

The heuristics described in section 6 are only applicable to colour and graph analysis since they are topological and colour based.

1.5 Thesis Structure

This work is divided into several sections. The first sections are the steps toward a context model and the application of this kind of model to a set of natural images. The final section is a conclusion together with a discussion of further work.

Chapter 2 deals with image representation and analysis. This section discusses both human and computer representations of scenes and images. The nature of symbols is also considered and the need for a symbolic syntax is expressed. Such a syntax is developed in chapter 3 where a minimal set of syntactic descriptors or atomic truths is described with predicates for their manipulation. Segmentation

is discussed in chapter 4. The problems of good general segmentation are addressed together with an absolute description of what is actually required for scene and context analysis.

The ideas behind the ICEnet, REDnet and COLnet networks are then developed in chapter 5. First algorithmic steps towards scene description is described then the parallel algorithm for segmentation and production of the Region Adjacency Graph (RAG) as well as automatic production of the base level set description for the syntax is also addressed.

Chapter 6 describes the assessment of context once the model has been derived along with the primary syntax. Many pictures of natural scenes have been analyzed in order to derive a general context model of the elements found in them. Chapter 7 describes that application and the results that have been derived. The training of several neural networks is also described.

Conclusions and further work are presented in chapter eight. Conclusions are drawn about the results of this work and the possible applications. Ideas are also expressed as to possible future directions for this work.

Chapter 2

Internal Image Representations and Scene Analysis

2.1 Introduction

Any algorithmic theory of human vision must by its nature start from a philosophically predefined viewpoint. Human thought may well be impossible to model and quantify directly (although methods of analogy are widely propounded and advocated, for example [61][158]), however the analysis of natural language may provide some clues to the high level computational schema adopted [47]. This viewpoint is both objectivist and representationist in nature. That is, we must presuppose that their exists a *language of thought* that is both derivable and that we can functionally emulate. The symbols of this language function as our internal representation of external reality. If such a language does exist then it must be capable of codifying human perception and manipulating symbolic codes to form conclusions about the external world. Consciousness and the process of thought itself may then be described as the parsing and manipulation of symbols held in some kind of small, fast, symbol cache. The attendant dependencies of these symbols must therefore take care of themselves in some pre-determined way.

Much work has been done to determine the types of memory that are used to codify visual perceptions. One example is the classic work carried out by Sperling [150] to deduce the existence of so-called *iconic* memory which has a storage time of only 0.5s. This memory can produce a very brief *sketch* from a very large amount of stimulation from, for instance, visual stimuli. This sketch contains not only the essential elements of the scene but also some of its important details.

Haber has claimed [151] that this iconic storage is irrelevant to normal perception on the grounds that the icons formed from one instant to another would obliterate each other before processing could take place. Haber is quite mistaken though, since the icon is not created at the offset of stimulation, rather at the onset [152]. This means that even in a constantly changing world, there are still adequate opportunities for this iconic vision to be used. Indeed, the mechanisms used for visual perception invariably operate on the iconic memory rather than directly on the visual environment itself. This

means that iconic store is an integral part of visual perception rather than simply an interesting curiosity. Further to this, it is proposed in chapter 6 that the information stored in such an iconic system can in fact be used as a pre-cursor and modifier for much of the recognition that takes place in visual recognition and scene analysis.

2.2 Wittgenstein

2.2.1 Overview of the *Tractatus*

Wittgenstein claims in his preface that the propositions in the *Tractatus Logico-Philosophicus* are true, unassailable and definitive. He starts by defining the principles of Symbolism and the relationships between words and things in a language. From this basis he begins an investigation of traditional philosophy.

He is concerned with the conditions required to fulfill the need for a logically perfect language:

- What happens in our mind when we use language to mean something.
- The relationship between thoughts, words, sentences and the things they refer to.
- Conveying truth rather than falsehood.
- What relationship must exist between two things when one can be used as a symbol for the other ? i.e What is the nature of Symbolism ?

2.2.2 Wittgenstein on Pictures

Importantly for vision are Wittgenstein's propositions on representation and the dynamic relationship between thought and pictures, beginning at sub-proposition 2.1[1]. The ultimate conclusion of these propositions is their superbly concise conclusion, proposition 3.

2.1 *We picture facts to ourself.*

This is an intimation of the sections main thrust, the relationship between visual sensation and thought.

2.12 *A picture is a model of reality*

2.13 *In a picture, objects have the elements of the picture corresponding to them.*

The important word in this proposition is 'corresponding', i.e there is a one-to-one relationship between the real element and the pictorial ones. Also, the picture must be *absolute* in order to be a true representation.

2.14 *What constitutes a picture is that its elements are related to one another in a determinate way*

2.141 *A picture is a fact.*

That is fact in the absolute propositional sense.

[1] The *Tractatus* is written using the Dewey-Decimal system. A proposition n may have an explanatory proposition n.1, which may in turn have an explanatory proposition n.11, etc.

2.15 *The fact that the elements of a picture are related to one another in a determinate way represents that things are related to one another in the same way.*

Let us call this connection of its elements the structure of the picture, and let us call the possibility of this structure the pictorial form of the picture.

2.151 *Pictorial form is the possibility that things are related to one another in the same way as the elements of the picture.*

2.15121 *Only the end-points of the graduating lines actually touch the object that is to measured.*

This is a concise description of the operation of segmentation which is utilised in chapter 5.

2.1513 *So, a picture conceived in this way, also includes the pictorial relationship, which makes it into a picture.*

2.171 *A picture can depict any reality whose form it has. A spatial picture can depict anything spatial, a coloured one anything coloured, etc.*

2.172 *A picture cannot, however, depict its pictorial form: it displays it.*

2.173 *A picture represents its subject from a position outside it. (Its standpoint is representational form). That is why a picture represents its subject correctly or incorrectly.*

2.174 *A picture cannot place itself outside of representational form.*

2.2 *A picture has a logico-pictorial form in common with what it depicts.*

3 *A logical picture of facts is a thought.*

This is the fundamental philosophical underpinning of the argument for a language of description. Once a set of language primitives is formed (referred to as *simples* or atomic propositions) then in a correct, compact, complete language all possibilities can be formed. It is these compound propositions *complexes* that are the topic of this chapter. It may be shown that the Computer Vision insistence on the search for objects is only one approach to the problem of scene understanding. A series of new complexes, known as contexts is defined that help in the appreciation of scenes and may be used in the construction of *thoughts* rather than of objects.

2.3 Information Processing of Visual Stimulation

Some useful conclusions may be drawn about information flow by using an information processing approach when addressing human cognitive abilities.

Assumptions

The major assumption of information processing is that perception is not the immediate outcome of stimulation but is the result of processing over time. This means that neither the visual experience or overt responses are immediate results of stimulation, they are the consequence of a sequence of processes which take a finite amount of time. This time interval may be broken down into a number of stages or processes, corresponding to a series of transformations in internal representations of the stimulus. Processes can be thought to be limited in the amount of information that they can hold at a given time. The magnitude of this storage can be determined empirically but results for humans vary

widely, from a small to apparently limitless amount ! [132]. Capacity limitation leads to selectivity - not all information can be processed to the same degree within the time available for processing. One of the major characteristics of information processing is to specify what determines the selectivity and the mechanisms to perform it.

The major elements in an information processing approach are:

1. *Storage or Memory.* Information may be deposited and retained at various stages in the processing sequence. This property is called memory. Different types of memory are often separated by their relative durations, for example short and long term memory. Information is then the medium of representation used to get the retinal image into memory.

2. *Processes.* Operations may be applied to information that will transform it in various ways as it is used by the perceiver. For information in a store or memory, it requires a process to put it there and one to remove it as well as processes to operate on it in situ.

3. *Information.* Information was defined by Shannon [139] in 1948 very specifically as the amount of uncertainty reduction in a particular communication channel. This measure is independent of both the information type and the channel type. This measure is used very often in psychological analysis.

2.4 The Splitting of Colour and Form

The eye provides a purely mechanical mechanism for focusing light onto the surface of the retina and as such its function is well understood [9]. The retina is responsible for the transformation of these fluctuating patterns of light into patterns of stimulation and then into the optic nerve (fig 2.1). These patterns of stimulation convey information about surfaces, objects and movement in the optical array. Many transformations are carried out in the retina before this information is then passed forward for further processing.

The Bauhaus artists of the 1920's and 1930's, particularly Kandinsky, believed that colour and form were linked directly and inextricably. However, recent neurophysiological evidence shows that human vision handles information about colour, acuity, speed and contrast using totally separate neural pathways [68]. This has been suspected for some time, and was finally demonstrated in the work of Kendrick [73] and others [74]. The fusion of the results from these pathways is then drawn together to form perception.

The idea of this division of labour came primarily from Hubel's early work [33, 68, 163] and others [69]. New work, however, has capitalised on advances in brain imaging techniques using positron emission tomography (PET) to gain a direct insight into the electrical patterns inside the brain when visual stimuli are presented.

The major part of the brain which deals with vision is the geniculo-cortical part or the lateral geniculate body. This has two obvious subdivisions, four parvocellular layers and two magnocellular layers, fig 2.2. Each eye projects to three of these six layers [70] in an alternating fashion and each

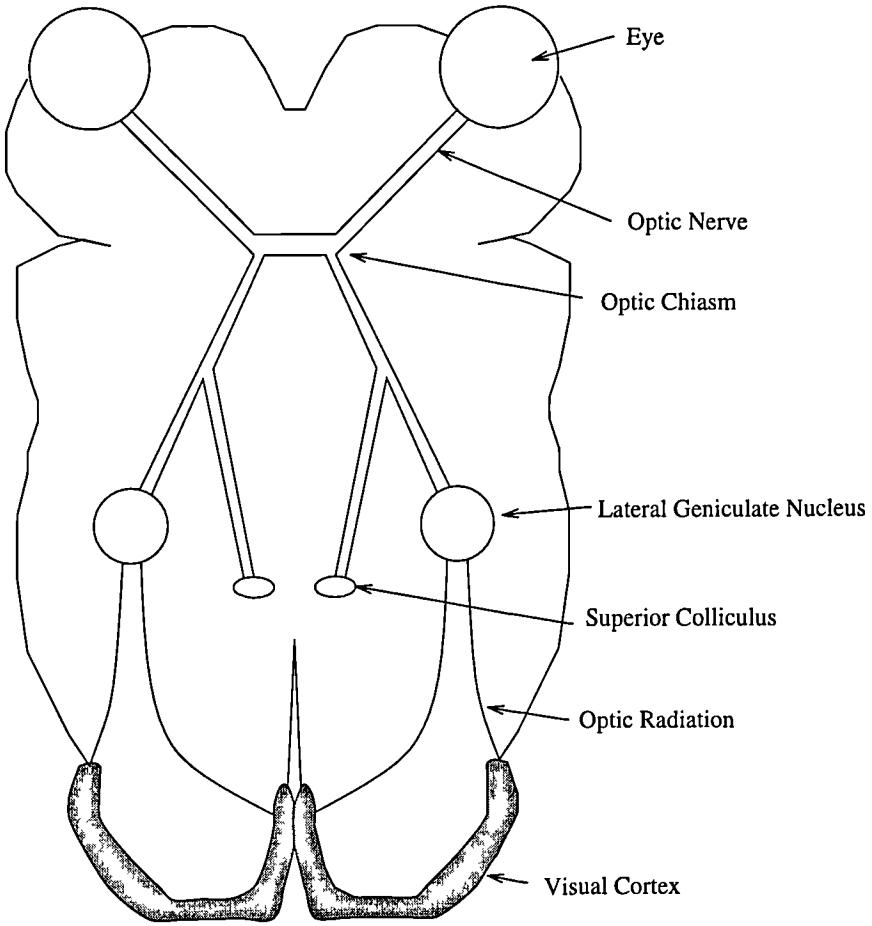


Figure 2.1: Pathway through the Visual System

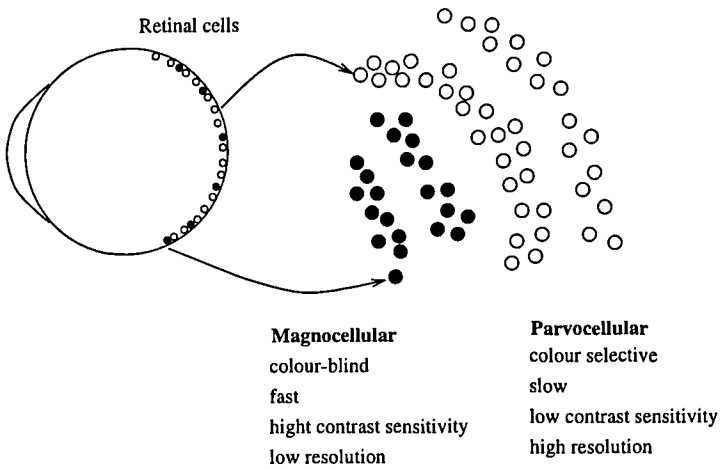


Figure 2.2: Segregation of the Visual System

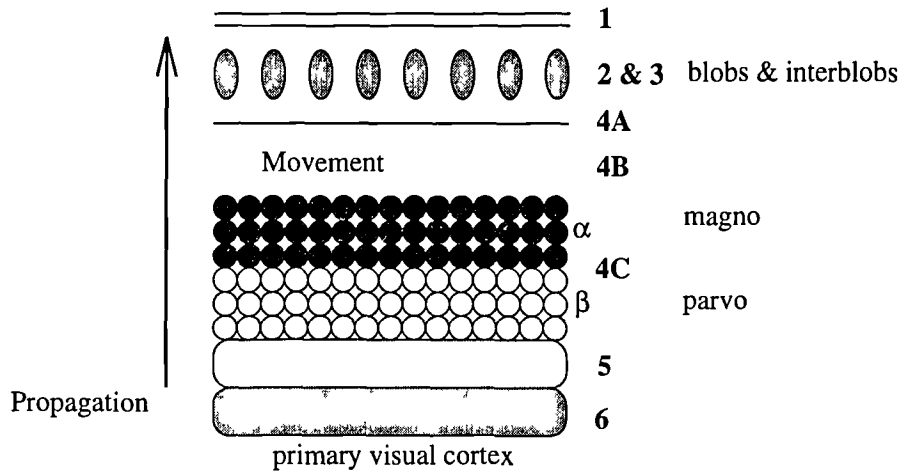


Figure 2.3: The Visual Cortex

half-retina is mapped three times onto one geniculate body, twice to the parvocellular layers and once to the magnocellular layer. All six topographic maps of the visual field are performed simultaneously. The magno and parvo divisions differ physiologically in the following ways:

- *Colour.* About 90% of the cells in the parvocellular layers are sensitive to colour, whereas the cells in the magnocellular cells are not. Magnocellular cells only perform a basic on/off function at all wavelengths and are thus colour-blind so two colours at the same relative brightness will cause the same response.

- *Acuity.* Magno and parvo cells have different field sizes. Magno cells are much larger than parvo cells although they both increase in size from the central point, the macula.

- *Speed.* Magno cells respond much more quickly than parvo cells. This means that they play a special role in the detection of movement. Many cells higher in the pathway sense direction as well as movement.

- *Contrast.* Shapley [69] has shown that magno cells are much more sensitive to low-contrast stimuli.

These results mean that it is very probable that the two different kinds of body contribute to different aspects of vision. Further evidence to support this is the secondary level of processing which takes place in the middle temporal lobe. The system for the parvocells is parvocells $\rightarrow 4C\beta \rightarrow$ blobs and interblobs $\rightarrow$ areas 2 and 3 [68], fig 2.3. The interblobs respond well to achromatic luminance and contrast borders. However, many of them also respond to appropriately oriented colour-contrast edges regardless of the colours forming the edge or the relative brightness of the two colours. Similarly, they usually respond to lines or borders of any brightness contrast (light-on-dark or dark-on-light). This suggests that much of the colour-coded parvocellular input is pooled in such a way that colour contrast can be used to identify borders but that the information relating directly to colours forming the border is lost.

The blob and interblob region work in entirely different but complementary ways. Blob cells are explicitly colour coded, excited by colours in one region of the spectrum and inhibited by others, and are not selective for stimulus orientation. Interblob cells are selective for stimulus orientation but mostly are not colour selective, responding to lines and edges, regardless of colour. The strategy of carrying information in a system that mostly pools colour information and colour-contrast information in a separate system that does not carry orientation is probably more efficient than having single cells selective for both orientation and colour of a border.

Clearly then, human vision operates a divide-and-conquer strategy for visual processing in which form and colour are split into two channels and are then operated on in partial isolation. Zeki [72] is also a proponent for hierarchical processing and has shown that some cells in area V4 of the cerebral cortex show colour constancy. This is our ability to perceive a surface as having a constant colour despite the changes in spectral light incident on and reflected from it. Zeki [73] has also reported that a number of cells responded selectively to a surface of a particular colour and maintained that response despite changes in the composition of the light falling on this surface, just as the observers perception of the surface remained constant. The magnocellular layer, however showed no signs of this kind of constancy.

More evidence for hierarchical processing comes from the work of Perretts et al. They have shown that some cells in the inferotemporal layer (known as V1) are selective for both simple processes, such as orientation as well as fairly elaborate processes. They have shown in monkeys that 10% of cells in this area responded to, that is showed a preference for, *faces* of either humans or monkeys. They also found that it made little difference to the outputs from these cells if the faces were distorted, were human or were monkey. The cells responded much less to images of scrambled facial features so were responding purely to the spatial configuration that represented a face. Thus this whole area of the brain responds more actively to topology and colour of the whole rather than to individual elements or features.

Similar but more detailed results have also been shown more recently for sheep by Kendrick [74]. In a sheep's brain, specific groups of faces or facial features are coded for by separate subpopulations of cells. Most cells respond only to faces with horns, and the larger the horns the greater the change in their output frequency. Even crude line drawings of sheep faces show this effect of horn size. A second group of cells responds only to faces of animals of the same breed and particularly familiar individuals. A third group of cells responds exclusively to the faces of humans and dogs. So sheep deal separately with information relating to dominance features, faces of other familiar sheep and faces of potentially threatening species, such as humans and dogs.

Brody [119] has developed a simple feedback model of the functioning of the pattern derivation sections of the lateral geniculate cortex. This model, although not agreeing directly with known physiology, has shown that robust, low-level pattern recognition is possible using only a simple emulation of the cortex $\rightarrow$ LGN $\rightarrow$ cortex feedback loop. Several models of the navigation regions of the cortex have also been proposed which explore the data passing mechanism from both information processing [121] and a hard modeling standpoints [122, 123, 133]. Although achieving differing levels of success, they

do show that by using both neural structures and fairly simple firing functions it is possible to create intelligent-like models of different elements of brain function.

No model of the magnocellular functioning yet exists, however certain facts are known from experimentation [124] which appear useful for producing a model.

- The pattern of connectivity, or ocular topology, is of primary importance to the information content of the array of retinal stimuli [125].
- The development of ocular dominance stems directly from the action of post-synaptic cells and is proof that if both eyes are seeing the same scene then any elements repeated are transmitted only by one eye instead of both [124].

Any model must then place importance on topology and patch connectivity but may relax the need for stereo processing.

2.5 Internal Representation

In an unnumbered Definition in *Opticks* [42], Newton states that

1. The physical constitution of an object (i.e, the primary qualities of the objects 'insensible parts') explains how the object has a disposition to reflect selectively various types of light rays.
2. The physical constitution of light explains how its rays have dispositions to excite various types of processes in the perceiver ("motions in the sensorium").
3. These processes explain how we have various types of colour sensations ("sensations of these motions under the forms of colours").

This is the basis of the *representationist* model of perception and states axiomatically that perception includes some internal medium in which features of the external world are re-presented to the perceiver. That is to say that an internal representation is formed and it is this symbolic representation that is used to derive such concepts as causality, distinction between forms, etc. This symbolic set has been termed *ideas* by Locke [43] as distinct from *qualities*. Ideas are what is formed by the process of thought whereas the quality is the basis of the power that causes the idea in the first instance. He also distinguishes between two types of qualities, the primary and the secondary. Primary qualities are those qualities which are inherent in an object and are unchanged by our perception, for example shape, size, mass, etc. Those qualities which are subjective are considered secondary, for example colour, sound, taste, etc. This view is therefore fundamentally representationist and requires a medium of representation.

It is still not known how the human brain encodes the information that is then stored in the memory. It is known that the brain transmits signals in the form of modulated nerve signals [72] but these are not simple digits. One of the largest problems currently facing neurophysiology is the search for the encoding mechanism and the units in which a record is written into the memory [41,61]. It seems

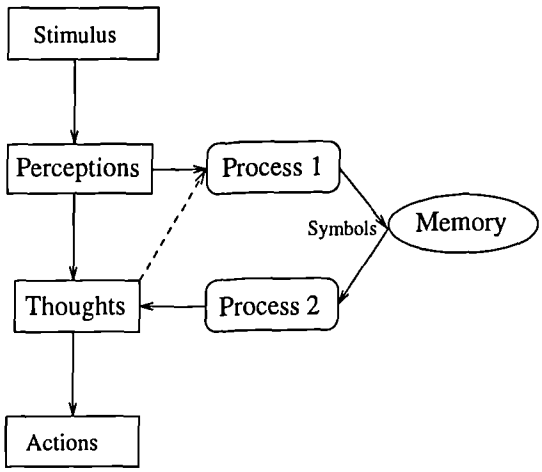


Figure 2.4: Internal Representation

that the code is likely to consist of various types of units that are genetically pre-established but are uniquely adapted to meet the needs of each individual.

Calvin [61] suggests a mechanism for this activity where Darwinian selection processes are employed in order to both form and utilise thoughts. Only perceptions can be stored since they are the only form of information gathering humans are capable of [41]. The coding is thought to be performed by some physiological transformation process (shown as process 1 in fig.2.4) perhaps involving frequency and topological coding and retrieval must occur by the matching inverse-transformation of that process (shown as process 2). In assuming this structure the main opacity comes from the internal symbols and constructions used in the representation. These, he has conjectured are stored in hexagonal units on the surface of the cortex (which is essentially 2-dimensional) in both a topologically ordered and relational fashion. Each hexagonal unit *is* an idea or symbol and corresponds exactly to a physical structure inside the brain consisting of a few thousandths of millimetres in area or the order of a minicolumn. The weight of Calvin's thesis can be summarised as follows:

- Human thought has a physical nature.
- That nature is directly emulable.
- Thoughts, that is representations, place themselves into the physical structure of the mind using a selection mechanism in order to reduce redundancy.
- The elements can be collected to form a hierarchy of thought.
- Given this hierarchy, symbolism and complex constructs such as metaphor are possible.

A representation is a symbol for an object or set of objects. There are thus many representations for any given scene, whether they are in the form of language, images or artistic expression, as shown in fig.2.5. All such symbols are valid provided that, within definable limits, they stand unambiguously for such a scene. Clearly, images are representations of scenes but Goodman [27] suggests that there need not be any physical resemblance between such a symbol for a scene and what is depicted in it.

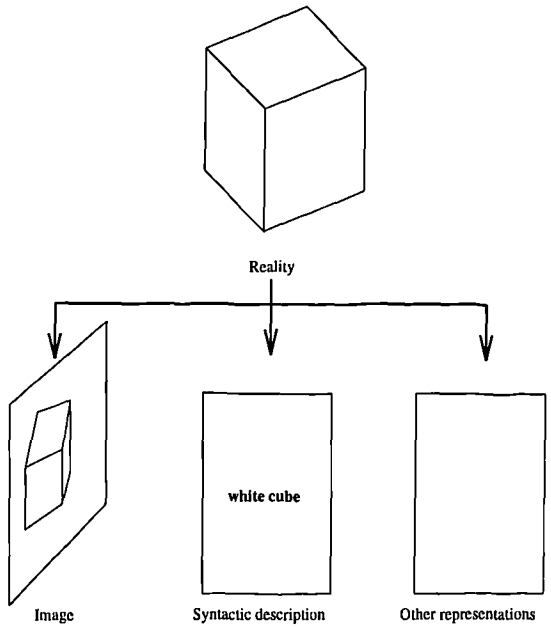


Figure 2.5: Representations

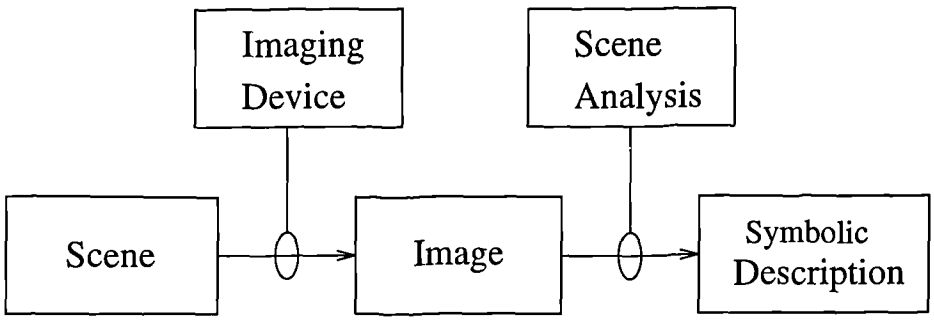


Figure 2.6: Traditional Scene Analysis

All sets of representational symbols are called *languages* [26, 158]. An image is a language containing such sets of symbols. Scene description is then a translation from the images symbolic set to, say, a natural language symbolic set. This process is referred to as *generating a symbolic description of an image* [1].

There are two types of information about the scene in an image, as made distinct by Marr [24]. The first of these is information about the structure of the scene and the second is information about how a particular scene looks from the given viewpoint. Pictures presenting structural information are divided into *object centred* representations, providing information about the structure of the objects in the scene and *array centred* representations, providing spatial information about the objects. Pictures giving information about the way a scene looks from a particular viewpoint are termed *viewer centred* representations. One important point about this representation is that it may not be good for conveying structural information about the object and scene.

The question remains, is the internal representation of images a real language which is open to interpretation. If this were the case, the analogy to the representation of natural (or verbal) language would be quite striking. Primitives could be found with properties analogous to those of words in the

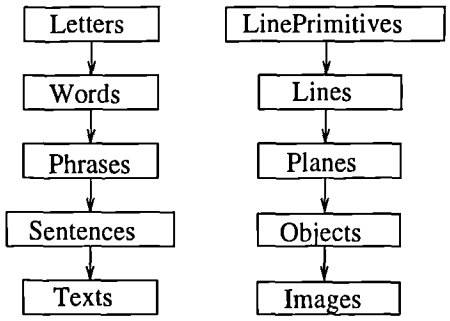


Figure 2.7: Natural Language Analogy

natural language model:

- Words can combine to form more complex entities which convey meaning, i.e *phrases.*
- These phrases may be built into meaningful *sentences* or *texts.*
- Meanings must eventually be traced down to truth conditions. Literally, is the import of the conveyed information true ?
- This tracing is a function based entirely on explicit interpretation rules which apply to abstract structures associated with the expressions.

The analogy being drawn is between the expressed intention of a natural language construct and structural information content in a visual construct. Human vision differs in that there is no expressed intention, rather intention is drawn internally from the stream of data which is presented. Focus of attention is therefore necessary, as is an ability to derive previously seen structures or forms from the scene. This visual language thus constructed must have analogous properties to natural language:

- Segmentation primitives can combine to form more complex entities which convey structure such as *lines* and *planes.*
- These planes may be built into meaningful forms or *objects.*
- Object structures must eventually be traced down to truth conditions. Literally, is the import of the conveyed information true or does it correspond to something which is recognised.
- This tracing is a function based entirely on explicit interpretation rules which apply to abstract structures associated with the expressions, perhaps in the form of topological or relational mappings.

In the next chapter a novel form of visual description language is derived as well as a set of manipulation functions, or predicates, which may be used to form some conclusions about the scene. Techniques for displaying, graphing and manipulating this language are also discussed.

Human beings understand images simply because the processes used in perceiving and processing the visual symbols in them are the same as those involved in perceiving the real scenes [28, 59]. Surrealist artists, for instance, have often exploited this fact by deliberately representing objects ambiguously

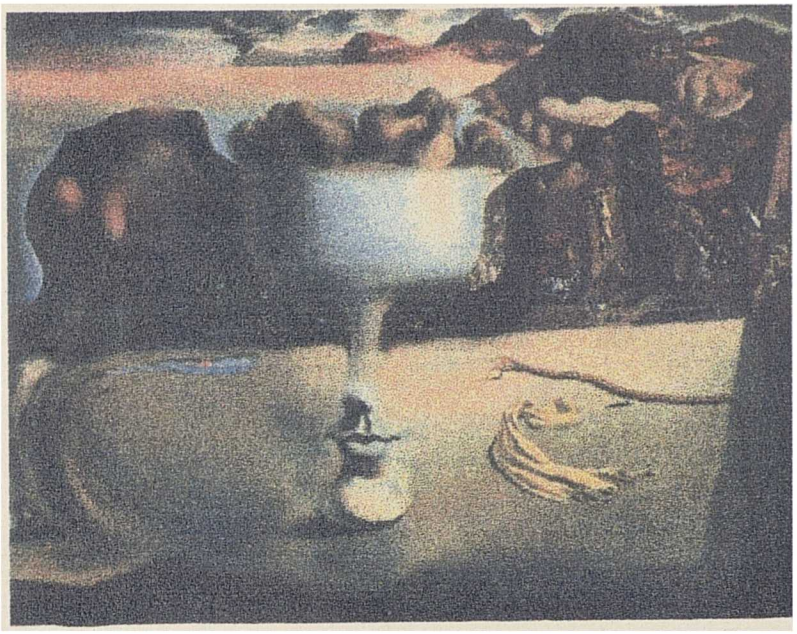


Figure 2.8: Dali: Apparition of a Face in a Fruit Dish

in order to disconcert the viewer, as in fig.2.8, Dali's *Apparition of a Face in a Fruit Dish*, which appears to be a face, a fruit-dish and a dog simultaneously. In this case the structural symbols representing part of objects are mixed to form the ambiguity. It may be noted that this is perhaps some indication of a caveat that may occur if an accurate model of human perception were established.

There are many perceptual theories based on representation, for example :

- Gibsonion ecological theories, suggesting that humans detect structures in pictures which they are trained from birth to detect in the external world [82]. He also refutes that humans can be divorced from the environment and operate on perceptions without reference to actions.
- Constructivist theories where perception is said to consist of constructing perceptual meaning according to past experience [41].
- Gestalt theories suggesting that perception consists of structuring visual inputs according to certain laws, as discussed in [156].

Each of these theories relies on internal representation to a large extent.

2.6 Human Symbol and Semantic Acquisition

Perkins [56] has remarked that the syntactic representation of words in a sentence is not actually useful for information handling and reasoning. Similarly for computer vision, deconstructing an image into solids does not produce understanding *per se*. The problem here is that parsing a sentence can only be achieved after experience has been accumulated. That is to say that a very large amount of contextual information must be gathered before useful processing can take place.

Gathering information (symbols) is a two step process. First information is extracted from incoming signals. Then correlations must be derived between and among different pieces of information. Language is thus acquired through symbolic processing. Simple symbologies are built into more and more complex ones until high information transmission rates are achievable. This kind of symbolic processing is different from ordinary data processing. The human brain is particularly good at symbolic processing and it is believed [57] that the complexity of the symbologies used to represent information increases from the front-end of the visual system to the back-end.

To acquire symbolic processing capability human beings must possess a symbol 'boot-strapping' capability to allow simple symbologies to build into more complex ones. The model used for symbolic acquisition throughout this work is called the "Symbol World" and is described in section 3.1.3. This model of acquisition can be seen as a generalisation of the work of Bruner et al. [126] which advocates a defining attribute categorisation theory based upon experimentation. They identified several different strategies used by subjects that could be used by people to acquire concepts in everyday life by using artificial categories.

2.7 Summary

Marr [9] proposed that different levels of theory must be distinguished if we are to understand visual perception :

- *A computational theory.* This theory should describe what is to be computed and why.
- *Algorithms.* Algorithms must be described to achieve the computation.
- *Representations.* These form the input to and output from the algorithms.
- *An implementation* of the algorithms.

Human vision consists of two parts, a purely physical process for deconstructing the image incident on the retina and then representation as thought which can be used in further mental processes. Initially, the physical part consists of two augmenting processes, one for derivation of form and another for derivation of colour and topology or structure. From this part onwards there is a hierarchical formation of processes which so far are only vaguely understood.

In considering the available neurophysiological, representational and objectivist literature the following conclusions have been drawn:

1. A realistic model of human vision must emulate the division into colour and form processing.
2. Such a model must be hierarchical.
3. Human thought processes and the physiological structure of human storage still prove to be opaque. They can, however, be modeled symbolically.

4. Representationism seems to be the most reasonable approach to solving the problem of human interpretation of visual stimuli. That is, description or symbolic representation must be introduced early in the computational processing scheme. A simple symbolic visual language, or *picture language* is the subject of chapter 3.

5. Syntax based rule sets are necessary in order to form conclusions (or truth tables) about what is described. This is discussed in chapters 3 and 6.

6. Any model that accepts both the representationist model and the natural language analogy must have predictive capability based on application to some real world task. This kind of application is the subject of chapter 7.

Chapter 3

Syntax

3.1 Description languages

It has been previously shown that there is a firm philosophical and biological foundation for the concept of "picture languages". However, it has not been discussed and there is no cross-discipline literature available on the subject. Representations both for input images and results have been made that have suited particular applications or computing hardware. In general, the problem of understanding has been left to one side in favour of model fitting. Even the world of artificial intelligence is split in two by the problem of representation [152]. On one side there is the *connectionist* school and on the other there is the *language of thought (LOT)* school. Although the gap appears to be wide, this is not necessarily the case and they do share much of the same ground. Traditionally, languages have been used as a representational symbology in the LOT school and they have been derived and parsed in a serial way using von Neumann architecture computers. As will be shown in chapter 5 and chapter 7, this is not necessarily the only, or best, way for either generation of a language or for later parsing of that language.

3.1.1 Aims

As discussed in the last chapter a schema for representation must be decided upon in order to progress toward a computational theory of the system model for vision. Kovalevsky (as quoted in [19]) has stated the traditional decision theoretic approach as follows:

"It is necessary to recognise a set of situations, k, by coming to various decisions, d. Decisions are based upon observable results, v, of an experiment conditioned by a situation. This dependence is characterised by a conditional probability distribution, $p(\frac{v}{k})$. Quality of reached decisions is evaluated by a magnitude of loss, $L(\frac{k}{d})$, which is specified for each situation, k, and for each decision, d. It is necessary to find a rule, $d(v)$, for reaching decisions based upon observations, v, which results in minimal mathematical expectation of losses."

When applied to computer based vision, the set of situations, v are features, the decisions, d are results of comparisons with models and the rules, $d(v)$ are rules based upon features. There are some

inadequacies in this approach, however. The attributes used in the classification procedure do not play any conceptual role other than for classification. A great deal of information is often available regarding object structure and other *derivable* semantics. It is clear that the structure of a scene must be analysed in order to compare it with another, not simply a derivation of feature elements.

What is required is a form of descriptive schemata that can be searched for features (*tokens*) belonging to classes in order that we can generate (*generative schemata*) or interpret them (*interpretive schemata*). Also, the relevant features must be brought out in the schemata and not obscured by irrelevant information [45]. Before we can search for patterns, however, we must be able to state what a pattern is. Consequently, we define a pattern as an *organisation* of sub-patterns, objects or elements. As Narasimhan [19] states:

"An organisation is a complex of relationships *that subsist between elements which are organised. This is one method of recognising a pattern, to see it as a* particular *organisation. That is to see it as particular objects satisfying particular relationships."*

This is referred to as a compositional description. Another method is to organise the patterns into primitives and analyse the relationships between them, often referred to as a transformational description. Marr [24] describes the goal of all research in vision as :

"to understand how descriptions of the world may efficiently and reliably be obtained from images of it. We must ask what kind of information does the human visual system represent, what kind of computations does it perform to obtain this information and why ? How does it represent this information and how are the computations performed and with what algorithm ?"

Finding an efficient method to describe the world is the touchstone of this argument. It is clearly the case that humans formulate ideas about the world use language as their internal representation. In fact, it has been proposed that the representations for images and language may be identical (from psychology [149] and from philosophy [158 proposition 3]). This section deals with some of the representations that may be used in order that a computer may make logical assessments and derivations about scenes. It outlines the basic elements (atomic propositions or simples) and method of construction of complex symbols (complexes).

As an example of a feature-based symbolic description set, Sloman [128] suggests that a large vocabulary of many different scene fragments is required. He gives a list of twenty-six possible types including concave and convex curves, corners, holes, etc. This could provide a basis for a very simple form of descriptional language.

3.1.2 Early Work

Guzman [6,7] has suggested classing vertex types and thus surfaces in order to compile a connectedness scheme for a scene. This resolves the issue of which objects are which. Models can then be applied singly to each of the objects. Waltz [2] has proposed a line classification method that uses an applied scheme of line orientation to tell one object from another and which objects occlude others. This approach is particularly useful for detecting shadows of objects. He severely constrains the way lines

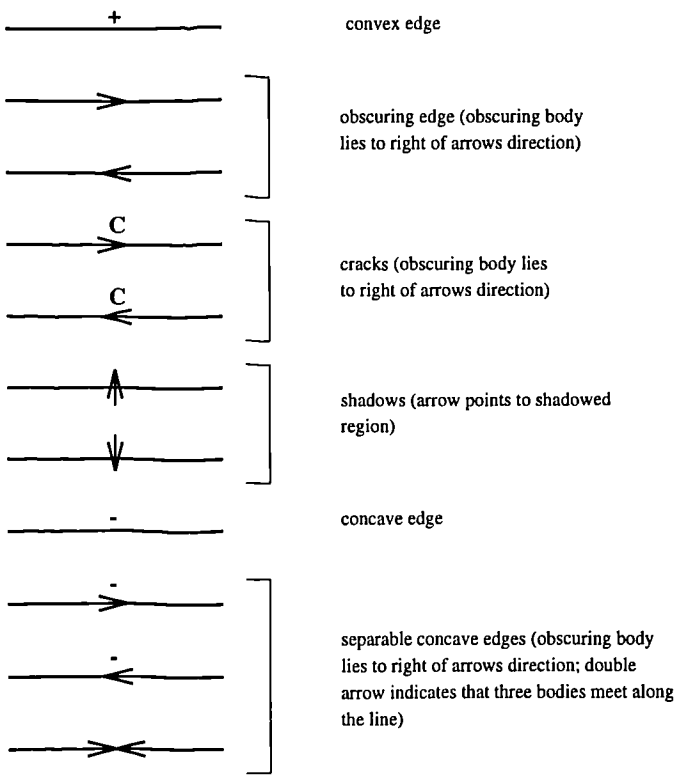


Figure 3.1: Edge Classifications

and vertices fit together in completed line drawings. He shows that deciding if a particular line in a drawing is a shadow, a crack, obscuring edge or internal seam can be done in a way that is analogous to the solution to a set of algebraic equations. In algebra, one has a set of variables with constraints in the form of equations, whereas in scene analysis each line corresponds to a variable and determining the line's physical origin. This corresponds to solving an algebraic equation for the variable values. The constraints, analogous to equations, are the given vertices. Waltz recognises eleven different categories of line, shown in fig. 3.1. Where two lines meet, the arrangement of the lines around a vertex is limited. It is reasonable to suppose that the number of forbidden combinations is enormous but this is not the case, the number of possible combinations is quite small, as shown in figure 3.2. As is shown in fig.3.3, a typical Waltz scene classification is relatively simple in nature.

Narasimhan [19] has shown that it is possible to form many different types of *generative* languages that can be designed for custom applications. One example of this is for the generation of written characters. Each of the generation schemes uses picture tokens of finite spatial extent in 2-dimensions, P, a set of attributes, A, a set of relations, R, and a set of composition rules and a set of transformations, T. A specification language, G is then formed by the 5-tuple :

$$G = (P, A, R, C, T) \tag{3.1}$$

P is a set of primitive actions, the performance of which is generated by picture fragments. These are the so-called picture atoms. Each of these atoms has a set of attributes, A which can assume values from a well defined range. Fixing any of these values is an *assignment*. P and A together

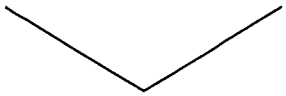
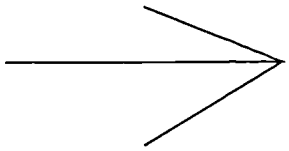
	Approximate number of combinatorially possible lables	Approximate number of physically possible labels
	3,249	92
	185,000	86

Figure 3.2: Line Possibilities

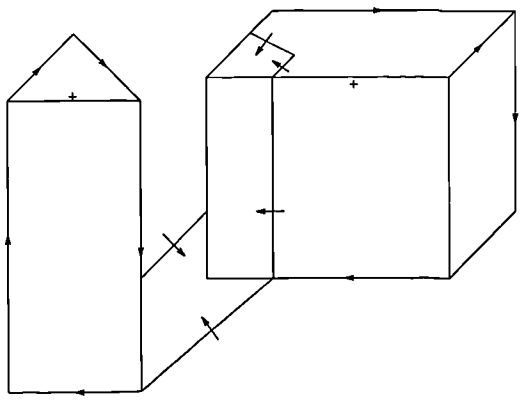


Figure 3.3: Example Waltz Classification

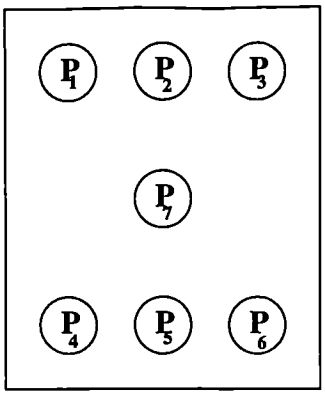


Figure 3.4: The Seven Primitive Regions of a Frame

allow generation of atoms. Picture fragments form a picture by being put together or *composed* using composition rules. Composition rules are specified using the set of relationships, R. Each of the relations in this set are m-ary predicates and defined over the attribute values of the constituents of the picture fragment. If p_1 and p_2 are atoms with assignment properties and r is a binary relation over some subset of the attributes of p_1 and p_2, then $r(p_1, p_2)$ defines a picture fragment whose constituents are the atoms p_1 and p_2 and whose assigned properties are the relationship, r. If the fragment is named f then

$$f \leftarrow r(p_1, p_2) \tag{3.2}$$

is a composition rule. To enable multiple compositions this schema can be generalised to

$$f() \leftarrow r(p_1(), p_2()) \tag{3.3}$$

Narasimhan applies these ideas to the generation of FORTRAN characters with such primitives as $ST(P_i, P_j; P)$ and $HZ(P_i)$ which are straight and horizontal. Each primitive has certain regions for its attributes, for example *starting* region, *terminal* region and *middle region.* The relations used in the composition rules are then : *left, right, above, slightly above, slightly below.*

Examples :

$$'7' \leftarrow HZ(P_1) + ST(P_3, P_5)$$
$$'+' \leftarrow HZ(P_7) + VT(P_7)$$
$$'L' \leftarrow VT(P_1) + HZ(P_4)$$

3.1.3 A Picture Language Critique

These so called *picture languages* are inadequate for the description of general scenes since, like so many scene analysis applications, they are need-driven by particular applications. There are many more examples like these from many different problem domains [19, 20, 21] although this kind of descriptional schema has fallen out of fashion in recent years. This may be because of the fall of

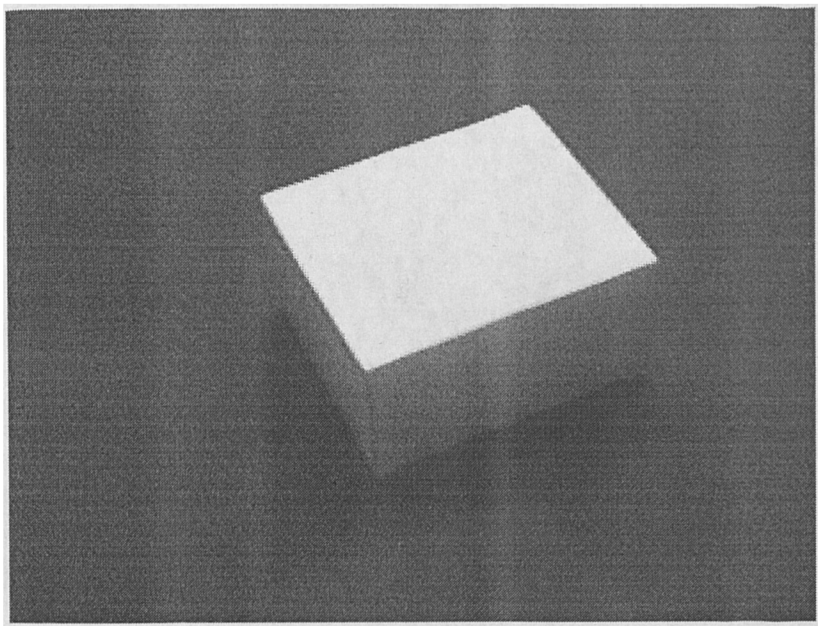


Figure 3.5: Toy World

interest in the problem of its analogue, natural language parsing, or the advent of "needs driven" vision research wherein target applications are the goal and elegance is sacrificed for brute force.

Waltz, Roberts and Guzman, like so many analysts have produced a system for classifying objects exclusively in a toy world of monochrome, high contrast, even shadows, geometric shapes and relatively easy segmentation. A typical toy world scene of this type is shown in fig.3.5. Amazingly, analysis of this kind is still performed with the justification that the analysis, once perfected, may be scaled up to the real world. The main problem with this is that even after a great number of test and re-implement iterations the application is still only good for analysing the toy-world since the real-world stubbornly refuses to scale down. Just how this descriptional scheme would perform on the image in fig.3.6 is another question entirely.

3.1.4 Rosenfeld

In his seminal text *Picture Languages* [162], Rosenfeld proposes an entirely digital geometry which has been extrapolated from cellular automata theory. He presents a collection of results based upon two-dimensional sequential and parallel (cellular) networks. At this time, of course, the results are purely theoretical and few cases were implemented. His picture grammar begins with a definition of an images as an array of discrete points (pixels), each of which has a set of neighbours. Lines and areas are then defined in terms of pixels and elements as are holes and other 2-dimensional structure. This was the first formal attempt at a picture language but was introduced simply to formalise the approach to 2D cellular automata.

Figure 3.6: A Pathological Waltz Classification Scene

3.1.5 A New Direction

What then are the unique elements which can unambiguously describe any scene ? The traditional idea of *features* is not able to do this because features do not possess the flexibility required to address scenes of very great complexity with the broad range of subject matter likely to be encountered, in effect they are 'too high a level' and not modular enough. What is required is a lowest-level data holder with the following properties:

- Must be able to contain generalised data. That is, structural and topological information as well as that collected by measurement.
- Must be able to cope with disparate data from a variety of sources.
- Must be able to expand and cope with possibly contradictory data.
- Must be able to be in some way pre-definable.
- Must be compilable. In this way data-holders may form a hierarchy.

A computational theory of information processing using such data holders, is proposed here, we call it a symbol world. A symbol world is a model of perception, symbol acquisition and semantic interpretation based on some ideas of human thought processes and internal structures [41,61,63,64,65,129,134]. Each of the symbols, for example, represents a construct inside the model proposed by Calvin [61], where selection based upon usefulness is also advocated. Inside the world only the following elements are allowed:

1. *Objects.* Objects are symbol storage elements analogous to objects in the real world. Objects are any conceptual entity which possesses information.

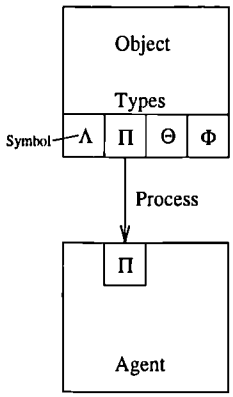


Figure 3.7: Symbol World

2. *Agents.* Agents are collectors of object symbols. These represent the embodiment of acquisition processes leading to ideas. Agents may structure and store symbols in any way which is suitable for retrieval, including generalisation.

3. *Symbols.* Symbols are low level information elements.

4. *Meta-symbols.* Meta-symbols are high order compound symbols which can be constructed using rules. They may be of different orders depending on their level of complexity.

5. *Types.* Types are assignable labels for groups of symbols with some property in common. These types may be assigned arbitrarily or learned.

6. *Processes.* A process is a method of symbol extraction used by an agent on an object. This is analogous to the perception processes humans use such as sight, touch etc.

It is also entirely plausible that these properties could include *motivations* or any of the ten fields proposed by Sloman and Beaudoin [129], in their tentative perceptual model of motive processing and attention, without noticeable redundancy.

Symbols

In order that rules can be formed symbols must possess properties that are:

- *Rankable* or able to be ordered in at least one way in order that they may be compared directly.
- *Assignments*, in order that they may be learned.
- *Rules*, to perform operations on rank information.

Agents

An agent then only needs the following processing capability :

- *Limited numbering ability* in order to assign ranks and prepare the rank information of symbols.

- *Comparison of types* to ensure that accumulation of ranks is correct and to remove redundancy when building meta-symbols.
- *Sorting according to rank* to ensure the rank hierarchy does not become disordered or incorrect.
- *Rule accumulation* for stereotyping and knowledge building.
- *Assignment accumulation* for data gathering before rule building or stereotyping.
- *Redundancy removal.* Once a rule covers accumulated data only special cases are needed and redundant data may be removed.

Symbol World Example

If an agent, by processes has derived the following symbols :

1. TYPE $\bigcirc$
2. TYPE $\star$
3. RANK $\bigcirc = 1$
4. RULE (TYPE,TYPE) + = ADD RANK
5. RANK $\star = 4$

Then not only would the logical step $\bigcirc + \bigcirc = 2$ be possible but so would the deduction of the entire positive $\bigcirc$ numbering system. The addition of assignment 4 means that the following are also possible deductions:

1. $\star + \star = 8$ etc.
2. $\bigcirc\bigcirc\bigcirc\bigcirc = \star$

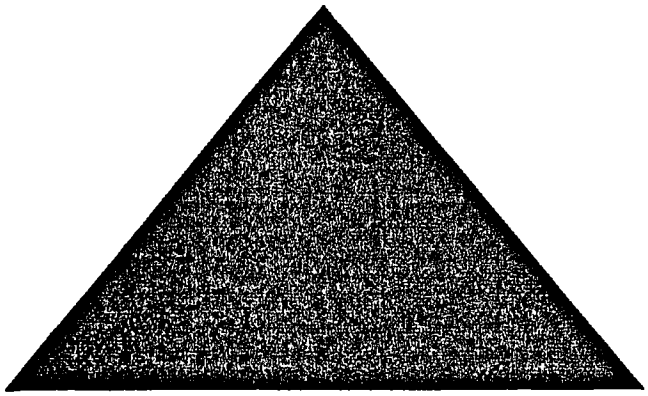
Without the assignment of TYPES it is possible to deduce item 2 above, that is because of absolute ranking in this symbol world deductions are possible without knowing whether $\bigcirc$ and $\star$ even belong to the same number system.

3.1.6 Picture Symbol World

It is proposed that the symbols required to describe scenes can be condensed into just two, line and colour. These two together form a first level meta-symbol called plane.

Line

A line in real space consists of an infinite number of indivisible points. It is rankable by virtue of some parameterisation, such as end-points, gradient, quadratic quotients or subtended angle.



Blue + Triangle
|
Line

Figure 3.8: Blue Triangle Symbols

Colour

A colour is normally associated with a vector in some colour space. An example is the (R,G,B) space as discussed in section 3.3. Colour and derivations from it such as saturation, chromaticity or hue (as defined in [1]) are rankable in several ways, for example shades of red may be ranked :

$$(255, 0, 0) > (200, 0, 0) > (172, 0, 0)$$

Example Picture Symbols

Two example objects are shown in figures 3.8 and 3.9. Processes by which these symbols may be extracted are discussed in chapter 5. Their descriptions are entirely in terms of our small symbol world. Note that in fig.3.9 the higher order meta-symbols such as face and mouth are used to indicate arbitrary labels for planes with particular properties: in the case of mouth, colour and line parameterisations; in the case of face, a collection of meta-symbols. As will be shown in chapter 6, each of these meta-symbols defines a local image context.

3.2 Pixelisation

Many different representations are used to describe our experience of colour, using different linear and non-linear combinations of primary colours. Any colour space can be represented by an n-dimensional vector space V^n with an orthonormal basis set $\{\underline{c}_1, \ldots, \underline{c}_n\}$ where each basis represents a vector direction in hue space. The vectors in this basis set may represent any of a group of different colour parameters, for example {R,G,B}, {X,Y,Z}, {Y,I,Q}, {L,a,b}, {U^*,V^*,W^*}. In general they are linear combinations of red, green and blue.

The human experience of colour, however, is bounded by physiological constraints since the three types of human cone pigment are maximally absorbent at just three distinct wavelengths[1] as described in [9],[10],[25]. These three wavelengths correspond approximately to the spectral values for red, green and blue [10]. Thus for humans the colour space is restricted to a three-dimensional vector space, V^3,

[1] 419nm, 496nm, 559nm

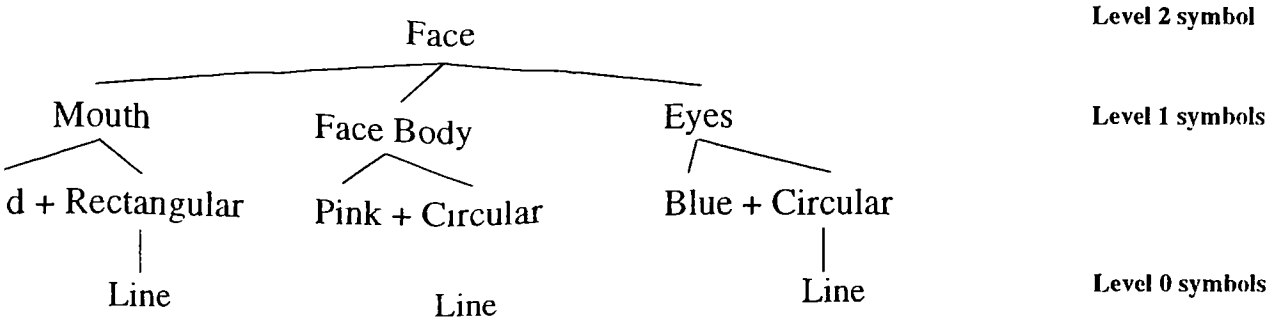


Figure 3.9: Face Symbols

with basis set $\{\underline{c}_1, \underline{c}_2, \underline{c}_3\}$. Furthermore, there is evidence to suggest that these detection cells have discrete activation levels so that the maximum number of colours perceivable by humans is around 9 million [11]. There is some wavelength variation (even among primates) [11] but three discrete wavelengths are invariably found.

Proposition 1 *A* pixel, $\underline{u}$, *is a vector representing a coloured spot in a scene. It is constructed by sampling and averaging the values for each of the vectors in the colour space basis set.*

Proposition 2 *A digitised* image, A, *is a representation of a real scene made up of pixels arranged in an* $(m+1) \times (n+1)$ *matrix.*

The methodology used to obtain these sampled colour spots and constructing the image matrix, pixelation, is well known and discussed in many texts, for example [1],[12]. Generally the method involves taking a regular rectangular array of input points from a real scene and generating an average hue and intensity for each point in the output array.

From definition 1, each pixel of A represents averaged values for each of the colour bases in the corresponding basis set and has the form:

$$\underline{u} = \{r\underline{c}_1, g\underline{c}_2, b\underline{c}_3\}$$

where r, g and b are scalars representing the intensity in the direction of each basis vector. If p bits are used in the sampling scheme, then a vector $\underline{u}$ has elements u_i such that

$$0 \leq u_i \leq 2^p - 1$$

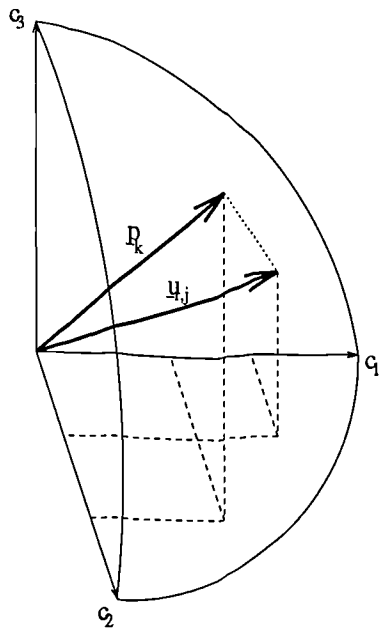


Figure 3.10: Colour Space

A then has the form

$$A = \begin{pmatrix} \underline{u}_{0,0} & \cdots & \underline{u}_{0,n} \\ \vdots & & \vdots \\ \underline{u}_{m,0} & \cdots & \underline{u}_{m,n} \end{pmatrix}$$

where $\underline{u}_{i,j}$ is the vector containing the red, green and blue values. For example if a pixel was pure bright red and sampled using an 8-bit sampling scheme, it would have a vector description $\underline{u}_{i,j} = (255, 0, 0)$

3.3 Image Segmentation

Proposition 3 *A* segmentation set *is any set of* ν *vectors,*

$$P = \{\underline{p}_1, \ldots, \underline{p}_\nu\} \text{ where } \underline{p}_i \in V^3 \text{ and } 0 \leq i \leq \nu$$

These vectors represent an exemplar set against which the image vectors are matched to form the segmentation. An example set of three vectors in an eight bit colour space might be the orthogonal set for red, green and blue, as shown in figure 3.10 :

$$\underline{p}_1 = (255, 0, 0)$$

$$\underline{p}_2 = (0, 255, 0)$$

$$\underline{p}_3 = (0, 0, 255)$$

The palette of colours for the image is then often chosen as the most frequently used 256 values from a range of 16.7 million colours. This can be done by colour-space partitioning, histogramming or similar techniques [25].

Proposition 4 *A* segmentation function, *f*, *is a function mapping vectors in an image, A, onto members of a segmentation set P :*

$$f : \underline{u}_{i,j} \mapsto \underline{p}_n \tag{3.4}$$

Essentially the segmentation function is the algorithm used to find the most efficient segmentation vectors in the segmentation set.

Proposition 5 *A* segmentation matrix, *Q*, *is an* $(m+1) \times (n+1)$ *matrix of elements* $\underline{q}_{i,j}$, *where* $\underline{q}_{i,j} \in P$.

The matrix Q is now a colour reduced *representation* of the image matrix A. Groups of pixels in the input image with similar colour characteristics will be mapped to the same vectors in the segmentation set and these will be placed in the segmentation matrix.

A segmentation matrix can also be used to produce false colourings and so illustrate which vectors in the input image have matched to which vectors in the segmentation set.

A vector in Q is generated from its corresponding vector in A by

$$\underline{q}_{i,j} = f(\underline{u}_{i,j}) = \underline{p}_n \quad if \quad \|\underline{u}_{i,j} - \underline{p}_n\| \leq \delta. \tag{3.5}$$

Where δ is some vector distance tolerance. The condition in (3.5) may also be defined as

$$\left\| \frac{\underline{u}_{i,j}}{|\underline{u}_{i,j}|} - \frac{\underline{p}_n}{|\underline{p}_n|} \right\| \leq \delta \tag{3.6}$$

This measure is a *hue* measure, which ignores the length, or brightness, of the vectors being compared and concentrates on the direction, or hue, of the vector in the colour space. If this measure is used it can eliminate faulty segmentation caused by quick changes in intensity, gradient effects or textured surfaces as described by Ohta [12]

No vector in A can generate a reference to more than one segmentation vector from P as long as δ can be set low enough to ensure that there is a one-to-one mapping. Once segmentation has taken place a segmentation matrix is formed as below. It can easily be seen that the definition in equation 3.4 can be extended to the matrix operation.

$$g : A \mapsto Q \tag{3.7}$$

where each element in A maps to the corresponding element in Q as defined in equation (3.4).

$$g(A) = Q = \begin{pmatrix} \underline{q}_{0,0} & \cdots & \underline{q}_{0,n} \\ \vdots & & \vdots \\ \underline{q}_{m,0} & \cdots & \underline{q}_{m,n} \end{pmatrix} \tag{3.8}$$

Each element in this segmentation matrix is then mapped to one of the vectors in the segmentation set.

Proposition 6 *An element of Q,* $\underline{q}_{i,j}$, *is said to be in* partition *n if* $\underline{q}_{i,j} = \underline{p}_n$

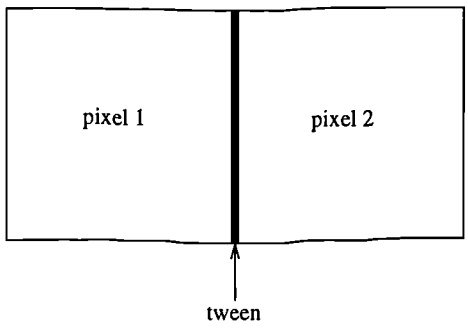


Figure 3.11: The Tween

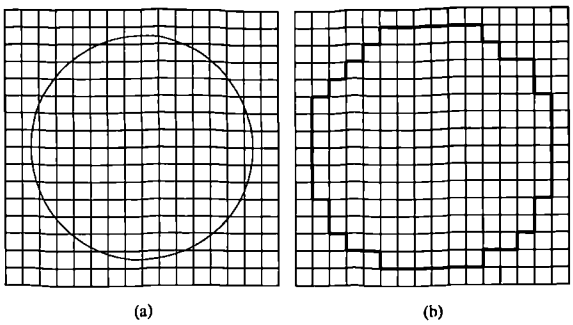


Figure 3.12: (a) Digitised circle and (b) Tween line set

This representation is relatively inefficient since for each vector in Q the three entries for each of its coordinates must be stored. A more efficient way of storing the segmentation matrix entries is to simply store an integer reference to that vector along with a look-up table of values. This would have the added benefit of providing a simple *colouring* showing which partition each of the elements in A belonged to. Thus we could give tween vectors in the form

$$\underline{t}_n = (\underline{q}_{i,j}, \underline{q}_{k,l}) \tag{3.9}$$

3.4 Well Formed Tween Sets

3.4.1 Tweens

The line descriptors are built up of *well formed tween sets* which are areas that lie between pixels rather than the pixels themselves (fig.3.11). The tween is particularly useful since the size of the segment is not artificially enlarged or reduced by adding a pixel boundary to it. This can be seen in fig.3.12 where a digitised circle can be seen. If the digitised bounding line is written outside the tween then the area is increased significantly, if it is written inside it decreases the area significantly.

3.4.2 Definition of Well Formed Tween Sets

There are three different groups of well formed tween sets, each describing a different kind of line.

1. A line consisting of a single tween.

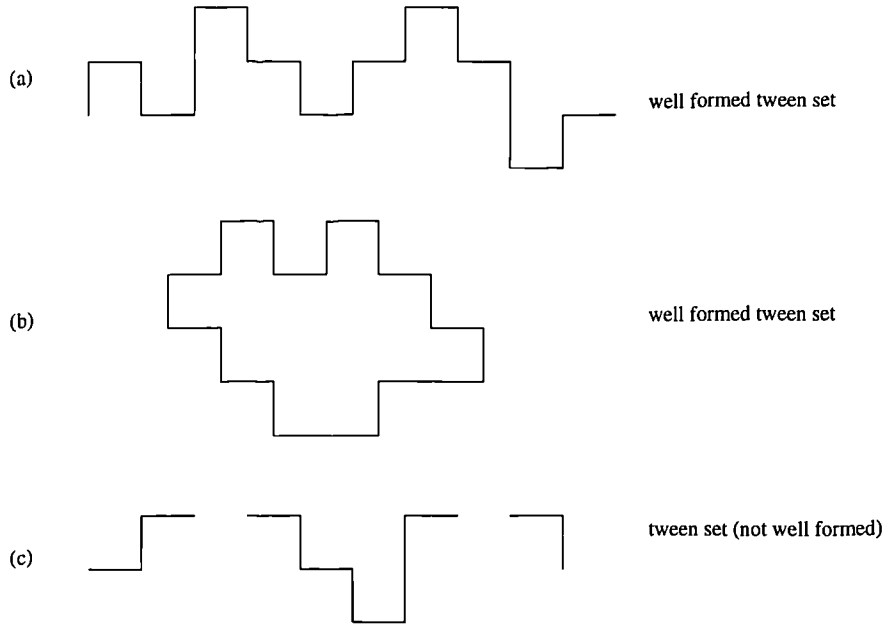


Figure 3.13: Examples of Tween Sets

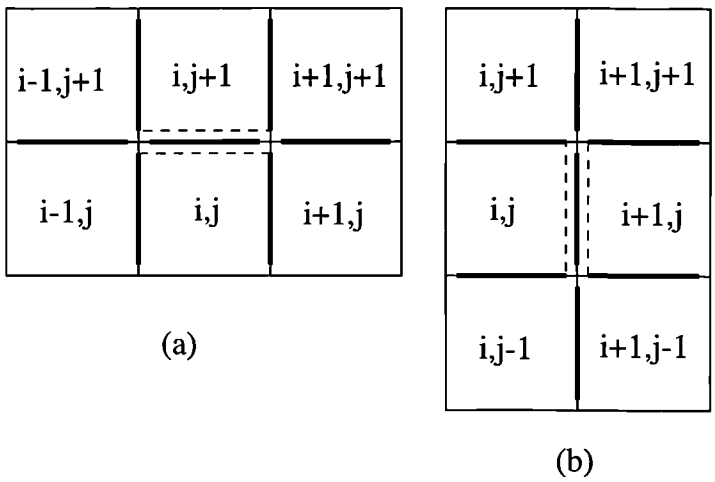


Figure 3.14: (a) Vertical and (b) Horizontal Tweens

2. A line which consists of a set of contiguous tweens, having two ends which are not joined, see fig. 3.13(a)

3. A closed line which consists of a set of contiguous tweens see fig.3.13(b)

It can be seen that the second and third groups are special cases of the first. Any segmentation can be described by a set of well formed tween sets. Any tween set which is not well formed does not represent a valid line, see fig.3.13(c).

Proposition 7 *A horizontal tween,* $\underline{t}_1 = (\underline{q}_{i,j}, \underline{q}_{i,j+1})$ *is* contiguous *to any of the six tweens* $\underline{t}_2$ *such that:*

1. $\underline{t}_2 = (\underline{q}_{i,j}, \underline{q}_{i+1,j})$
2. $\underline{t}_2 = (\underline{q}_{i+1,j}, \underline{q}_{i+1,j+1})$

3. $\underline{t}_2 = (\underline{q}_{i,j+1}, \underline{q}_{i+1,j+1})$

4. $\underline{t}_2 = (\underline{q}_{i,j}, \underline{q}_{i-1,j})$

5. $\underline{t}_2 = (\underline{q}_{i-1,j}, \underline{q}_{i-1,j+1})$

6. $\underline{t}_2 = (\underline{q}_{i,j+1}, \underline{q}_{i-1,j+1})$

A vertical tween, $t_1 = (\underline{q}_{i,j}, \underline{q}_{i+1,j})$ *is* contiguous *to any of the six tweens* t_2 *such that:*

1. $\underline{t}_2 = (\underline{q}_{i,j}, \underline{q}_{i,j+1})$

2. $\underline{t}_2 = (\underline{q}_{i,j}, \underline{q}_{i,j-1})$

3. $\underline{t}_2 = (\underline{q}_{i+1,j}, \underline{q}_{i+1,j+1})$

4. $\underline{t}_2 = (\underline{q}_{i+1,j}, \underline{q}_{i+1,j-1})$

5. $\underline{t}_2 = (\underline{q}_{i,j+1}, \underline{q}_{i+1,j+1})$

6. $\underline{t}_2 = (\underline{q}_{i,j-1}, \underline{q}_{i+1,j-1})$

Proposition 8 Tween connectivity $C(\underline{t}_i)$ *is the number of tweens contiguous to a given tween,* $\underline{t}_i$ *in a tween set. This can be between zero and six, from defn. 7 above.*

The idea of the *well-formed* tween set is now necessary in order to specify exactly what a line is in this description scheme.

Proposition 9 *A* well formed tween set *is a set of tweens with the following properties:*

1. *It is non-repeating, i.e each tween is unique.*

2. *Each tween is contiguous to at most two other tweens in the set.*

3. *Each tween has connectivity of at least one. This follows directly from the definition.*

4. *There are at most two tweens with connectivity one. These tweens are the termination points for the set. Again from the definition.*

3.5 Lines and Edges

Once a segmentation matrix has been derived from an image, the symbolic description still has to be formed. This is built using a set of primitives, in this case line and plane descriptors combined by the use of operators.

Proposition 10 *A* line *is a conceptual division between two partitions* i *and* j *and is the set of all vector pairs of elements from* Q*,*

$$l_{i,j} = \{\ldots, (\underline{q}_{a,b}, \underline{q}_{c,d}), \ldots\}$$

or

$$l_{i,j} = \{\ldots, \underline{t}_n, \ldots\}$$

where $\underline{q}_{a,b} = \underline{p}_i$ *and* $\underline{q}_{c,d} = \underline{p}_j$

This is the tween set that defines the edges of a segment and is the first primitive that is used to describe the segments in the image segmentation. All lines are well formed tween sets.

Proposition 11 *A* plane, *D, is an area in Q enclosed by a set of lines and having a descriptional colour vector from P associated with it :*

$$D^1 \equiv (\{l^1_{0,1}, l^2_{0,2}, \ldots\}, \underline{p}_d)$$

or

$$D^1 \equiv (L^1, \underline{p}_d)$$

Once all the segments in the image are described in this syntax then the image is fully described in terms of this syntax. A brief example, after a segmentation algorithm has completed the mapping from a 4×4 input image, is shown below.

Given the example segmentation matrix

$$Q = \begin{pmatrix} 1 & 1 & 1 & 2 \\ 1 & 1 & 2 & 2 \\ 1 & 2 & 2 & 3 \\ 2 & 2 & 3 & 3 \\ 2 & 3 & 3 & 3 \end{pmatrix}$$

a description can be formed as follows :

$$l^1_{1,2} = \{(\underline{q}_{2,0}, \underline{q}_{3,0}), (\underline{q}_{2,0}, \underline{q}_{2,1}), (\underline{q}_{1,1}, \underline{q}_{2,1}), (\underline{q}_{1,1}, \underline{q}_{1,2}), (\underline{q}_{0,2}, \underline{q}_{1,2}), (\underline{q}_{0,2}, \underline{q}_{0,3})\}$$

$$l^2_{2,3} = \{(\underline{q}_{3,1}, \underline{q}_{3,2}), (\underline{q}_{2,2}, \underline{q}_{3,2}), (\underline{q}_{2,2}, \underline{q}_{2,3}), (\underline{q}_{1,3}, \underline{q}_{2,3}), (\underline{q}_{1,3}, \underline{q}_{1,4}), (\underline{q}_{0,4}, \underline{q}_{1,4})\}$$

then

$$D^1 \equiv (\{l^1_{1,2}\}, \underline{p}_1)$$
$$D^2 \equiv (\{l^1_{1,2}, l^2_{2,3}\}, \underline{p}_2)$$
$$D^3 \equiv (\{l^2_{2,3}\}, \underline{p}_3)$$

It can be seen that once the line descriptors are complete, the plane descriptors are very simply constructed. A description of parameters derivable from this syntax is given in the next section.

3.6 Properties

3.6.1 Lines and Parameterisations

Since the line descriptors can describe any shape of line they are superior to simplified geometric line descriptors and parameterisations, as used in other work [2,3,4,5,6]. Although they have a much

greater size, there is no loss of information forced by fitting an artificially simple line to a complex border. Few lines in the real world are straight or easy to parameterise without severely compromising reality. Horizontal and vertical lines, for instance, are simply special cases of line descriptors with the properties stated below. It is also possible to incorporate them into a plane descriptor. Given a line descriptor :

$$l_{i,j} = \{\ldots, (\underline{q}_{a,b}, \underline{q}_{c,d}), \ldots\}$$

The line, or part of the line, is horizontal if for a subset of consecutive elements in that section there are elements such that a=c. Similarly a line, or part of a line, is vertical if for a subset of consecutive elements in that section there are elements such that b=d. Similar simplifications can be made for lines at different angles, circles and curves.

3.6.2 Topological Properties

Simple topological properties are easily derivable from the line and plane descriptor directly.

- *Euler number.* This is given by $E = C - H$, where C is the number of components in an image and H is the number of holes in it. In our syntax, if a plane is fully described by one line it must be a hole. If this line is included in the plane descriptor of another plane then the number of holes of that plane must be increased by one.
- *Perimeter length of segments.* This is simply computed at the time of description by counting the number of elements in each line descriptor included in the plane definition.
- *Geometric relationships between planes.* Topology graphs or tree relationships showing which planes are encapsulated and which planes border other planes need not necessarily be derived from the description, since they are *inherent* in it.

3.6.3 Basic Topological Expression Symbols

Relative topological positioning can only be used in logical arguments once symbols are defined to express properties. The most simple symbols are defined in this section.

$\ominus$ - Borders. For example $D^1 \ominus D^2$ means D^1 borders and is bordered by plane D^2

$\odot$ - Contains. For example $D^1 \odot D^2$ means plane D^1 contains plane D^2

3.6.4 Predicates

The $\ominus$ Predicate

Two planes, D^n and D^m border each other, written $D^n \ominus D^m$ iff $L^n \cap L^m \neq L^0$, where L^0 is the empty line set.

The $\odot$ Predicate

Plane D^n contains D^m, written $D^n \odot D^m$, if $D^n \ominus D^m$ and there is a single member l_i in the line set D^m and that line segment is in both line sets, $l_i \in L^n$ and $l_i \in L^m$

Also, if some plane D^m is contained in another plane, it is known as a *hole* and is described by a closed line segment.

3.6.5 Length Operators

Proposition 12 *The* brightness *of a palette colour vector is given by its Euclidean length in the vector space. If :*

$$\underline{p}_n = (r\underline{c}_1, g\underline{c}_2, b\underline{c}_2)$$

then its brightness is given by :

$$\| \underline{p}_n \| = \sqrt{r^2 + g^2 + b^2}$$

Proposition 13 *The length of a line segment,* $\| l_n \|$*, is the number of tweens it contains.*

Proposition 14 *The* line set length, $\| L_n \|$*, is the number of line segments it contains.*

Proposition 15 *The perimeter length of a plane* D^n*,* $P(D^n)$ *is the sum of the lengths of each of the line segments in its line set:*

$$P(D^n) = \sum_{\forall i, l_i \in L_n} \| l_i \|$$

Proposition 16 *The area of a plane is the sum of the pixels inside the bounding perimeter length and is given by:*

$$A(D^n)$$

3.6.6 Concatenation Operators

Heuristic operations may require the addition of two planes in order to form a third. This can be achieved by defining the necessary operations for addition.

Colour Concatenation, $(p, +)$

Two palette entry vectors, $\underline{p}_1$ and $\underline{p}_2$ may be concatenated in many different ways, for example by averaging their vector entries:

$$\underline{p}_1 = (r_1\underline{c}_1, g_1\underline{c}_2, b_1\underline{c}_2)$$

$$\underline{p}_2 = (r_1\underline{c}_1, g_1\underline{c}_2, b_1\underline{c}_2)$$

then

$$\underline{p}_1 + \underline{p}_2 = (\frac{r_1 + r_2}{2}\underline{c}_1, \frac{g_1 + g_2}{2}\underline{c}_2, \frac{b_1 + b_2}{2}\underline{c}_2)$$

Addition on the palette space is well defined since the addition and division of regular vectors with a basis in R^3 is well defined [62].

Line Segment Concatenation, $+(l_i, l_j)$

The concatenation of two line segments forms a single new line segment containing the tweens from each of the constituent segments.

$$l_i = \{t_i, \ldots, t_n\}$$

$$l_j = \{t_j, \ldots, t_m\}$$

$$l_i + l_j = \{t_i, \ldots, t_n, t_j, \ldots, t_m\}$$

Theorem 1 *The addition operation $+(l_i, l_j)$ is well defined if there is a tween in l_i which is contiguous with one in l_j.*

Proof 1 *By proposition 10, l_i and l_j are well formed. Any addition of l_i and l_j can take place only at one point. This point must be at a tween in each set with connectivity 1 from proposition 9. There can only be one tween in l_i contiguous to one tween in l_j and this must be the joining point. Once the sets are added the only change in connectivity is the change at the joining point, each tween has increased in connectivity to 2. Since this is acceptable by defn. 9 then the complete set $l_i + l_2$ is still well defined.*□

Plane and Line Set Concatenation, $(L, +)$ and $(D, +)$

When two planes are concatenated the line sets and colours of the planes must be concatenated to form one single bounding line set and one uniform colour description vector.

Two planes may be concatenated to form a single plane with gross properties that combine each of its constituents. This can only be achieved if either $D^1 \ominus D^2$ or $D^1 \odot D^2$. In the case shown in fig. 3.15, the concatenation procedure is described below:

$$D^1 \equiv (\{l_2, l_4, l_5, l_7, l_8, l_9, l_{12}\}, \underline{p}_1)$$

$$D^2 \equiv (\{l_3, l_4, l_6, l_7, l_8, l_{10}, l_{11}\}, \underline{p}_2)$$

At stage 2 (fig 3.16), each of the lines common to the line sets of D^1 and D^2 are removed and the planes are merged. Similarly, the plane colour vectors are concatenated, as described in section 3.6.6.

$$D^1 + D^2 \equiv (\{l_2, l_3, l_5, l_6, l_9, l_{10}, l_{11}, l_{12}\}, \underline{p}_1 + \underline{p}_2)$$

At stage 3 (fig 3.17), each of the line segments which coincident end tweens are merged to form the new bounding lines.

$$D^3 \equiv (\{l_{13}, l_{14}, l_{15}, l_{16}\}, \underline{p}_3)$$

where

$$l_{13} = l_2 + l_3$$

$$l_{14} = l_5 + l_6$$

$$l_{15} = l_9 + l_{10}$$

$$l_{16} = l_{11} + l_{12}$$

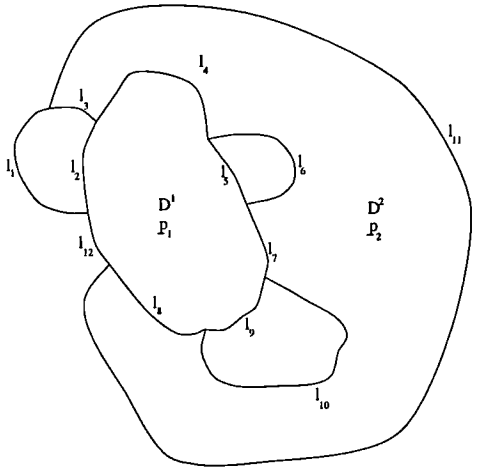


Figure 3.15: Concatenation of Planes. Stage 1

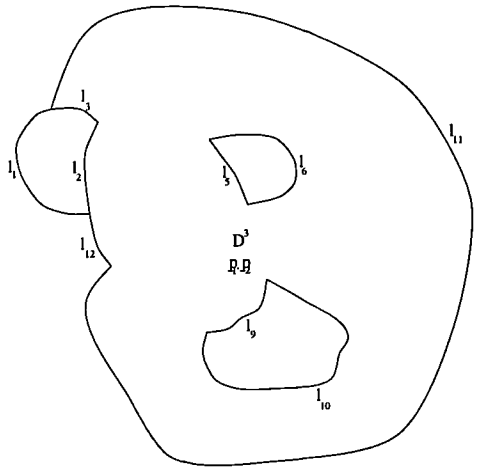


Figure 3.16: Concatenation of Planes. Stage 2

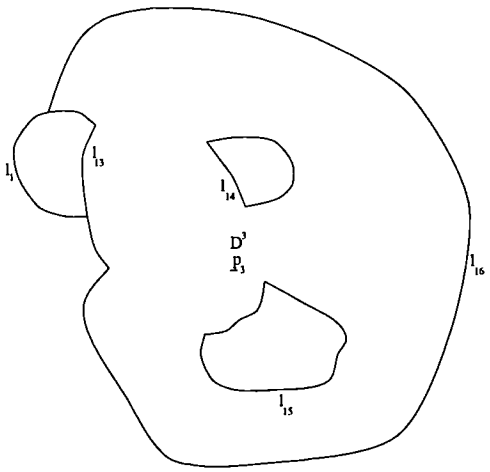


Figure 3.17: Concatenation of Planes. Stage 3

Theorem 2 *If two planes* $D^i \equiv (L^i, \underline{p}_i)$ *and* $D^j \equiv (L^j, \underline{p}_j)$ *fulfill* $D^i \ominus D^j$ *then the operation* $+(D^i, D^j)$ *is well defined.*

Proof 2 *Since*

$$D^i + D^j \equiv \{L^i + L^j, \underline{p}_i . \underline{p}_j\}$$

$$L^i = \{l_m, ...\} \text{ and } L^j = \{l_n, ...\}$$

then

$$L^i + L^j = L^i = \{l_m, ..., l_n, ...\}$$

which is a regular set operation. So $+(D^i, D^j)$ *is well defined.* □

3.7 Context

3.7.1 Definition

We now have a universe of discourse of the form proposed by Rosenfeld [162] but which has been significantly expanded and much more generalised (away from the concept of cellular automata). Elements are now much more free and less attached to their method of production.

An image, A now consists of a set of planes,

$$A = \{D^i, \underline{p}_i\} \tag{3.10}$$

where $\underline{p}_i \in P$, as defined in proposition 3.

Sub-groups of these planes form contexts, as shown in fig. 3.18. If all of the planes form a context, it is global, if a selection of planes in a neighbourhood form a context, it is local.

Proposition 17 *Two planes* D^n *and* D^m *are* related *if* $D^n \odot D^m$ *or* $D^n \ominus D^m$

A local contextual assessment can only take place among sets of related planes.

Proposition 18 *A* local context rule, T *is some relationship which holds between planes in a related plane set. For example* $T(O)$, *where* $O = \{D^n, ..., D^m\}$

Context rules may overlap in an image and affect each other incrementally, as shown in fig. 3.18. All of the local contexts are influenced by the global context, however.

3.7.2 Examples of Contexts

If we consider the example of a face in fig. 3.9, each element (eye, nose, body) is a local context and together they provide an overall context (face). This is the nature of the symbols that they represent. Each of these contexts has many other factors associated with it, such as geometric and topological arrangement and colouring.

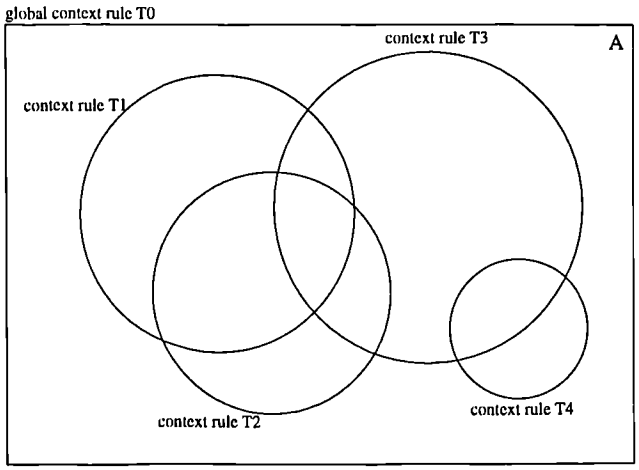


Figure 3.18: Local Contexts

An example of a global context is that of ambient light of a given colour. In light at sunset, for example, the planes which comprise a face will take on colours which lie outside of their normal set of hues. The impact of contexts is discussed further in chapter 7.

3.8 Tweens in 3 Dimensions

The idea of the tween supergrid is easily extensible to the third dimension by the inclusion of another orthogonal tween type, as shown in fig. 3.19. This then gives a supergrid which can occupy 3-D space and describe pixelized cubes in space. This idea is summarised briefly in this section.

3.8.1 3-D tweens

In the monocular expression of scenes so far pursued in this thesis, only 2-D tweens are needed. However, in applications using 3-D representations the idea of the 3-D tween and its hierarchy is necessary. The 3-D tween has three forms, as shown in 3.19. For these forms of tweens, the degree of connectivity (proposition 8) is higher but the idea of the well-formed tween set is substantially the same.

The syntactical representation is particularly useful if one is segmenting, for instance, range data from laser striping equipment (for example [140]) or constructing . A typical 3-D scene is shown in fig. 3.20, consisting of planes and lines. We can now also form a higher level descriptor, *volume.*

Proposition 19 *A* volume *is a section of space in 3 or more dimensions in the form:*

$$V^i \equiv \{D^m, \ldots, D^n\}$$

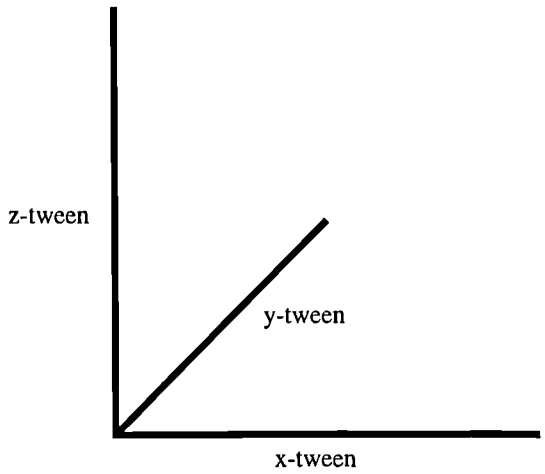


Figure 3.19: 3-D Tween

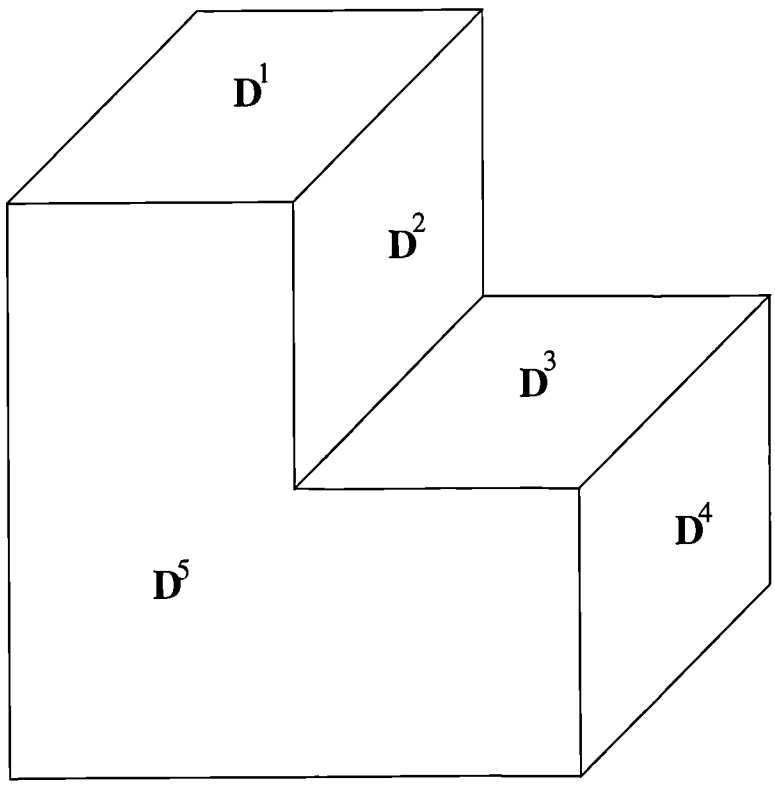


Figure 3.20: 3-D Tween Scene

3.9 Summary

The syntax described in this chapter formalises and provides a logical basis for the segmentation of colour images in the human visual cortex and to provide a language of description that may be used in further logical argument. The syntax has both descriptional and predicate parts which may be used for formulation, comparison and the construction of truth tables. These truth tables may then be used for identification of local and global context rules, as discussed in chapter 6.

A summary of the properties of the syntax:

- It provides an unambiguous scene description with a pre-defined information loss.
- It provides a set of description symbols that may be used in logical argument.
- It is relatively simple to derive topological properties as well as geometric relationships between planes.
- It is hierarchical and may be used in upward construction as well as downward de-construction.
- It provides a basis for the construction of local and global contexts.

Before rule construction, description or logical analysis can take place, images must first be divided into regions of similar colour. Processes (with reference section 3.1.3) for this must first be established which are analogous to those in the human visual cortex both in function and form. Such processes, at both high level and low level, are discussed both in the next chapter and in chapter 5.

Chapter 4

Segmentation and Models

4.1 Grey-Scale Segmentation

Before a scene, or *object assembly* can be represented symbolically, small patches from it must be identified and compiled into areas with similar characteristics. There are a wide range of techniques for achieving this which conventionally rely on the following processes:

- Derivation of joined areas or *regions*.
- Edge enhancement.
- Thresholding to find lines.
- Line thinning to form a binary image.
- Derivation of lines (object boundaries).
- Fitting geometric models to the derived structures.
- Attempting to derive information about lighting.
- Deriving information about orientation of individual elements.

A comprehensive treatment of current scene analysis techniques is found in Duda and Hart [4] and a brief description of methodology is found in [16]. There are many other texts which cover this subject.

4.2 The Region Based Approach versus Edge Detection

Region derivation is a crucially important element in the computational problem of scene description. Indeed, D'Zmura [44] has said that :

"to find the loci of responses that correspond to different objects, one must have already segmented the scene to establish which lights come from which objects. This begs the question of the purpose

of colour vision, which we believe plays an important role in the discrimination among objects and in their identification."

Region detection, (or rather the detection of boundaries between regions) is generally the detection of large gradients in colour reflectance between one area and another. Methods of boundary detection tend to be simple and often involve scanning an image looking for changes in brightness or colour gradients. A rectangular filter, or mask, is scanned across an image and has values such that, when convolved with the attributes from the image, produce maximum excitation when an edge or corner is encountered. Methods based on detecting gradient changes, such as those of Roberts [29], Prewitt [8] and Sobel [30] are all similar in nature and involve setting threshold values for excitation and applying different operator masks.

4.2.1 Edge Detection

Edge detectors are filters which emphasise high spatial frequency components. The idea of using operators giving large results in regions of rapidly changing brightness is generalised in these gradient operators. The size of the result depends not only on the location but also on the direction that the rate of change is calculated. A suitable compromise is to find the maximum rate of change and its direction. For a continuous image $f(x, y)$, this is easily found using the calculus:

$$\nabla f = \frac{\delta f}{\delta x} i + \frac{\delta f}{\delta y} j \tag{4.1}$$

This describes the vector that points in the direction of the maximum rate of change and with a magnitude equal to rate of change. In many ways, all gradient operators result from attempts to find digital approximations to this classical result. For example Robert's gradient, shown below, is given by:

$$\nabla a_{i,j} = \sqrt{(d - a_{i,j})^2 + (b - c)^2}$$

and the angle, θ is given by :

$$\theta = \frac{-\pi}{4} \tan^{-1} \left(\frac{b-c}{d-a} \right)$$

Immunity to noise is improved if a larger area is used to calculate a gradient. For example, the Prewitt and Sobel operators are defined on the 3×3 mask.

Prewitt's gradient is defined by:

$$S_x = (a_0 + a_3 + a_4) - (a_0 + a_7 + a_6)$$
$$S_y = (a_6 + a_5 + a_4) - (a_0 + a_1 + a_2)$$

and Sobel's gradient is defined by:

$$S_x = (a_2 + 2a_3 + a_4) - (a_0 + 2a_7 + a_6)$$
$$S_y = (a_6 + 2a_5 + a_4) - (a_0 + 2a_1 + a_2)$$

Irrespective of the operator used, the magnitude, M, is given by :

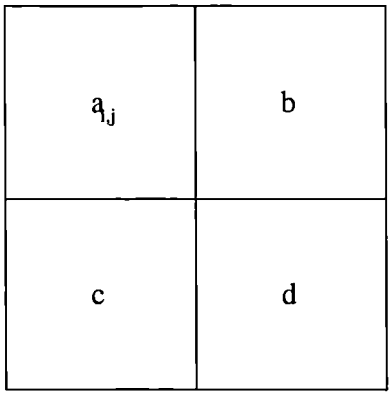


Figure 4.1: Robert's Gradient Operator

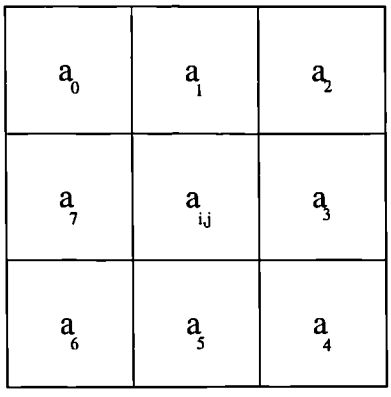


Figure 4.2: 3×3 Gradient Operator

$$M = \sqrt{S_x^2 + S_y^2}$$

and the angle, θ, is given by :

$$\theta = \tan^{-1}\left(\frac{S_x}{S_y}\right)$$

Newer methods, in use since the 1980s, are based on more complex operators and require much faster hardware. For example a Gaussian filter [31, 32] which designates certain pixels as *propagating* pixels. The intensity of each pixel in the vicinity of the propagating pixels is increased on the basis of its current intensity and its distance from the propagating pixel. The function of intensity is based on the familiar Gaussian (normal) curve, with the peak representing zero distance. A Gaussian filter is then applied with every pixel as a propagating pixel, thus all pixels bleed into their surrounding pixels. This has the effect of smoothing the image. Another filter, known as the Laplacian is then applied to the image which replaces each pixel with its second derivative relative to the surrounding pixels. These two methods are combined into a single filter called the "Laplacian of a Gaussian convolver" which, since it is shaped like a Mexican hat, is often called the sombrero filter. When this filter is applied, places where the Laplacian is zero are edges since it is there that values change from positive to negative. It has been shown [33] that this method is similar to the primary image processing in the human eye and that specialised edge detector cells exist in early (outer) layers of the visual cortex. However, it is unlikely that this can be achieved in real time on conventional hardware since it has been shown [34] that for edge detection in detail similar to that in human eye would require the order of 100 trillion computations per eye per second. Theoretical results on the use of Gaussian filters are discussed in [108]

One of the best known and most sophisticated edge detection methods is the Heuckel's operator [35, 36]. The edge detection models described previously are orientation specific to the proposed line whereas Heuckel's operator uses a general model of a line that includes not only its step size but its orientation. To make this possible, the circular neighbourhood of its edge is considered. The edge is defined by four parameters, b - the grey level on one side of the image, h - the size of the edge, r the radius of the neighbourhood and a - the angle of the edge in the circular neighbourhood. To use the algorithm, all that has to be done is to compute the values of the parameters which make the model fit the section of the image that falls in the circular neighbourhood. In practice this is very difficult [37]. Heuckel used an approach whereby a series expansion approximation was fitted (a radial Fourier series, in fact) to the image rather than the exact model. In practice it is thought to be the best that can be achieved as a raw edge detector. It is, however, very difficult and time consuming to apply and simpler schemes are often used in the line detection phase.

4.2.2 Critique

It is critical to address the motivation behind edge detection. If it is to be performed in order to fit geometric models, then there are the following assumptions underpinning the line finding process :

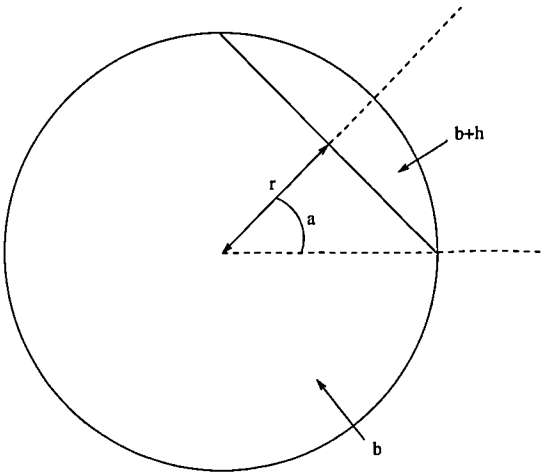


Figure 4.3: Heuckel Operator Neighbourhood

1. Geometric models of objects likely to be found already exist. Without this assumption there would be nothing to match found objects against, rendering the process meaningless. The image can only be understood in terms of a single symbol.

2. The problem domain has been reduced to a very simple level where the *only* objects which are findable are premodeled. This makes success relatively easy.

3. It is allowable to alter the analysis environment until a high level of success is achievable.

4. It is allowable to manually alter thresholding parameters in order to achieve a (subjectively) good segmentation.

Such filters are designed to emphasise high spatial frequencies and unfortunately this means that they also tend to emphasise noise. They also have a blurring effect upon the data, irrespective of the filter window size. The fitting of geometric models is the supposed endpoint of line-finding. Edge detection is therefore thought to be edge-enhancement. It is questionable to what extent it does, in fact, help. Operators create more interesting pictures for humans but maybe not for computers. Having edge-detected one is left with a 2-D array of values representing probabilities of edge being at a given position. No data reduction has been performed on the original image, since it is still in the same form, also no description has been performed of any kind. Now it must be decided where the edges lie in the probability array. More intelligent algorithms can use all the data present to decide upon the edges whereas less intelligent ones use a break-point method. This means that any probability over a given limit is mapped to one, all others are mapped to zero. This produces the binary image where all illumination effects have been removed. The lines produced must now be thinned and higher level algorithms applied to find structures. With the blurring and noise carried forward through this process, an accurate binary image from a real image becomes very unlikely. Also, with an image containing low contrast data if one sets an inappropriate threshold the binary image will be very sparse indeed. In general, what result is achieved is entirely dependent upon just how well the original threshold has been set. Too low and there will be large quantities of un-connected blobs in addition to the real edges.

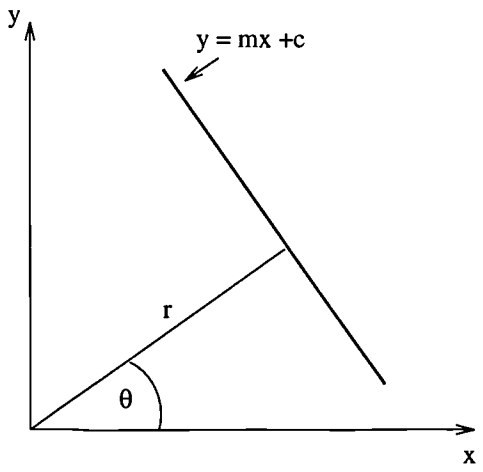


Figure 4.4: Hough Transform Quantisation

Too high and there will be little in the way of edges at all.

It should also be noted that the more ideal the source image then the easier it will be to set the thresholding level. The more realistic the source image then the harder it will be to produce binary images without noisy and blobby results.

Algorithms that do make intelligent use of the probability array include the Hough transform [90, 91]. This transform first quantises the entire dataset into (r, θ) notation, firstly to avoid infinite gradients and secondly in order to represent the hypothetical lines to be found by their normal, which passes through the coordinate origin, and the distance of the hypothetical line from the origin along this normal (fig 4.4). One can now represent all possible quantised lines in a table of (r, θ) values. This transform table can be filled by adding all the probability values along that hypothetical line in the probability space. Having performed this step, the transform table indicates the relative probability of a line being present in the edge detected image. Since the Hough transform considers all lines to be of infinite length it is vitally important to determine the exact endpoints of probable lines from the edge-detected image. The final result then is proposed to be all real lines from the edge-detected image.

There are however several problems with this scheme. The first is its sensitivity to noise. If several elements of noise lie along a given line, which is possible if the noise is random and probable if it is not, then the line through these points will be assumed to be present. Secondly, since the origin of the data arises from probability measurements there will be many possible lines going through the same collection of probability data points. Furthermore short lines and long lines going through the same points will be equally valid. This causes much confusion and one must arbitrarily remove lines below a given threshold based on whether they conform to suitable lengths or probabilities. Finally, and much more significantly the Hough transform can *only* detect straight lines. This may be overcome using the generalised version known as the Radon-transform [92, 93] where high dimensional transform tables are used. Unfortunately a phenomenal amount of both storage space and computational resources are required. In all of these filtering techniques the essential element is the user-correction feedback loop

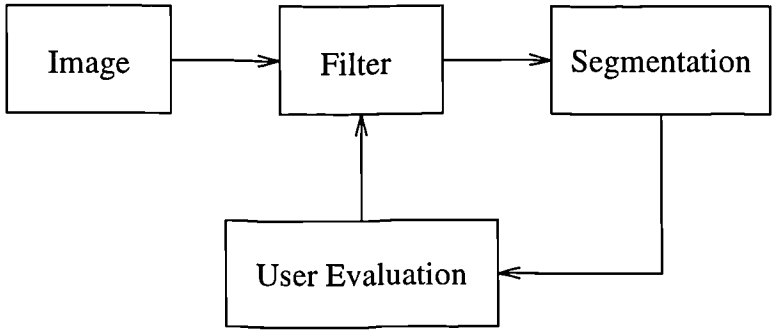


Figure 4.5: Feedback Correction Loop

(fig. 4.5)

Many different criteria for assessment of segmentation performance have been formulated, for example Abdou and Pratt [107], and De Micheli et al.[106]. These assessment metrics rely upon the following information:

1. Probability of false edges
2. Probability of missing edges
3. Error in the estimation of edge angles
4. Mean square distance of the edge estimate from the true edge
5. Tolerance of distorted edges such as corners and junctions.

The performance of the edge detector is then evaluated in two stages : count the number of false and missing edges and measure the variance for the estimated location and orientation. It is this entirely subjective efficiency evaluation which negates the reputation of edge detection as a means of model fitting in a general sense and brings into question the entire applicability of geometric model fitting.

Region based approaches include the statistical methods based on region growing and merging [97, 98]. In these methods pixels are chosen as seed points and then propagated depending upon statistical likelihood that bordering pixels belong to the same region. In monochrome images this is relatively straightforward, however when colour images are used the process is significantly more complex [99, 100].

4.3 The Primal Sketch and Marr's Theory of Human Vision

How humans actually form the internal representation of the scene they perceive is a particularly complex issue. Based on experimental work, the medical psychologist Poppelreuter [40] has proposed that visual perception is based on following schema. There is a series of part systems:

1. dark-light system

2. colour system

3. space, topology system

4. motion detection/tracking system

5. direction system

Each of these systems can be disordered and can work in relative isolation. No form or body perception is actually achieved at this low level. The conglomeration of the outputs of these functions is referred to as the *Gestalt* and its elements (*Gestalten*) are proposed to be formed on each of the following levels:

- *First level*: The visual field is simply an expanse without form.
- *Second level*: The visual field differentiates into diffuse illumination, left, right, etc.
- *Third level*: Size appears.
- *Fourth level*: Awareness of direction.
- *Fifth level*: Form differentiation.
- *Sixth level*: Several perceived image elements are seen simultaneously and discretely.
- *Seventh level*: Precise differentiation of straight, curved, etc.

4.3.1 The Primal Sketch

The work of Marr [24] is used as starting point in many modern approaches to human visual perception and scene analysis. He attempted to formalize the idea of the *primal sketch* which was widely discussed by perceptual psychologists in the early part of the 20th century. Stauffenberg [39], for instance, wrote as early as 1918:

"If we want to analyse the normally very fast and subliminal course of the process [of human vision], we may assume that the first overview, a sketchy outline, as it were, is obtained at diffuse attention, which integrates only the outstanding characteristics, such as form, size, colour, spatial position, into an unclear picture; this occasions the arousal of general concepts, leading to further analysis by attention under the direction of general categorical images."

The human (or animal) eye takes as its input an array of intensities from a large number of locations. These intensities come from light reflected from physical structures. These reflections represent the shapes of objects and surfaces, their orientations and distances from the viewer. Marr contends that it is the representation of these structures that is the function of early visual processing. This early form of representation is termed the *primal sketch*. It is the function of the primal sketch to make global structures evident from changes in intensity across an image. As has been shown, discontinuities in an image often suggest edges or boundaries. The *raw primal sketch* is a representation that consists of a messy statement of edges and blobs present, their locations and orientations etc. From this complex

representation, larger structures are found using grouping procedures. This description is termed the *full primal sketch.*

The full primal sketch represents the contours and textures in an image. This is the culmination of the early processing and gives the viewer-centred representation, termed the $2\frac{1}{2}D$ *sketch.* This is found by analysis of depth, motion, shading and structures found in the primal sketch. The $3D$ *sketch* or $3D$ *model representation* is then found by applying models to the stored set of objects.

4.3.2 The Marr-Hildreth Algorithm

The Marr-Hildreth algorithm [38] for finding the raw primal sketch begins by applying a number of different smoothing filters to an input image in order to produce different representations of it. These filters are simple versions of the Gaussian filter described above, with two or more different widths. The effect of these filters is to restrict the spatial frequencies in the resulting images, however the more the spatial frequencies are restricted the more spatial information is lost due to neighbourhood blurring. The optimal trade-off between frequency reduction and information loss is achieved using the Gaussian function, G. After passing through two or more of these filters, the grey-level intensities, I, are replaced by arrays of sets of values of $G * I$; the Gaussian weighted averages of the neighbouring values of I.

The second operation in the Marr-Hildreth operation is involved with location of intensity changes by differentiation. For economy, Marr and Hildreth consider only the second derivative. The crossing points are then found by taking the Laplacian (∇^2), which gives the sum of second derivatives taken in two different orthogonal directions. This is applied independently to each of the Gaussian filtered images, resulting in a set of arrays of $\nabla^2 G * I$. Zero crossings are then compiled into *zero-crossing* representations which are combined to form the primal sketch. No evidence of systematic testing is given for images but a set of asserted rules is given for combination methodology.

The first rule for combining inputs of the $\nabla^2 G$ filters is to put an *edge segment* into the primal sketch wherever zero crossing segments from adjacent filters match. If a zero-crossing segment in a wide channel is matched by two parallel ones in a narrow channel this is represented by a *bar* in the primal sketch and the ends of bars are represented by *terminations*, closed loops of edge segments are represented by *blobs.*

There is some physiological evidence that a scheme similar to the $\nabla^2 G$ filtering is actually taking place in the magnocellular mammalian visual pathways (as discussed in [24]). It also has the following advantages:

- The Gaussian is rotationally symmetric and will not bias segmentation in any particular way once it has been applied.
- Edges have components at both high and low frequencies and these will not be distorted by the removal of high frequencies by the Gaussian.
- There is a simple relationship between the parameterisation of a Gaussian and its smoothing effect.

- Large Gaussians can be implemented very efficiently due to the fact that Gaussian functions are algebraically separable. This means the degree of computation grows linearly instead of quadratically.

4.3.3 Colour Segmentation

Segmentation of colour images is problematic. In grey-scale images there is a well defined goal, an edge is defined as a discontinuity in the image and as such may be relatively easily found using a wide array of tools (as described above). This kind of discontinuity is precisely the one which is detected by the magnocellular layer and has given rise to the biologically motivated tools described in the literature.

The problem of segmentation of colour images is not quite so simple, however. The problem is initially ill defined on the colour image since:

- Coloured objects may consist of several patches of different colours.
- Different colours interact differently with different coloured light sources.

It is concluded [56] that coloured images should be thought of as consisting of the triple { hue, saturation, intensity } and the following operations performed on those three images:

- Grey level algorithms should be performed on the intensity component.
- Segmentation should be performed on the hue component.

4.4 A Model of the World as Local Contexts

4.4.1 Summary

Marr's primal sketch ideas focus on only one area of Poppelreuter's categorisation of the human visual system, the dark-light or contrast system. This corresponds to the form perception taking place in the magnocellular layer of the visual cortex, as discussed in section 2.3. Space and motion and direction are also discussed but in less detail and colour and topology are not discussed at all. In this section a context model is introduced as an extension to the idea of the primal sketch which amplifies the role of the parvocellular region and its interaction with the higher layers of the visual cortex.

4.4.2 Biological Motivation

The biological motivation for this model is strong. At the farthest extreme from the eye is the visual memory [134]. In this area there exists tiny pattern recognition templates often referred to as icons which are associatively linked. Thus they can be activated or potentiated by different routes. Firstly, by the process of pattern matching with incoming signals and secondly by the activation of other templates through associative learning. Interestingly, this is the area of the visual cortex which shows least retinotopic ordering or structure [135]. A small section of this scheme is shown in fig. 4.6. and illustrates the two types of connections found :

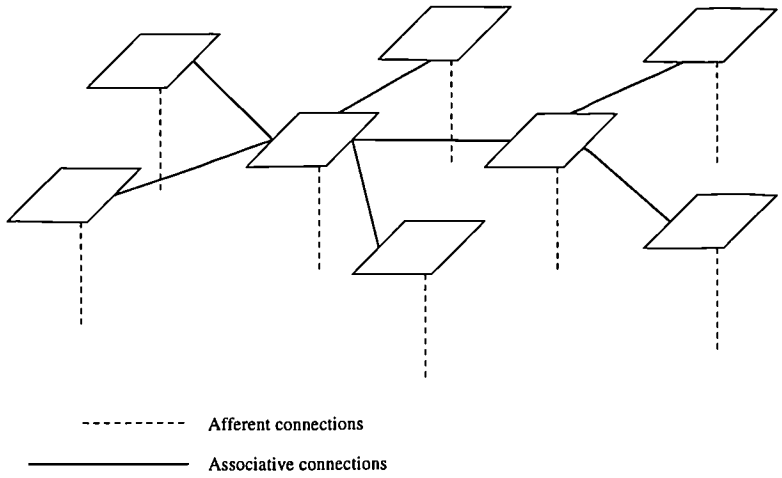


Figure 4.6: Visual Memory Schematic

- *External*, or afferent connections, where information enters the space.
- *Internal*, or associative from where pressures to consolidate and order the information is embodied.

It is asserted by Nakayama [134] that the experience of seeing is dependent upon the activation of these iconic nodes and without such nodes visual perception cannot exist. The essence of this theory is that although these areas contain little information, they capture the essential properties of some fragment of an image in very few 'bits'.

The most surprising part of this assertion is this low information content in these templates or icons, since it is known fact that the visual memory is made up of an associative network of elements [136]. Plausibility for the theory comes from experiments conducted in reading [138], which is a procedure where systematic attention is applied along a line of print. Rayner [137] constrained the amount of intelligible information provided by following eye movements around the text and then replacing all but a few letters around the fixation points with meaningless symbols. His finding was that if one made more than around 5 letters visible then reading is not substantially improved. This, coupled with the work of Legge et al. [138], showing that reading did not improve where more than 3.5 letters are visible, shows that the visual system cannot process more than a small number of letters at a time. The implied icon size from these tests is between 100 and 125 'pixels'.

4.4.3 Definition

Segmentation in the form-differentiating levels of the human vision system is vital and must be performed quickly and accurately. For this reason the cells performing it have become greatly specialized. However, segmentation in the colour-differentiating levels is much less developed and the detection of certain types of edges is lacking in acuity. Topology, however, is preserved and the first tentative steps in visual recognition after birth are almost entirely colour/topology based [163]. This also is useful in the case of failing eyesight and explains how fair perception is still achievable even with greatly

reduced visual acuity [163]. The finding and analysis of the *dominant* colours and planes is the vital element of this whole parvocellular layer and its attendant post processors.

A colour context model closely following the functions of the human visual cortex requires the following properties:

- *Speed.* The parvocellular layer is not as fast as the magnocellular layer but is still very quick in neurological terms.
- *Parallelism.* Obviously the system is parallel in nature although it is basically synchronous possessing an overall timing mechanism analogous to a system clock [68].
- *Colour differentiation* or colour store. This is to aid in narrowing the probable connections and identifying possible regions as well as possibly identifying objects. This would be analogous to the areas found in experimentation by both Zeki [72], Perrett et al.[74]. There is also evidence to suggest that colour is used to identify objects exclusively in some animals [75].
- *Topology Construction.* This is also a requirement for rotational invariance which exists in human vision.
- *Scale Invariance.* This is a useful function of human vision. It reduces the need for large and complex models of objects, allowing for a more general model.
- *Hierarchical Symbolic Output.* This is required in order to represent the thought space as discussed in section 2.2.

The model above is scalable in that at a very low level all of the main global information is still present (a global or primal model) and at a very high level all of the local specific information is present (a local or specific model).

4.4.4 Development

The kind of model outlined above is difficult to develop from first principles because a staggering level of complexity must be avoided. To return to the idea of a symbol world (section 3.1.3) it can be seen that the hierarchical aspects of the system are easily applicable to the development of a general contextual model. If the primary elements or *types* of scene are developed as high order meta-symbols then elements in that scene may be thought of as lower level meta-symbols and so on until line and plane level is reached (primary symbols). This means that a hierarchy of models will then be produced with greater and greater specificity where the lowest level is a model of *any* scene and the highest level is the exact scene which is analysed. In terms of context, this means that the highest level meta-symbol is the global context, with each of the lower order symbols representing local contexts. The interaction of these contexts then forms the scene.

Development of a global context then consists of building it probabilistically from many specific examples of the kind of scene to be modeled. This can be done by analysing the *nature* of the important

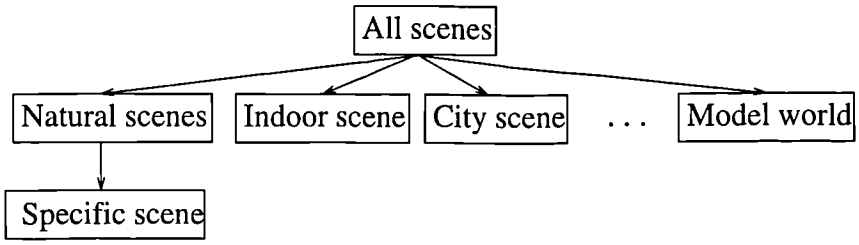


Figure 4.7: Model Hierarchy

elements in the scenes and their relationships to each other, their type, position, neighbours, connections and so on. The issue of which are and are not important, however, is not a trivial one.

To begin with, let us consider a general scene consisting of N planes of varying sizes which may be connected in any number of ways. If we consider that the largest planes are the most important then it is a simple matter to deduce their connectivity (even if separated by smaller planes) and to glean some idea of the general layout of the scene. We may also consider that brightness, saturation or any other rankable attributes are the most important and order them in that way instead. The leading planes in this ranking we term *object* planes, in the sense that deductions will be made using these planes as a base rather than them being the basis for object construction.

Once we have constructed an object plane set we may, using heuristics and search methods construct many different local interpretations of the entire plane set. What we are interested in is how two sets may be compared and a generalisation formed to match against further data against. In this case we are considering :

- *Colour range.* This will give us information to form the colour context of the scene.
- *Areas of object planes.* These areas will indicate relative importance of planes in the scene.
- *Degree of fragmentation.* Fragmentation is a good indicator of both sampling noise and textures present.
- *Topology of object planes.* This is a good indicator of the relationships between objects in the scene and their relationship to the background.
- *Connections of object planes to smaller planes.* Shape from shading assessments are possible from this information.

Each of these elements may then be tabulated to form the general model. It may be noted that the intermediate symbols constructed by examining the topology may map to specific objects in the scene, although this is not necessarily the case. Details of general model derivation are discussed in section 6.3.

4.4.5 Example Colour Primal Model

Examples of outdoor natural scenes are shown in fig. 4.9 and fig. 4.10. Many of the elements in these two scenes are similar and are immediately recognisable, for example sky, foliage, clouds and grass.

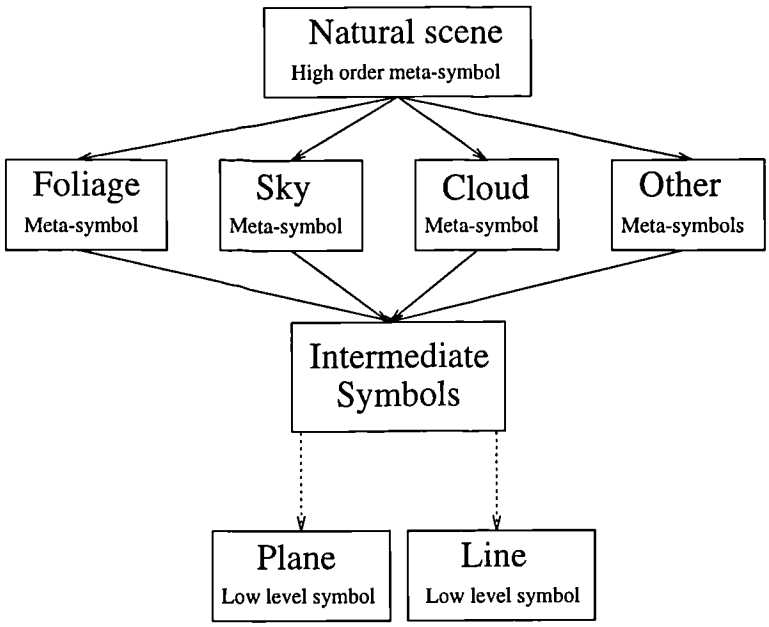


Figure 4.8: Natural Model Meta-Symbols

Shown in the following figures (fig. 4.11 to 4.13) is the development of a simple general colour model which incorporates both clouds and low level shading. If this were applied to either of the two example images we would expect the fit to be very good.

In the first model (fig. 4.11) only two planes are represented, one for sky and one for foliage. They are topologically connected with blue above green. In figures 4.12 and 4.13 a cloud element is added which may be topologically connected to both the sky plane and the foliage plane or just the sky plane. It may not, however, be connected only to the foliage plane. In fig. 4.14 a more complex model is generated with varying levels of shading to represent a more realistic model. In this case the cloud plane may be topologically connected to any or all of the sky planes but only one of the foliage planes.

4.5 Summary

Edge detection in order to find bounding lines causes blurring in the image, re-inforcement of noise and is time consuming even its most simple application. In the worst case, we may not even find the edges we are interested in. There is also a slightly more subtle problem when segmentation is used as a tool for the understanding of an image. That is the nature of the *expected* outcome. In the parameter setting, evaluation and re-setting loop, one is superimposing the expected result upon the outcome since a *good* detection phase is purely subjective. Region derivation by growing and/or merging techniques, however, may be used without previous specific information and the setting of parameters. To what extent this may be used without some explicit knowledge base, however is a moot point. It is believed that every aspect of perception uses domain specific information extensively [94]. Explicit application of knowledge for segmentation has been addressed in many systems [101, 102, 103, 104] but without the generation of a general system model.

Figure 4.9: Natural Scene

Figure 4.10: More Complex Natural Scene

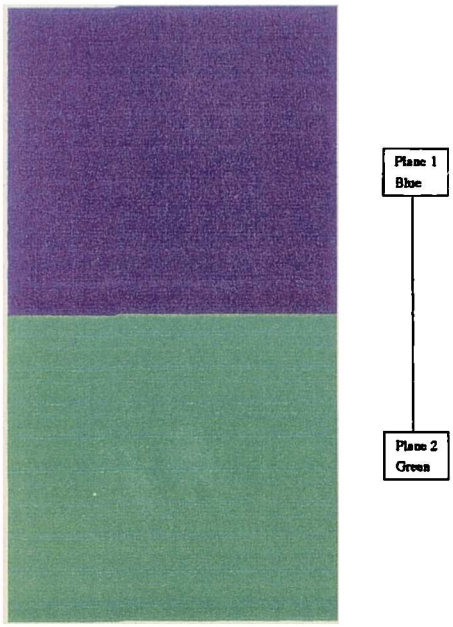


Figure 4.11: Simple Natural Scene Model

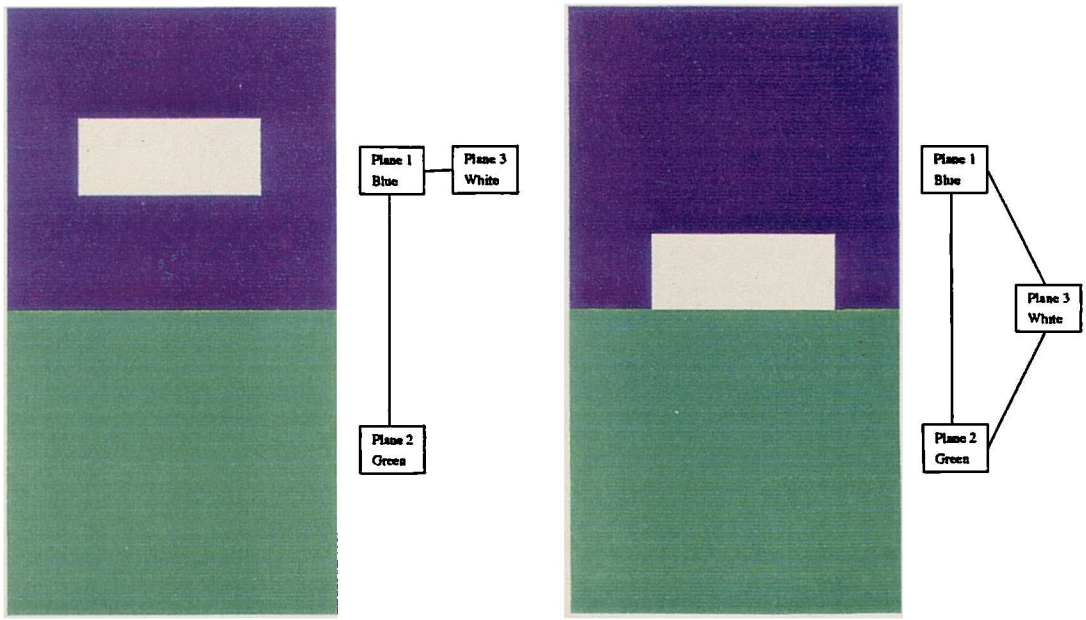


Figure 4.12: Simple Natural Scene Model Incorporating Cloud

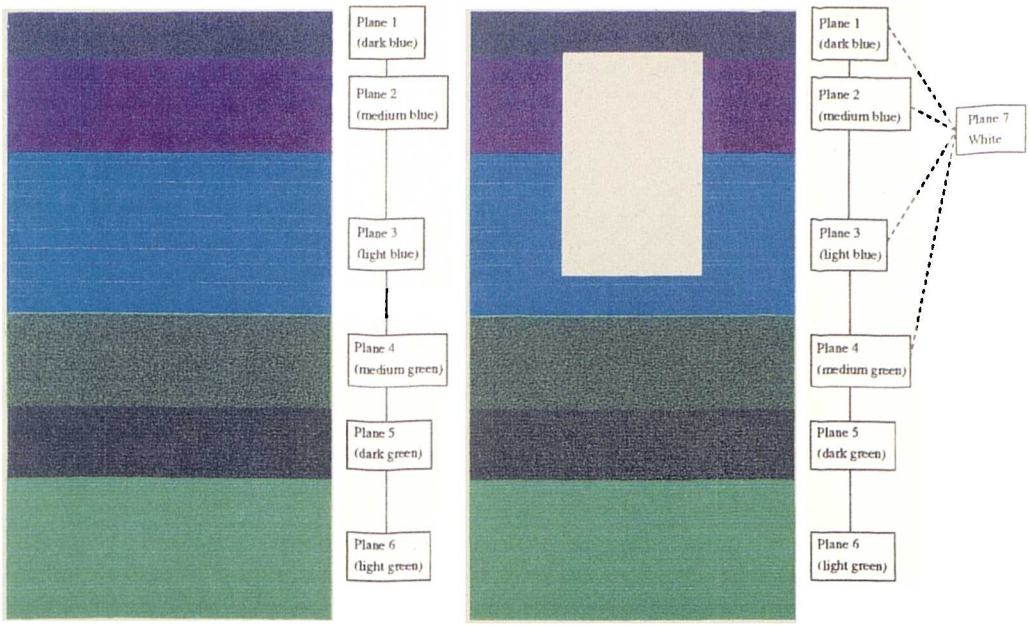


Figure 4.13: Simple Natural Scene Model with Shading and Cloud

Whilst it is important to segment the image into regions bearing similar properties it is unlikely that edge-detecting to produce a binary line image will generate any information which would be lost using the most simple gradient detecting algorithm. Expending significant time on edge detection per se is then avoided and processing is back-weighted, as it seems to be in the human visual cortex.

For true understanding of images more subtle structural information is required. Marr's theory and implementation of the full primal sketch has gone some way towards a true emulation of part of human vision. However, there are many missing elements in Poppelreuter's scheme which still must be addressed. The most important of these is colour. In this thesis, a colour context model is proposed, based upon the part of the geniculate which was neglected by Marr, the parvocellular layer (as described in section 2.3). This layer, although not as well evolved to deal with form perception, is a vital component for the derivation of colour patches and topology at all levels. This kind of size scalability and the rotation and size invariance of topology is exploited in the model.

In the next chapter details of two pipelined parallel networks are presented which can quickly derive the elements required for the model. In subsequent chapters, the application of these networks is described with some suggestions for further work and directions.

Chapter 5

Model Networks

The biological significance of splitting colour and form was discussed in section 2.3. This division is the primary motivation for the construction of the networks described in this chapter. It must be noted however that form has closely followed function in that the networks have used only mechanisms which are present in neural processes [76], namely the modal types:

- rhythmic output
- synaptic input/output with either brief or prolonged signal
- graded response output to a given input
- impulse output
- all or nothing (or latch)

Some of these types bear a close resemblance to standard electronic components, which is not surprising since the information transport used in the brain is electrochemical in nature.

The ICEnet (Independent Computing Element network) has been constructed to emulate the first part of the parvocellular mechanism and the interblob region of the visual cortex. It is in this region that segmentation of the image into similar hue areas takes place. The topology, size and nature of these areas are then passed on and merged with the magnocellular layer information later in the relevant processing areas of the brain (V1 and related areas as discussed in section 2.3).

5.1 Earlier Work

Region growing, using merge or split techniques, is widely used for image segmentation and there is a great deal of literature regarding it (for example [97, 98]). The goal in region growing is to expand from a set of seed points in an image until the boundaries of each region are reached. For an image of reasonable size (say 320 pixels by 200 pixels), an exhaustive search of all possible segmentations is a practical impossibility since it is an NP-hard problem [12]. Modern statistical methods for image

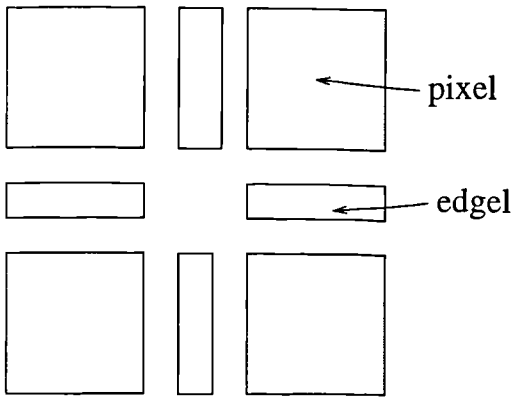


Figure 5.1: The Edgel

segmentation such as the Karhunen-Loeve often require several parameters to be set by the user as well as a careful choice of input data representation.

Recent work [13,14,86,87,88,89] has shown that neural networks can be used for fast segmentation of monochrome images. Booth [15] has also constructed a simple high-speed binary segmenting chip design for use as part of a passive vision system. The output from these segmenting algorithms is either in the form of a connected corner list (CCL) or a simple chain-coded boundary line. The segmentation is performed by small computing elements called *edgels* which compare pixels values and then form an activation. This activation is then compared to a cut-off level and the edgel is considered on if it is above this level.

Booth has taken this scheme one level further and added another computing element, the *boundel* which is a weighting layer which forms a segmentation. If a boundel is stimulated then it sends a signal along any edgel that is in an on state. Therefore any boundel which is activated must be in the same segment.

The network itself is constructed as a series of two dimensional arrays of nodes, each of which feeds forward to the next. The first layer contains an array of $N \times M$ light sensitive components which propagate a signal whose strength is proportional to the intensity of light falling on them. The second layer contains a collection of *boundel stimulating cells*. These cells are so named because their outputs directly control the cells of the next layer, which is the *boundel* layer (fig. 5.2). The stimulating cells work on two input stimuli from the two adjacent light sensitive cells in the first layer. They then perform a signal strength comparison. If the relative difference between the two signals is greater than some limit then the cell has detected a regional discontinuity and the cell output reflects this. The discontinuity signal is then propagated to the next layer.

This next layer consists of an arrangement of three different types of cells: boundels; boundel nodes; and stimulus-centres. These are illustrated in fig.5.2. The most functionally complex component in this layer is the boundel. The boundel is controlled directly by the boundel-stimulating layer associated with it and its behaviour is different depending upon whether an area of discontinuity has been detected or not. The change in operational characteristics of this element give it its name (boundel = boundary element). The boundel acts as a path for signals between two stimulating centres. If it lies next to

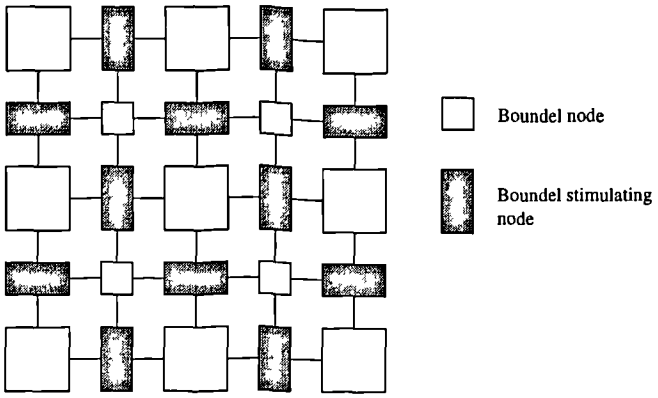


Figure 5.2: The Arrangement of Boundel Stimulating Cells

an edge it will prevent the passage of any signal from one stimulus-centre to another. In addition, if a signal is received from an adjacent stimulus-centre and the boundel is underneath an edge then this signal is used to drive an output from the boundel to the next layer. If the boundel does not lie below a discontinuity then it simply acts as a direct connection between the adjacent stimulus centres.

The stimulus-centres are cells which when triggered try and stimulate further stimulus-centres by sending a signal to the associated boundel. Thus region growing can be achieved by simply activating one stimulus and observing which other stimulus-centres become activated. Thus a complete image can be segmented into distinct regions. Also, since boundels situated over discontinuities propagate signals down the network these represent the edges in the image. These edges also have spatial information associated with them that indicates which region the edge is associated with. Thus, region growing and edge detection are performed in parallel.

5.1.1 Critique

Although a theoretical version of Booth's segmenter has proved to be reasonably fast, the quality of the results produced are mixed. The passive imaging component of the hardware is capable only of resolving binary images. This is useful only for the simplest applications involving pixel matching and some flat-bed applications, as was discussed [15]. Obviously the next step forward is the expansion to colour, either true 24-bit colour or 8-bit pseudocolour.

The only output from this network is a connected corner list (as described in [162]) which is fine for producing monochrome segmentation images but not particularly useful for further computation. This would be especially true in very complex and coloured images, primarily because no information about the planes themselves is included, all we know is that there is a division between them. Because of this, we could never know the plane areas, for instance. These problems stem from the fact that the analysis domain for the scheme falls between an image analysis and scene description but cannot fulfill either. The idea is sound but lacks both neurophysiological motivation or specific target problem domain since it springs from a cellular automata base.

In order to produce a better form of neural segmentation, a device capable of full colour segmentation has been designed. The next section describes the new algorithm for plane detection using functionally

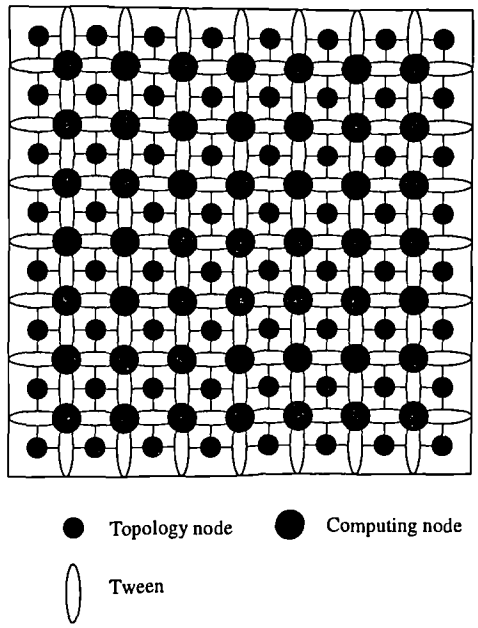


Figure 5.3: ICEnet Construction

complex boundary elements, with results of its application to several simple scenes. Output in the form of topological maps and text-based description, as well as colour images and segmentation images are also included.

5.1.2 Description of ICEnet

The ICEnet construction, shown in fig. 5.3 has been designed to address all of the outstanding problems with Booth's neural segmenter and to provide a computationally inexpensive segmentation. There are intensity measuring nodes connected directly to the pixels in the image. Between every intensity measuring node in the horizontal and vertical direction there is a boundel-stimulating node which is a neuron that can actively make comparisons between them. This is analogous to the primary parvocellular layer which segments purely to find edges regions which contain similar colours. Once again, the specifics of the geometric model are relaxed in order to attempt to understand which areas are related and in what ways.

The boundels themselves have been increased in complexity, however, and can perform a higher degree of computation. Since they no longer perform an autonomous stimulation and reaction they have been renamed computing nodes. The transfer function determining the state of the boundel-stimulating nodes, has been changed from boolean to continuous in order that further computation can be carried out. Since they no longer react in a binary style and are elements of an 'in-between' or supergrid they have been renamed *tweens* as described in section 3.4.2.

5.1.3 Algorithm, Coding Strategy and Implementation

The code to perform this emulation was heavily object oriented and it is for this reason that the algorithm can easily be employed in parallel with a fairly low grain size. Each of the computing

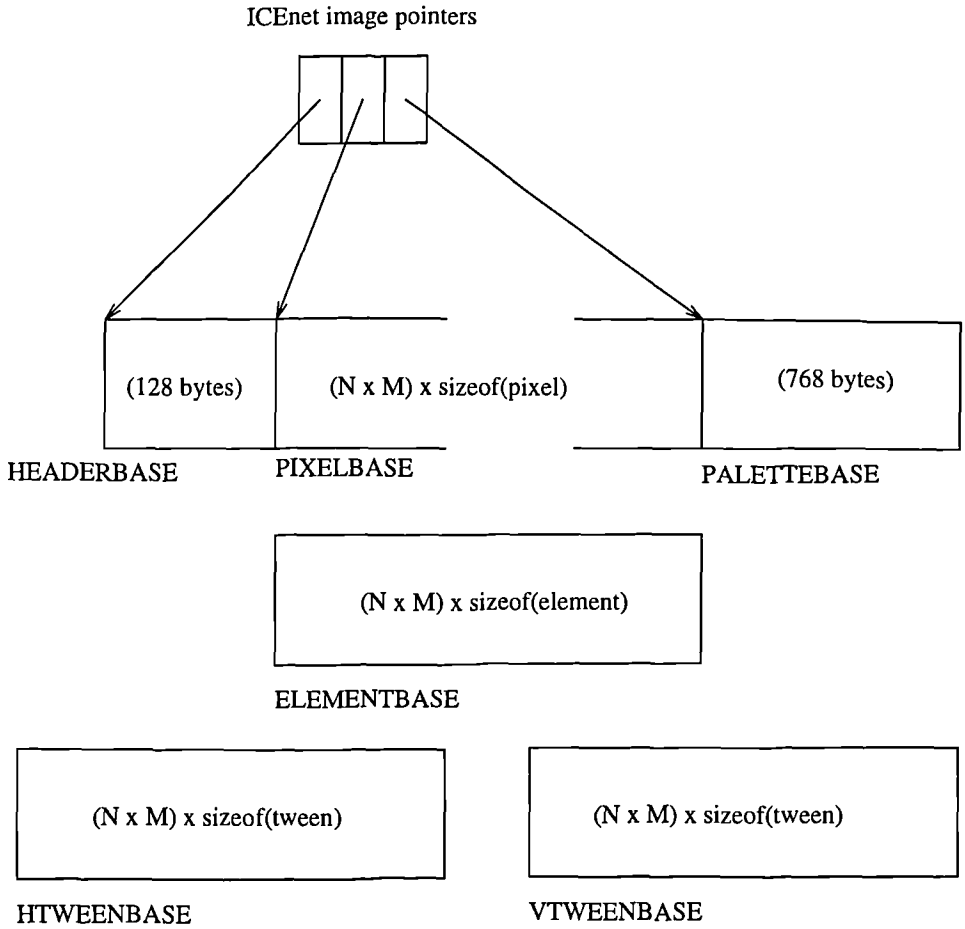


Figure 5.4: ICEnet Memory Storage

elements is self sufficient and can perform its small area of segmentation independently of the others provided it has access to the image data for its own area.

Before the algorithm is performed the following global values are set:

1. Gaussian smoothing kernel size.
2. Number of exemplars in the exemplar set or percentage exact coverage, discussed in section 5.1.4
3. Method of segmentation, by normalised hue or (R,G,B) exact colour reference
4. Smoothing level for the image

ICEnet Object Instantiation, Memory and Information Passing

The network has control over its own attached objects and is capable of creating them as required (fig. 5.5a). These objects may also be made persistent if required. The general form of instantiation is described by the pseudocode below. The user first selects the segmentation method, either by histogramming or by thresholding. Appropriate threshold parameter values may then be passed. These should be set as low as possible by the user, as discussed in 4.2.1-4.2.3. The suggested number of exemplars is also required if histogramming is used but is not vitally important, as will be discussed in section 5.1.4.

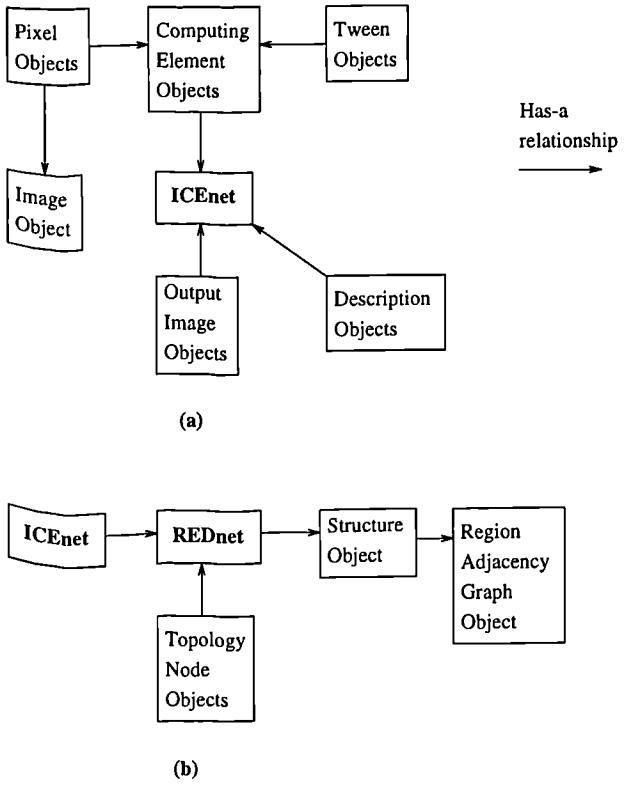


Figure 5.5: ICEnet Algorithm Construction

```
    begin CONTROL
    Read segmentation METHOD from user
    Read SOURCE image from user
    if METHOD = HISTOGRAM
        read HISTOGRAM data
    else if METHOD = ORDINARY
        read ORDINARY data
    endif
    Instantiate image object
Pass SOURCE to image object
        Pass METHOD, HEADERBASE,
        PIXELBASE and PALETTEBASE to image object
    Instantiate ICEnet object
        Instantiate N x M computing elements
            Pass PIXELBASE and TWEENBASE based values
            to each computing element
    Send PROCESS signal to computing elements
    Instantiate output image objects
    Instantiate REDnet
        Pass PIXELBASE and TWEENBASE to REDnet object
```

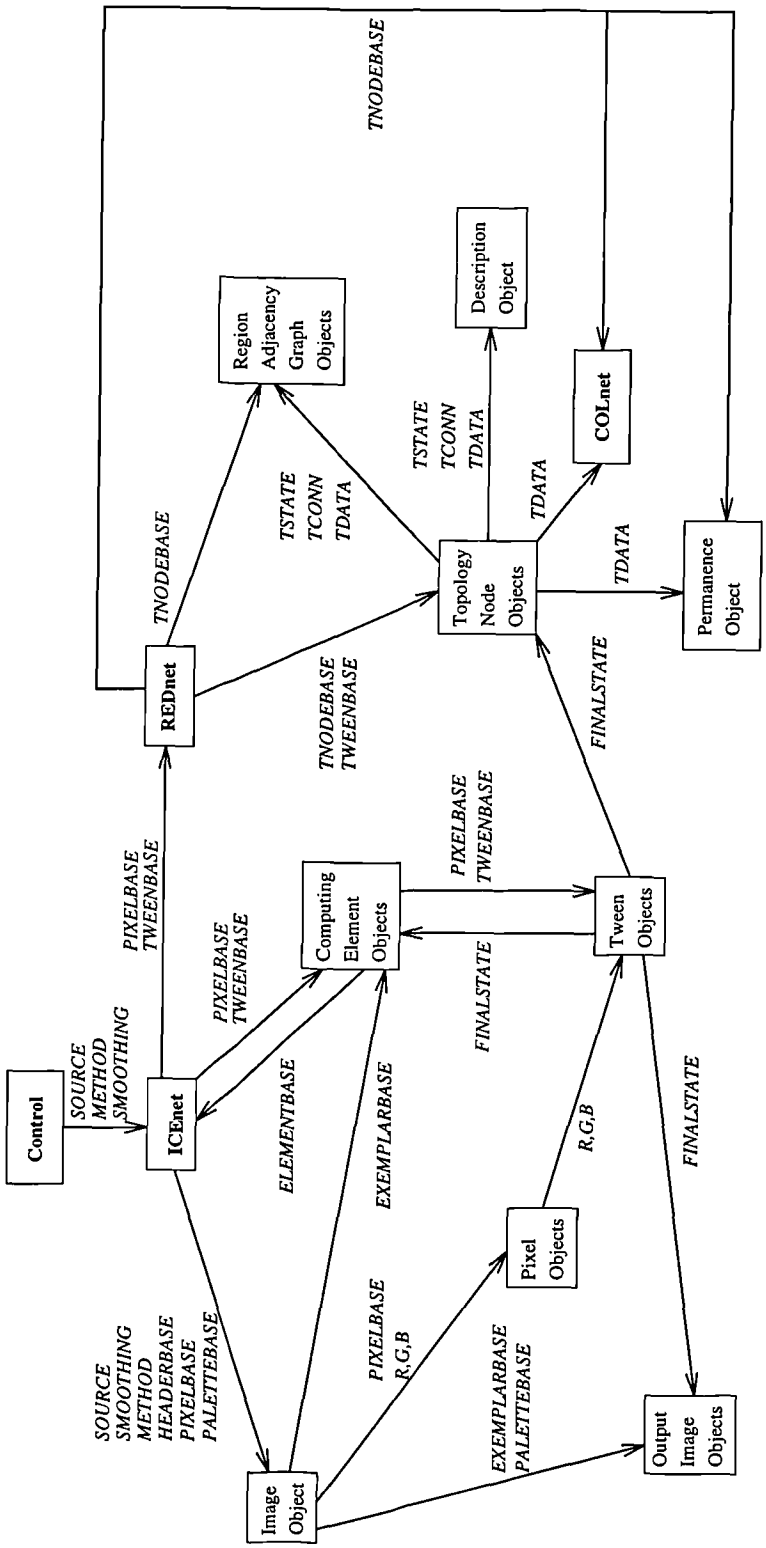


Figure 5.6: Data Passing

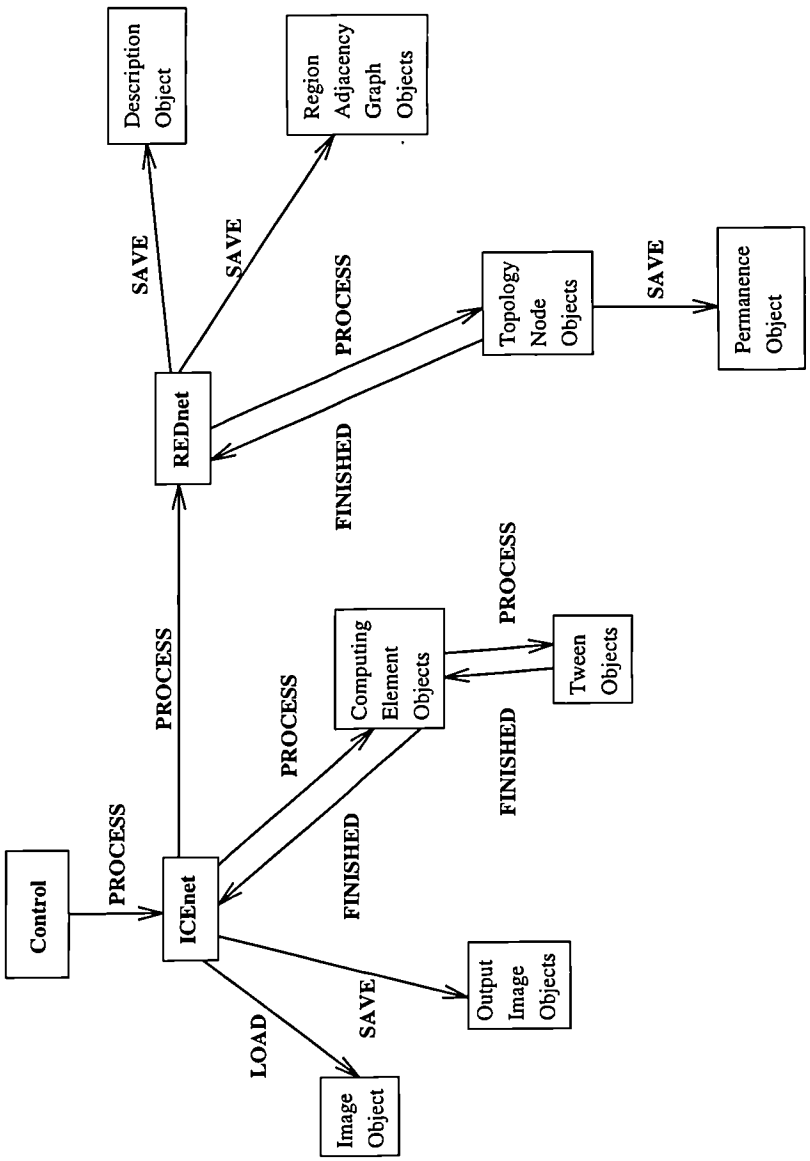


Figure 5.7: Process Message Passing

```
Send PROCESS signal to REDnet
```

On the **Instantiate image object** call, sections of memory are allocated and labeled as IMAGEBASE, HEADERBASE, PIXELBASE and PALETTEBASE to house the elements of the image in the form of a header, individual pixels and the palette lookup table (fig 5.4). The header for a PCX format image is 128 bytes and consists of the following information :

- PCX file identifier
- Version compatibility level
- Encoding method
- Number of bits per pixel, or depth
- X, Y position of left edge
- X, Y position of top edge
- X, Y position of right edge
- X, Y position of bottom edge
- X screen resolution of source image
- Y screen resolution of source image
- PCX color map, for 16 colour images
- Information byte for resolution, should be 0 or 1 if it is a standard resolution fax
- Bit planes in image
- Byte delta between scanlines
- Palette information byte 0 = undefined, 1 = color, 2 = grayscale
- Filler to make header size 128 bytes

The pixel reference array will occupy $(N+1) \times (M+1)$ bytes and the palette will consist of 256 (R,G,B) references for a 256 or 8-bit colour image occupying 768 bytes. The reference for the source image pointer, SOURCE, is then passed to the image object and the LOAD signal is sent.

On the **Instantiate computing elements** call, memory is allocated for the ICEnet computing array and labeled as ELEMENTBASE, together with HTWEENBASE and VTWEENBASE for the horizontal and vertical tween sets. The computing elements are then instantiated in the relevant positions relative to ELEMENTBASE and are passed their pixel locations relative to PIXELBASE and each is then responsible for finding the data for each of its four associated pixels either from its neighbours or from the pixel array. The default is a memory reference copy.

Once all of the major objects are created, the **PROCESS** signal is sent to the computing elements and the segmentation is performed. When the segmentation is complete, each of the computing elements sends a FINISHED signal back to the ICEnet. When all elements are finished, REDnet is instantiated.

Image Object Instantiation

The image object is designed to be a flexible translation object that takes a compressed image file and expands it into a range of different formats. This means it can easily make data available for processing or for storage.

```
Wait for SOURCE to be passed.
Wait for LOAD signal
    Open image SOURCE file
    Decompress image into pixel objects
    Smooth image if required
    Histogram incoming values into 256 pots
    Do comparison to find representative colour vectors
```

The default format for images was PCX. Once the header has been removed from the file, it is stored from HEADERBASE onward and the pixel values are decompressed. The compression format is a simple run-length encoding scheme where palette references are stored instead of the pixel values. These references are decompressed into the objects pointed to from PIXELBASE onwards.

Subsidiary functions as part of the object class are also available such as image smoothing by median filtering, as described in [141], and histogramming as described in section 5.1.5. The base memory location, EXEMPLARBASE, for exemplar vectors are then passed to the computing elements as required.

Tween Object Instantiation

Tween objects perform primary discrimination between pixel objects, they pass a value back to the computing element which is the Euclidean distance either in (R,G,B) space or in hue space. When the END signal is propagated to the tween it will set a value of FINALSTATE which depends upon its segmentation value and the value it receives from its linked computing element. This value will be binary on or off to indicate that the pixels do not or do belong to the same segment. Pseudocode for their instantiation is shown below.

```
Link to 2 x pixel objects
Interrogate pixels for R,G,B values
Wait for PROCESS signal
    return FINALSTATE
Send FINISHED signal to computing element object
```

Computing Element Object Instantiation

The computing elements are the most important part of the ICEnet and are responsible for the actual segmentation.

```
Link to tween objects
Link to neighbouring CEs
Wait for PROCESS signal
On PROCESS signal
    if HISTOGRAM is set
        Perform colouring from exemplars
        Send PROCESS signal to tweens
    elseif ORDINARY is set
        Send PROCESS signal to tweens
    endif
 Wait for FINISHED signal from all linked tweens
    On FINISHED signal
        Send FINISHED signal to ICEnet
```

The first function they perform is to link both to their tweens and to their neighbouring CEs. The elements then wait for the PROCESS signal from the ICEnet algorithm object. On this signal they send a PROCESS signal to their tween objects which then set their state, FINALSTATE. When all of these states have been formed the FINISHED signal is sent from the tweens and the states are available to be sent to output objects. The FINISHED signal is now sent to ICEnet which has completed the segmentation and sends the PROCESS signal to REDnet.

Output Image Object Instantiation

The output image objects are able to make persistent objects describing both the false colouring and segmentation states of the ICEnet.

```
Create pixel objects
Interrogate tween objects for their FINALSTATE values
Wait for SAVE signal
Make values persistent
```

5.1.4 Histogramming with Palette Reduction

As discussed in section 3.3 exemplar selection is necessary to segment the image One way to perform this is to first histogram the frequency of occurrence in the image of each of the palette entries Then a representative exemplar set must be generated by creating a histogram of sorted palette entries with their relative number of hits An example from a real image is shown in fig 5 8 A required yield of pixels in the image is then decided upon and a corresponding number of exemplars chosen

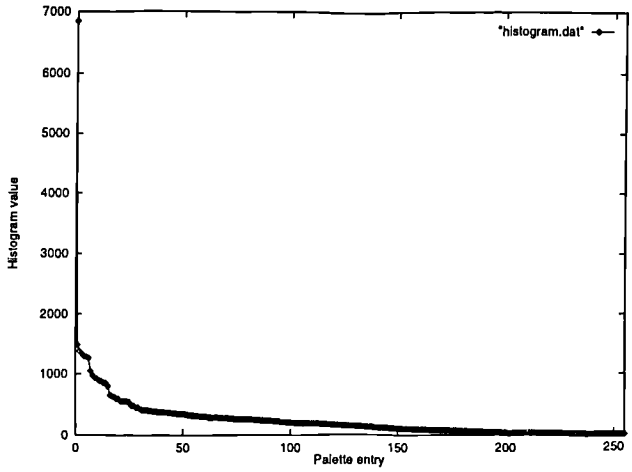


Figure 5.8: Example Sorted Palette Histogram

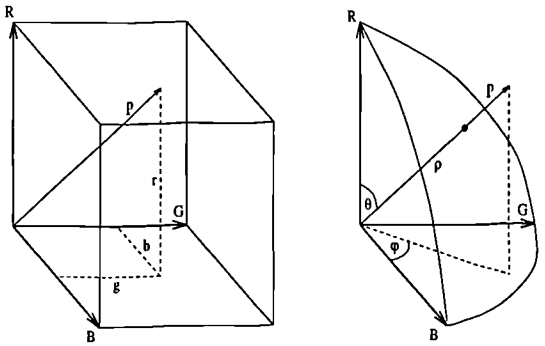


Figure 5.9: (a) (R,G,B) space, (b) (ρ, θ, ϕ) space

The hues present are the most useful measure of colour differences, as discussed in [12]. Images direct from the camera come in a compressed form as vectors in (R,G,B) space. While this is useful for faithful reproduction of the sampled scene it is less useful for simply identifying the hues since intensity (overall brightness of the colours) is bound up in the measurement. In order to normalise these vectors they may be shifted into a spherical coordinate space using the standard transformation :

$$\rho = (r^2 + g^2 + b^2)^{\frac{1}{2}} \tag{5.1}$$

$$\theta = \tan^{-1}\left(\frac{g}{r}\right) \tag{5.2}$$

$$\phi = \cos^{-1}\left(\frac{b}{\rho}\right) \tag{5.3}$$

This has the effect of normalising the vector length (brightness), as shown in figure 5.9.

In order to illustrate the effect of this normalisation, a benchmark image is illustrated in fig. 5.10. This standard image [4] contains a full range of colours and gives useful limits for subsequent plots.

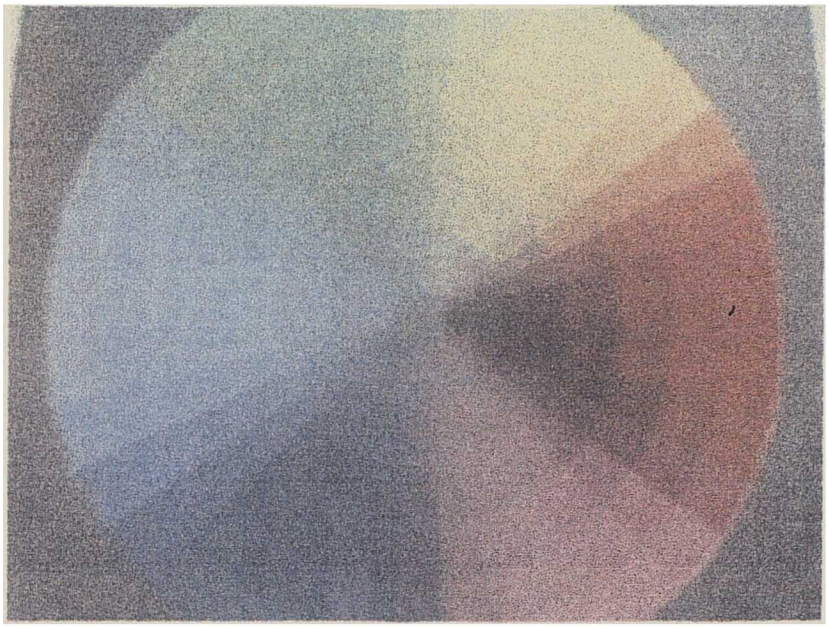


Figure 5.10: Benchmark Colour Image

It can be noted that the shape of the data in the hue spectrum histogram (fig. 5.11) is similar to the standard C.I.E chromaticity diagram [4].

For the real example image shown in Appendix III the hue spectrum histogram is shown in fig. 5.12 below. It can clearly be seen that several distinct peaks are present and that a majority of pixels are classified into only few groups.

Palette Reduction

Once the (R,G,B) palette set has been derived from the image, the representative exemplar set must be generated. This is done using palette histogramming and a histogram of sorted palette entries, with their relative number of hits is produced, for example fig. 5.13. A required yield of pixels in the image is then decided upon and a corresponding number of exemplars chosen.

In our case a palette reduction heuristic is proposed that allows a wider range of exemplars to be chosen and a greater level of representation of the original image colour variation. This is done by comparing the relative similarity of exemplars in the exemplar set and removing any that are similar enough to map to approximately the same vector at the classification stage. With a level of tolerance set sufficiently high, say 0.8 this allows a greater number of different vectors while still being a representative set. Similarity is judged using the following criterion. Given two vectors (r_i, g_i, b_i) and (r_j, g_j, b_j) their similarity coefficient, $S_{i,j}$, has three components, $S^r_{i,j}$, $S^g_{i,j}$ and $S^b_{i,j}$ given by :

$$S^r_{i,j} = \frac{\sqrt{(r_i - r_j)^2}}{max(r_i, r_j)} \tag{5.4}$$

$$S^g_{i,j} = \frac{\sqrt{(g_i - g_j)^2}}{max(g_i, g_j)} \tag{5.5}$$

$$S^b_{i,j} = \frac{\sqrt{(b_i - b_j)^2}}{max(b_i, b_j)} \tag{5.6}$$

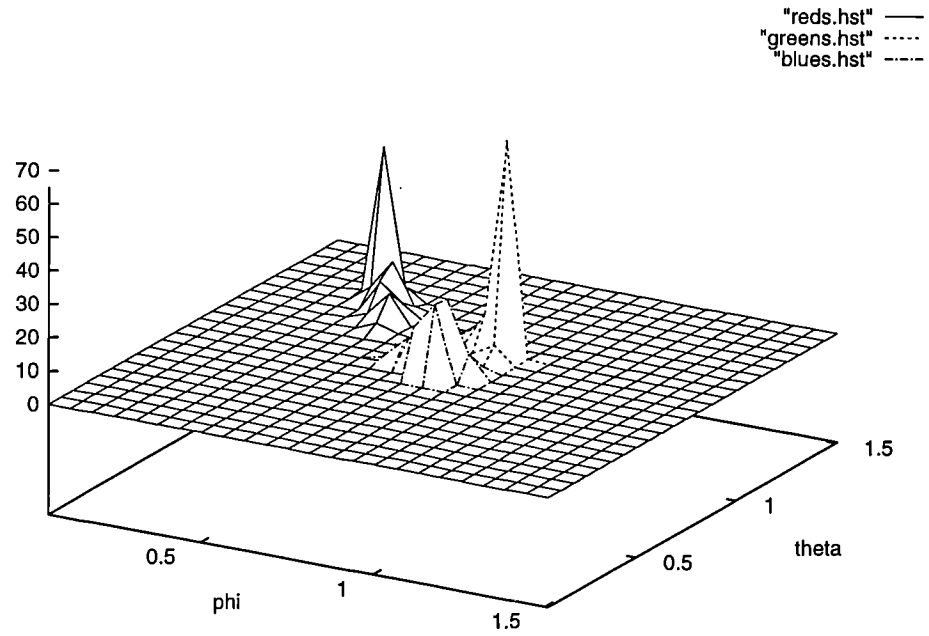


Figure 5.11: Benchmark Hue Spectrum Plot

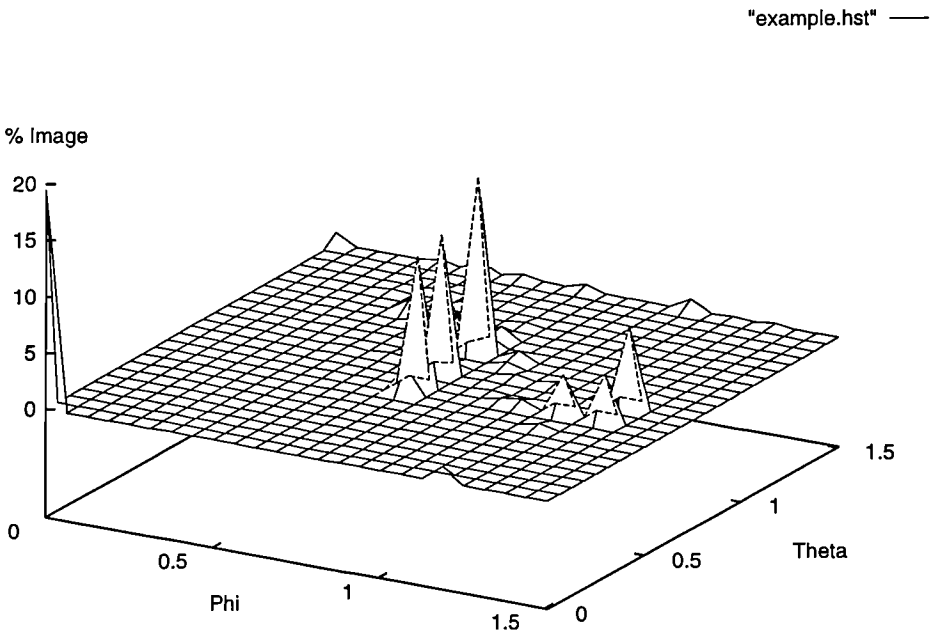


Figure 5.12: Example Hue Spectrum Histogram

And finally,

$$S_{i,j} = S^r_{i,j}.S^g_{i,j}.S^b_{i,j} \tag{5.7}$$

Two exemplars are then considered to be similar if :

$$S_{i,j} < \gamma \tag{5.8}$$

Where γ is the tolerance level for similarity.

An example of the output of the palette reduction heuristic is given below. In this example 50% coverage was chosen which would normally require 33 vectors but with palette reduction this was reduced to 15, allowing the choice of a further 18 vectors. It should be noted that normally γ would not be set as low as 0.5.

```
Picked 50% exemplar coverage.
This requires 33 vectors.

Palette reduction heuristic start ..
Gamma = 0.5

232 240 248 == 208 232 232 [s=0.811]
232 240 248 == 224 248 232 [s=0.874]
232 240 248 == 176 208 208 [s=0.551]
232 240 248 == 176 216 232 [s=0.639]
232 240 248 == 248 248 248 [s=0.905]
232 240 248 == 200 232 216 [s=0.726]
232 240 248 == 184 208 192 [s=0.532]
232 240 248 == 192 208 216 [s=0.625]
 56 128 192 ==  56 112 184 [s=0.839]
 56 128 192 ==  40 112 176 [s=0.573]
 56 128 192 ==  72 144 208 [s=0.638]
 56 128 192 ==  80 160 208 [s=0.517]
 56 128 192 ==  80 136 192 [s=0.659]
  0  16  56 ==   0  24  72 [s=0.519]
  0  32  40 ==   0  40  56 [s=0.571]
 40 104 160 ==  32  88 160 [s=0.677]
 40  64  72 ==  32  56  56 [s=0.544]
 40  64  72 ==  32  72  88 [s=0.582]

No. disposed of is  18 out of  33
```

In this example, a 70% coverage was achieved using only 25 vectors from the original palette. That is 70% of the pixels in the target image are mapped to only 25 exemplars.

5.1.5 Discussion of Testing

The algorithm was tested on 120 examples of human faces each of size 125×174 pixels, giving 21,750 possible plane areas. Results for an example image are shown in table 5.1. The false colourings are a function performed by the CEs after segmentation and show the colour which each of its four areas has been mapped to form the segmentation set, as discussed in section 3.3. Different smoothing levels and different numbers of exemplars were used and the figures to display these tests are shown in appendix

Smoothing	Exemplars	No. planes	Active tweens	Av. tweens/plane
9	2	7	994	142.0
9	4	26	2298	88.4
9	8	210	6550	31.0
9	16	414	9293	22.4
7	2	26	1478	56.8
7	4	261	5524	21.2
7	8	374	8378	22.4
7	16	797	12242	15.4
5	2	31	1828	59.0
5	4	385	7266	18.9
5	8	567	10722	18.9
5	16	1185	15525	13.1

Table 5.1: Table of Results for Example Segmentation

III. All of the images were segmented using hues rather than using R,G,B coordinates. This has been shown [12] to reduce the number of exemplars required and reduce misclassification of small planes. To a lesser extent gradient changes due to shading are also reduced in this way.

Whilst it is not vital in this work that the number of planes detected is kept as low as possible, a large value of active tweens per plane is obviously desirable. This is because small areas of colour are generally considered as noise in the imaging process and are therefore not as important as larger areas. In the pursuit of the general model small planes can be held for further processing at a later stage while larger planes are taken to be *object planes*, as will be discussed later.

The top 16 most important exemplar vectors for the two images have been plotted in fig 5.12. With this number of exemplars, this image has 78% of pixels mapped exactly to palette values. It can be seen that the basis for a general model for this type of image is already apparent since the range of the vectors is relatively low. This suggests the palette exemplars that might be used for the general model.

Fig.5.13 is a graph of the percentage of correctly covered image pixels with a given number of exemplars averaged over all 120 images.

The results from this processing with discussion are in the first section of Appendix III.

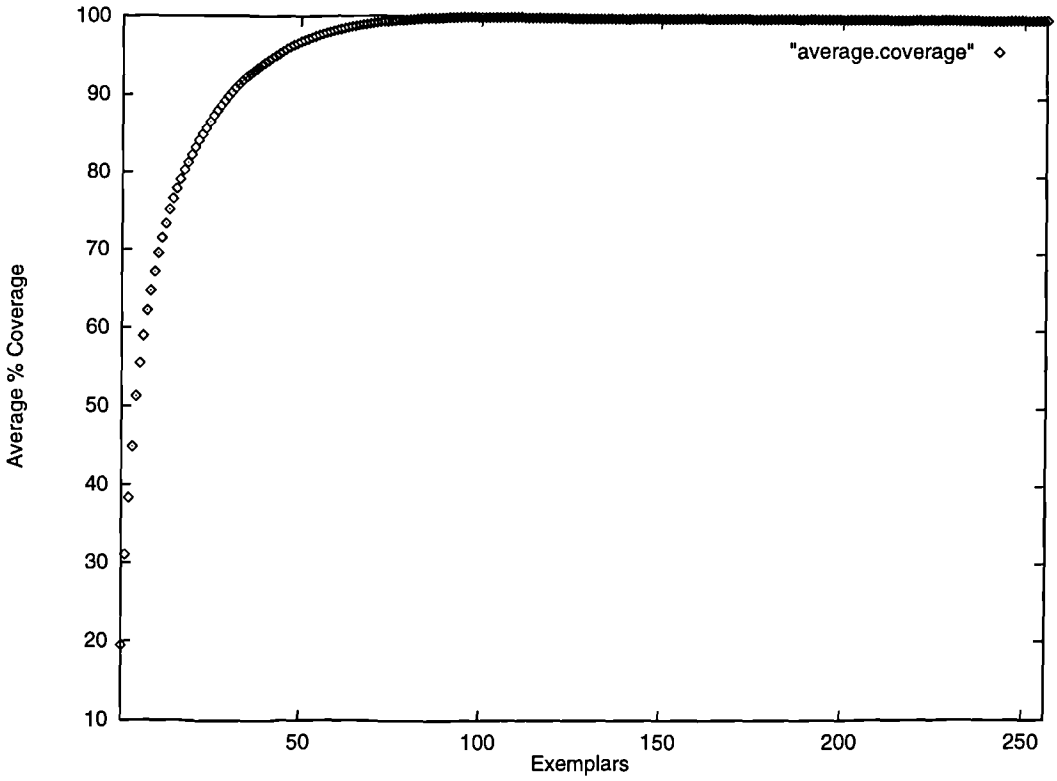


Figure 5.13: Average Percentage Coverage against Number of Exemplars

5.2 Post-processing by REDnet

5.2.1 Physiological Motivation

Once an image has been successfully segmented, its topology may be extracted by a second network using only the state information from the computing nodes. The kind of propagation versus inhibition used at this stage is a very common feature of true neural processing [76]. Topology can be built up in a very few passes in this way.

5.2.2 Description of REDnet

The image reduction network (REDnet) works by processing the computing element output from the ICEnet immediately after it has been evaluated. First the tween values are loaded, then a propagation signal is sent so processing in this level can begin. Initially, one REDnet cell exists for every pixel in the original image (fig. 5.15(a)) and it is these cells that represent the topology after the collapsing has taken place. The process is shown by the creation of objects in the next section.

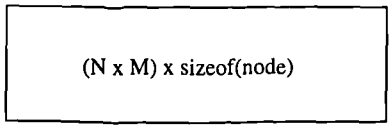


Figure 5.14: REDnet Main Usage

5.2.3 Algorithm, Coding Strategy and Implementation

REDnet Object Instantiation

Initially, the REDnet object receives a PROCESS signal from ICEnet and builds a network of topology node objects which then produce the final topology graph. REDnet first instantiates a section of memory to hold the topology nodes, TNODEBASE which is $N \times M$ nodes in size.

```
Instantiate one topology node per image pixel

Instantiate structure objects for output

Send PROCESS signal to topology nodes

Wait for FINISHED signal from topology nodes

When no more FINISHED signals left

    Send SAVE signal to region adjacency graph objects
```

Topology Node Object Instantiation

The topology nodes have two separate functions. The first is to collapse into neighbouring nodes until no more are present and the second is to compute data about the nature of the plane which they represent. The first of these functions is performed by sending a propagating (PROP) signal to neighbouring nodes. If there is a response then they collapse into that node by adding their link data to that node and deleting themselves. Once no further collapsing is possible they compute their data, TDATA, set their state, TSTATE, to ON and send a FINISHED signal to REDnet. Area is also computed at run-time.

```
Link to neighbouring tweens and topology nodes

Wait for PROCESS signal

On PROCESS signal

    Destroy all links where tween has FINALSTATE = ON or equivalent

    Send PROP signal to neighbouring nodes

    If a node responds (receive order shown in fig. 5.15(d))
```

```
        Collapse into that node

        Remove your connection to it from the link table

        Add your link table to its link table

        Set TSTATE to OFF

    Else

        Node is single representative node

        Compute TDATA values

        Set TSTATE to OFF

        Send FINISHED to REDnet
```

The TDATA construct contains the following information :

- Bounding line set
- Plane connection set
- Plane number
- Perimeter length
- Palette index colour
- Area
- Percentage of image area
- TSTATE, ON or OFF
- Top left X position
- Top left Y position
- Object node status, on or OFF
- Red, green and blue components of it colour
- Colour intensity
- Colour saturation
- Colour values as (ϕ, θ, ρ)
- Relative depth, estimated by saturation

The structure object is then quite straightforward :

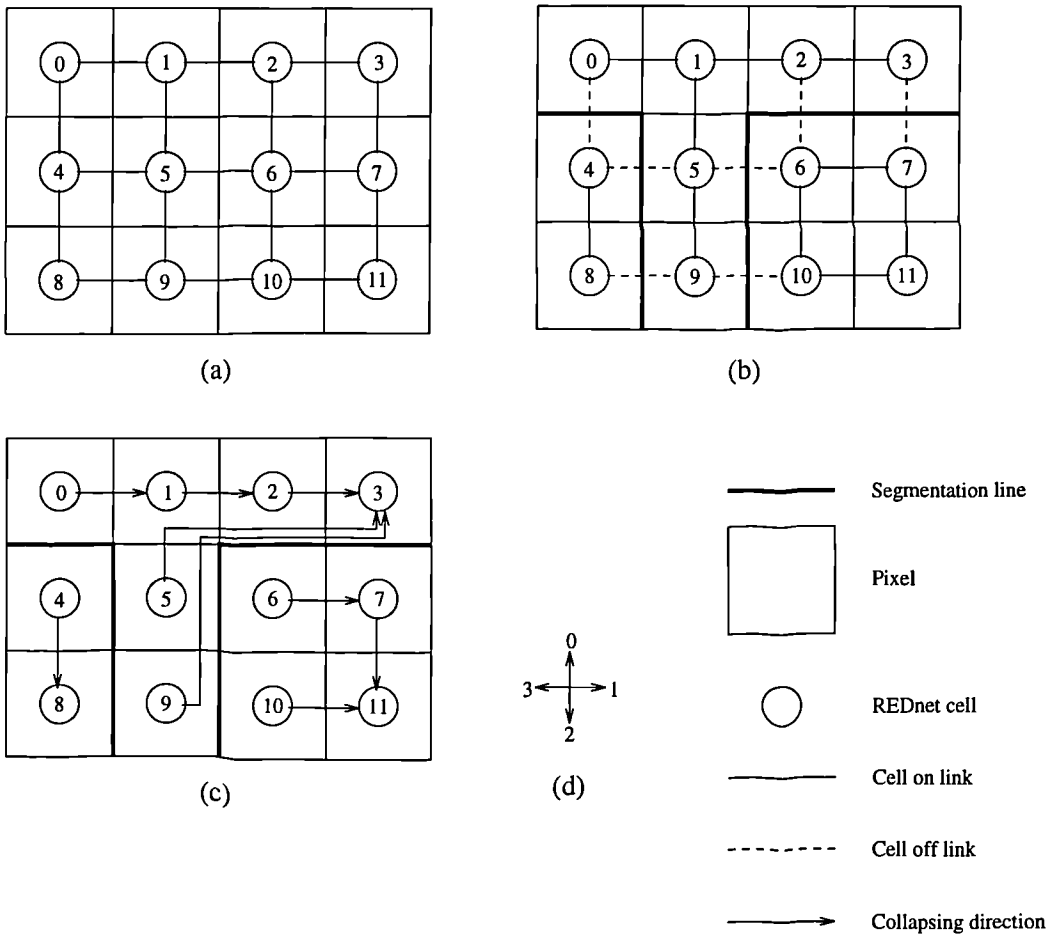


Figure 5.15: REDnet Construction and Algorithm

Region Adjacency Graph (RAG) Object Instantiation

There are several different types of RAG object which save different pieces of the TDATA construct and allow graphs to be constructed with them. Any information in the TDATA structure may be saved in this way for further analysis. They are saved in the XVCG package standard [142].

```
Instantiate region adjacency graph object

Interrogate topology nodes for their TSTATE and TDATA

Convert TSTATE and TDATA into node and edge information

Compile into XVCG format data file

Wait for SAVE signal

Make data persistent
```

Description Object Instantiation

The description object fulfills the function of scene description as described in chapter 3. The persistent version of this object is in LaTex [143] format and consists of topological descriptors and relationships. Pseudocode for the instantiation is as follows :

```
Interrogate topology nodes for their TSTATE and TDATA

Convert TSTATE and TDATA into node and connection information

Compile into Latex format data file

Wait for SAVE signal

Make data persistent
```

5.2.4 Inputs and Outputs

Results are automatically generated at all stages of the processing through the network objects. The files produced are in a variety of formats for different display packages. Files and formats are as follows:

1. Input files. These files conform to the run-length encoded PCX file format. This format is widely used in graphics files and contains a 128 byte header and the pixel data is then run length encoded. A palette of 256x3 entries follow the main data.
2. Segmentation output. There are three types of segmentation files output at the first layer:
   (a) *Tween file.* This file is in LaTex format for processing by the LaTex publishing system and not intended for further examination.
   (b) *Segmentation image file.* This is in PCX format.
   (c) *False colouring image file.* This is in PCX format.
3. Topological post-processing output. These files are as follows:
   (a) *Description file.* This file is in LaTex [143] and is intended for documents.
   (b) *Topological description file.* This conforms to the VCG visual compiler graphs format [142]. This output is loaded to the package and has its layout performed automatically.

5.2.5 Discussion of Testing

Results were generated by directly pipelining ICEnet to the REDnet. The results from the faces tests were used and example topologies are shown in Appendix III. As will be discussed in chapter 6, many more levels of information were also extracted at this time.

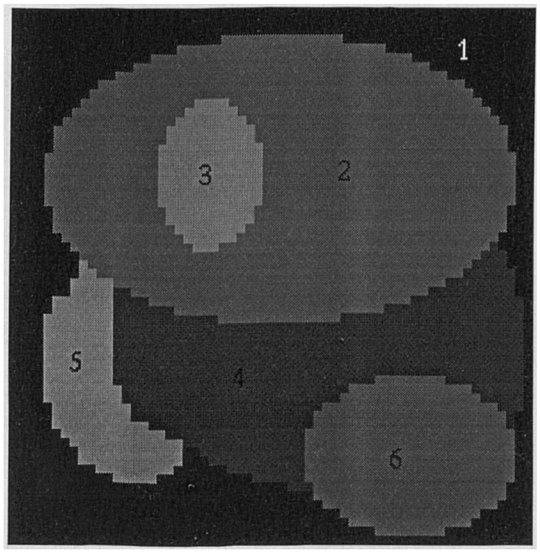


Figure 5.16: Example Segmentation

5.3 Textual Description Example

An example segmentation is shown in fig.5.16 and 5.17.

The line descriptor W represents the outer edges of the viewport, or image edges. A listing of the automatically generated line segments and plane descriptors is shown below:

$$D^1 \equiv (\{l^0_{1,4}, l^1_{1,4}, l^2_{1,5}, l^3_{1,2}, l^8_{1,6}, W\}, \underline{p}_1)$$
$$D^2 \equiv (\{l^3_{1,2}, l^4_{2,3}, l^5_{2,4}, l^6_{2,5}\}\underline{p}_0)$$
$$D^3 \equiv (\{l^4_{2,3}\}, \underline{p}_3)$$
$$D^4 \equiv (\{l^0_{1,4}, l^1_{1,4}, l^5_{2,4}, l^7_{4,5}, l^9_{4,6}\}, \underline{p}_2)$$
$$D^5 \equiv (\{l^2_{1,5}, l^6_{2,5}, l^7_{4,5}\}\underline{p}_3)$$
$$D^6 \equiv (\{l^8_{1,6}, l^9_{4,6}\}, \underline{p}_0)$$

The tree structure for connectivity is shown in fig.5.18. This diagram shows both the encapsulation and bordering relationship in this case. Direct inference is :

$$D^2 \odot D^3$$

5.4 COLnet - Colour Classification Layer

COLnet is a trainable second stage network which maps colours to possible classifications of type. This network may vary in complexity depending upon the application space and the complexity of the objects presented. A decision surface is generated for the colour space in a contextual way in order to achieve classification of planes generated by REDnet. Physiological motivation for this is the colour analysis capabilities of the V1 area of the visual cortex, which is the colour post-processing area which can perform classification based only on colour.

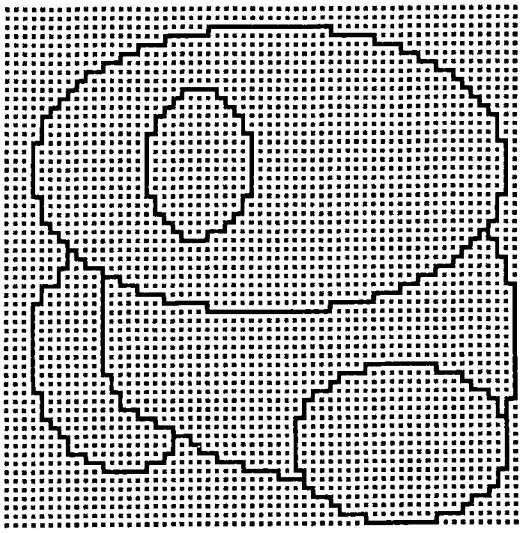


Figure 5.17: Segmentation Diagram Showing Computing Elements and Active Tweens

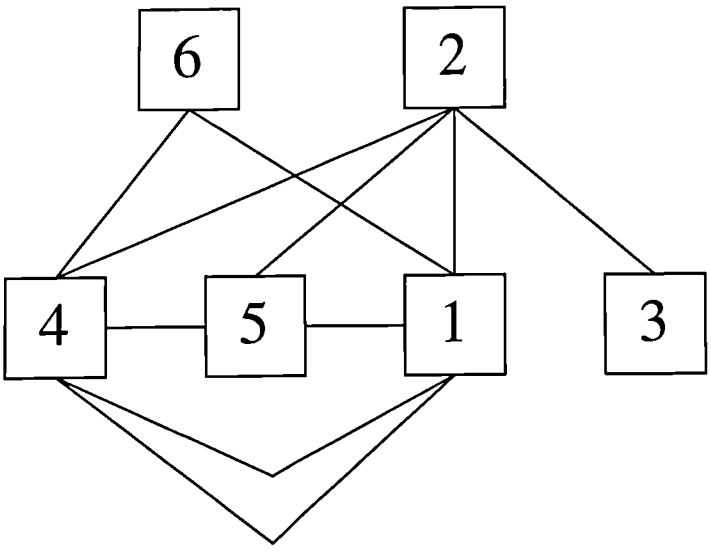


Figure 5.18: Topological Structure Output

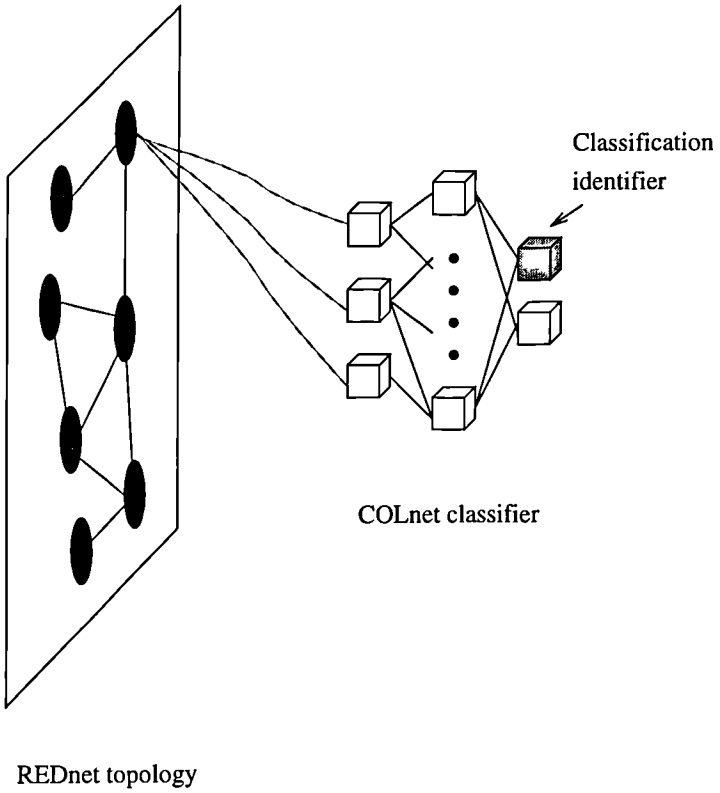


Figure 5.19: COLnet Structure

5.4.1 Algorithm, Coding Strategy and Implementation

In order to achieve a degree of generalisation over many different input exemplars a neural network is employed which may be trained using a variety of feedback training algorithms as detailed in [78] and [79]. The training regimes used in this work were quick-propagation and back-propagation with momentum, which produce well balanced, robust networks with a good degree of generalisation. Detailed discussion of the application of training is shown in chapter 7 and in [80].

5.4.2 Neural Network Analysis

Neural networks were used for this layer rather than conventional statistical methods because:

- The decision boundary for the data is not simple. Neural algorithms bring with them a model-wide non-linearity that is particularly useful for modeling complex boundaries.
- The kind of neural networks used provide input to output mapping inherently in their design. This is particularly useful in this context.
- The outputs give *confidence* output rather than a set classification. This is particularly useful if the network is to be embedded in a real-time system.
- A hardware implementation of this system would be inherently fault-tolerant. Extensive damage has to take place before the distributed learning is affected.

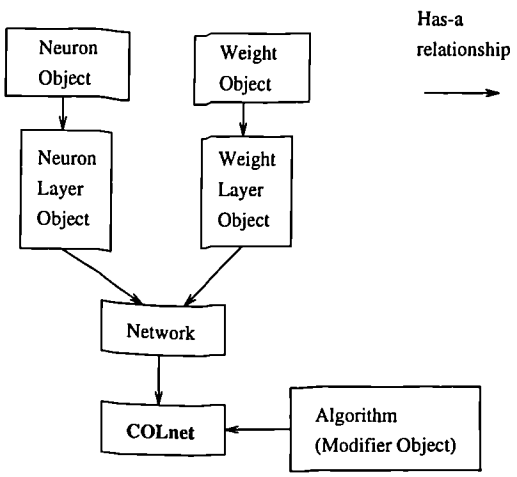


Figure 5.20: COLnet Object Hierarchy

- VLSI implementability. Real-time systems can be built which exactly mimic the parallel nature of the networks.

5.4.3 Discussion of Testing

Specific testing of this structure was not performed for the face data but was performed comprehensively for the natural scene data presented in chapter 7. Discussion is therefore delayed until then.

5.5 Summary

The networks described in this chapter are based on the transport of information from layers $4C\beta$ to the blobs/interblobs region and then to the V1 area of the visual cortex. They employ a very simple discrimination algorithm in parallel and so in hardware would be operable in a single time step for the entire image provided a sufficiently large network of tweens were used. Since there would be a certain amount of noise inherent in this scheme a second, post-processing, layer has been designed. This layer consists of smart data holding elements which can merge together and perform computations using information derived locally from their neighbours, such as possible depth in 3D and shading from colour. Three pipelined network objects were introduced.

The first network object is used for finding simple edges in an input colour scene. This is achieved by activating small elements between the pixels laid out in a supergrid. A pre-processing stage consisting of both palette reduction and median filter smoothing is also performed.

The second stage network is a parallel network capable of condensing the segmentation stage outputs in to a 2D planar topology. This network is described and an example is given of the software simulation.

The third stage network is a standard back-propagation neural network. This network is capable of colour discrimination and classification by generating a decision boundary based on training inform-

ation provided by the user from previously segmented input images. These classifications form the basis for the colour context described in chapter 6. Methodology for the training and application of this form of network is delayed until chapter 7.

The next chapter details the generation of a heuristically driven second phase that, using context based models, can both describe and analyse scenes while evaluating hypotheses about content. This second phase consists of several separate entities :

- Speckle (noise) reduction based on size. This is also useful for plane collapsing to form the most important primal-sketch like elements.
- Topological mapping based on intensity, for attention and analysis of psychological phenomena such as the Aubert-Foerster phenomenon [22].
- Topological mapping by hue for object description.
- Topological mapping by saturation, of general use.
- Contextual spectral hue matching by neural networks and Bartlett-type scripts as detailed in [114].
- Attention graphing based on intensity topology.
- Natural language description based on the described syntax.

In the next chapter, the idea of visual contexts is described. Visual contexts are the rules discussed in section 3.7 and define areas of scenes where inherent rules apply that may fundamentally change both the structure and nature of that part of the image relative to the rest.

Chapter 6

Context

6.1 Introduction

Recognition of objects and scenes requires that a model of some kind must be stored in the brain against which a match may made. Many proposals have been put forward for a storage and organisational scheme which may achieve this but few have been able demonstrate a system that exhibits all of the mechanisms that are required. The main features of such a scheme are:

- *Data holders.* These must be of reasonable complexity and be expandable to accommodate new data as it is received.
- *Creation.* The scheme should be capable of creating new data holders when the need arises.
- *Rule formation.* It should be capable of forming rules in order to achieve compaction, since storage is a finite and limited resource.
- *Recall.* A method for retrieval of data holders into a thought process is obviously required.

The philosopher Kant was the first person to propose the idea of what he called *schema* in his *Critique of Pure Reason* of 1787 [115]. Schema are a scheme for acquired symbols that group together concepts, ideas and classifications (or schemata) that are related in some way. Elaborate examples of schema have been explored by Schank and Abelson [116], for example, who have applied the scheme to text comprehension. *Scripts*, on the other hand, are an extension of schema that have predictive capabilities and aid in the formation of understanding since they help anticipation of expected schemata. They encode general or *generic* information that can be applied to specific situations if those situations are instances of schema. Kant's schema, together with Minsky's *frames* (discussed in section 6.4) provide the basis for the context model that is proposed in this chapter.

Recognition of objects also requires a method for performing the matching and this is where the complexity in any vision system is to be found. Correct model identification is hampered by one or

both of the following problems:

1. The stored model has been transformed in some way relative to the target object.

2. The target model has been transformed relative to the stored model.

Simple toy-world systems with limited model spaces and hand-built canonical models largely avoid such problems. However, in the real world they are unavoidable and any realistic model must take them into account. In advocating the use of *contexts* as a *look-ahead* strategy some of the problems of de-transforming the target objects may be addressed. Customising algorithms to fit the appropriate contexts, where even local customisation is possible, will also make identification possible even in extreme circumstances.

The brain area most suitable for context assessment is the iconic store where rapid sketches of visio-spatial perceptions are stored for a very short time and are quickly overwritten [72, 73, 133, 136, 149, 155]. The function of this iconic store has been moot for some time and it is partially the way in which it is refreshed that has caused debate. It seems to preserve visual information from a brief glimpse of a scene for a period of 500ms or more. During this time it appears to decay passively. Information in the iconic memory appears to be in a vaguely interpreted form since only physical cues can be used to give a partial report advantage [155]. It has been thought that this is the area where successive views are integrated as observers fixate different portions of a scene [150]. Some have argued strongly that it cannot since it is tied anatomically to retinal coordinates [157]. It can thus serve no useful integrative function since we would be replacing the problem of comparing retinal snapshots with comparing iconic snapshots. Thus it would seem that a memory area at a more abstract level would be needed to serve this integrative function.

So, we have an iconic store which exhibits the following properties:

- It derives a crude snapshot of a scene.
- It is tied to the retinal co-ordinate frame and thus has little use as a real-world co-ordinate representation.
- It is heavily linked to the post-iconic visual store [156].
- It is masked by the presentation of bright light immediately after the presentation of a test pattern [156].
- It is not affected by level of pattern complexity.
- It bears some similarity to the notion of Minsky's frames [157] in that it incorporates a projection mechanism.

From these properties it seems that the iconic store is used as both a stable model of the scene since it averages many inputs over a 500ms period. It is heavily linked to the post-iconic visual store, suggesting it is in some way connected with the recognition procedure but is not affected by

pattern complexity. For these reasons I propose that iconic memory is used as a context propagating mechanism which is used for the stable assessment of scene conditions to other parts of the vision system.

The shape of the retina is fairly complex. Being tied to the retinal co-ordinate frame would not be a hindrance to the gross assessment of context however, since properties such as connectedness and localized object/patch placement are preserved. Further, colour and brightness assessment are not affected by the co-ordinate system. In fact, the only properties affected are specifically geometric ones.

6.2 Requirements of the Visual Model

A model of any part of the brain function must exhibit the three fundamental requirements :

1. *Reproduction* of the given behaviour. It is not sufficient to suggest a model without showing clearly that some aspects of the modeled function are demonstrated.

2. *Principles* similar to the original which make the mechanism possible. The behaviour of the model must at least be analogous to (preferably directly analogous to) the original function.

3. *A mechanism* for the models implementation. That is a demonstration must be performed on a non-trivial example that shows the function at work.

Even the simplest of models of brain function miss at least one of these three clear and reasonable requirements. Widely propounded ideas are often unprovable, miss caveats and features of the real thing and are just complex enough that they cannot be disproved (for example the four accounts of mental structure given in [144]). All are basically Cartesian, many are so-called Neocartesian, which means that there are few who dare argue with them and thus they are loaned a degree of respectability. There are, however, problems arising from this approach which leave logical holes which need to be filled.

6.3 Cartesian Ideas of Brain Function and Learning

Descartes holds a central place in most models of brain function because he was among the first modern philosophers to propose any formal theory. One example of his ideas is the famous analogy between automata in the Royal gardens in Saint-Germain and human reflex. That is, given a set of circumstances (those of one standing on a trigger paving stone) then there will be a given response (automata movement). He attempts to describe human responses to virtually all stimuli in this way. His description of reflex, for example, (as shown in figure 6.1) is as follows:

> "As for example if the fire A is near the foot B the particles of this fire which as you know move with great rapidity, have the power to move the area of skin which they touch; and in this way drawing the little thread that you see attached there, at the same time they

Figure 6.1: Descartes' Illustration of Reflex

> open the entrance of the pore d,e, at which this thread terminates, ... Now the entrance to the pore or little conduit d,e, being thus opened the animal spirits in the cavity F enter within and are carried by it partly to the muscles that serve to withdraw this foot from the fire, partly into these that serve to turn the eyes and the head to look at it, and partly into those that serve to advance the hands and to bend the whole body to protect it."

The ventricles in the brain were therefore thought to pump so-called animal spirits which had been previously manufactured in the heart. The little strings opened pores in the brain and this fluid was let into tubes to inflate muscles and allow them to move. The relationship between stimuli and reaction is described as a mechanically (hydraulically) *causal* relationship.

After much more of this pontification on animal spirits flowing in the muscles, he finally arrives at mechanical model of learning and memory:

> "When the mind wills to recall something, this volition causes the little (pineal) gland, by inclining successively to different sides, to impel the animal spirits toward different parts of the brain, until they come upon that part where the traces are left of the thing which it wishes to remember; for these traces are nothing else than the circumstance that the pores of the brain, through which the spirits have already taken their course on presentation of the object, have thereby acquired a greater facility than the rest to be opened again the same way by the spirits which come to them; so that these spirits coming upon the pores enter therein more readily than others."

Cartesian tenets have had many advocates and in their time were politically and socially very acceptable. In fact it has been suggested that these ideas are even modern [147] and that the strings and tubes are analogous to neural pathways. However, through manipulation of these ideas Descartes was also able to describe the actions of animals to be no more than the reflex actions of automata, the screams of tortured pigs to be "no more than the ticking of a clock". This is where the idea of mind-body dualism was introduced with the proviso that humans have a mind, that is higher intelligence,

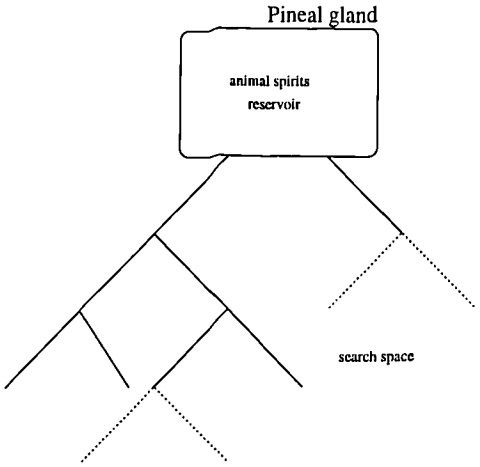


Figure 6.2: Descartes' (Parallel) Brain Search Procedure

and animals do not.

The idea of perceptions being in some way laid down or deposited in the brain does have some merit and if the quote is interpreted correctly, he is suggesting a parallel biurified search mechanism (fig 6.2).

The idea of animal spirits as a storage and retrieval model has the following structure :

- *A method for the transportation of perceptions.* He suggests that perceptions are transported to a central location.
- *A method of storage.* Once perceptions are at the central location they are stored for later retrieval.
- *A method of retrieval.* However different the hydraulic mechanism is from modern ideas of neurophysiology, it still embodies a method for retrieval.

There are some elements which are lacking :

- A cohesive structure for perception storage. As it stands an enormous space would be required to store *all* the perceptions of objects one would meet, even in a relatively short 17th century lifetime.
- A mechanism for the manipulation of perceptions. Once one has identified an object or situation there is no explanation for the the manipulation process.

6.4 Minsky's Ideas of Brain Function and Learning

Some modern thinkers on artificial intelligence prefer the tested ground of Minsky[148] when addressing learning and identification. Although the fundamentals of Minsky's framework lie in the ideas of Kant he has significantly expanded the idea of schema/schemata to the point where it has become strained and dogmatic.

6.4.1 Frames

The idea of a *framework for representation of knowledge* was put forward by Minsky [148] and has found many advocates since. A frame is a data structure for representing a stereotyped situation. Attached to each frame are several different types of information, some about how to use the frame and some about likely expectations.

Frames are represented by networks of nodes and relations. The top levels of these frames are fixed and represent things that are always true in the given situations. Lower levels have many terminals that must be filled by specific instances of data and which bear the conditions that must be met for this filling. Conditions may be specified for objects that must be in specific locations with more complex conditions specifying relations between them.

Collections of related frames may be built into frame systems and actions are mirrored by corresponding transformations between the frames of a system. These transformations may be used to perform calculations about the system. For scene analysis, different frames may describe the scene from different viewpoints and the transformation between one frame and another may represent the moving from one place to another. When the terminals of a frame are not filled then a default assignment may be made which can be easily displaced if a new item seems to be a better fit in the situation.

Frame systems are said to be linked to an information retrieval system which proposes frames to be filled. Once a frame has been proposed, a matching process is undertaken (serially) to try and match scene data with expected data.

6.4.2 Critique

Minsky argues that at higher levels of brain function the idea of parallel processing is not useful. He suggests that the then modern ideas of symbolic processing using hypothesis testing necessarily take place in a serial fashion and that it is hard to solve a complicated problem without giving it full attention. His most interesting assertion, however, is that the brute force parallel approach can be beaten by the serial approach using symbolic structures as data holders. It was suggested in the introduction to chapter 3 that this may not be the case and further discussion of this is given in the next section.

6.5 Visual Contexts as an Approach

Iconic vision, as discussed in chapter 2, is used almost exclusively for the primary processing of visual stimuli. Importantly, this processing takes place not in the associative cortex but in the early levels of the visual cortex. This gives further weight to the thesis that symbolic processing is used for identification at the associative level, rather than low level filtering processes. Simple descriptional symbols alone, as described in earlier chapters, are not enough to perform identification however. Scenes in the real world suffer a series of transformations from their canonical representation which change their

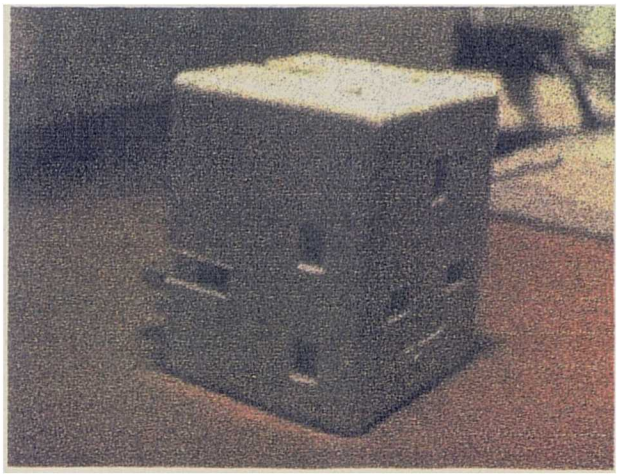


Figure 6.3: Example Acanonical Image (1)

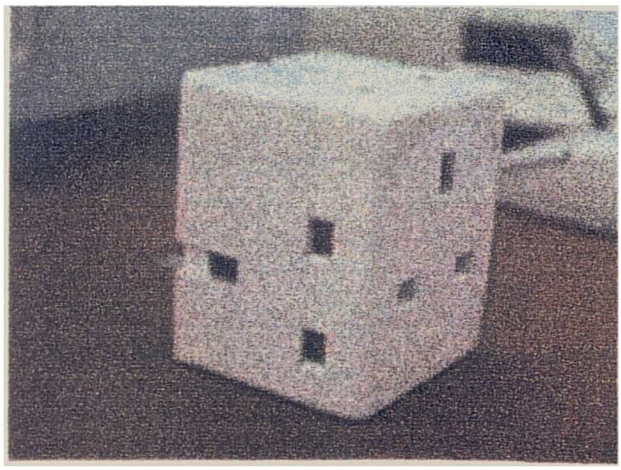


Figure 6.4: Example Acanonical Image (2)

nature to such an extent that any regular (that is to say conventional) method for description or for categorisation can never work in all circumstances. I believe this is the reason why most algorithms have a severely restricted problem domain and many complex, often sensitive, parameters to set.

What I propose is that the human cortical iconic method uses a *look-ahead* or *peeking* mechanism to sample a very brief sketch of many aspects of a scene before and after processing takes place in parallel. This peek can be used to tailor parameters of low-level processing and may also be responsible for some physical reflex changes in the sampling process as well, for example pupil dilation or adaptation of the focusing mechanism. As an example, consider the two pictures shown in figures 6.3 and 6.4. Clearly they show the same object in the same scene with only different lighting.

The visual context is the setting in which the analysis of a scene can take place. If there are obvious local clues to the nature of the context then it is likely that this context is applicable to the local surrounding structures. The visual context is likely to be smoothly continuous in most circumstances, for example when locomoting around a scene it is unlikely that dramatic lighting changes will take place or that the position of viewing will change rapidly or repeatedly (unless one is falling over). The

remainder of this section will detail the commonly held properties or local contexts in which analyses take place.

6.5.1 Some Visual Contexts

Visual contexts fall into three main groups: gross scene colour properties; geometric area properties (scene topology); apparent local area properties. These three groups I shall term colour context, geometric context and local (or object) context and their properties are outlined below:

Colour context

This is the main controlling influence of the scene and would generally be the immediate classification made by the viewer at the iconic level, whether the scene was a natural (pastoral) scene, street scene, internal scene etc. The elements that make this context up are as follows:

- *General palette arrangement* and content for large generalised areas.
- *Hue* spread and content.
- *Saturation* spread and content.
- *Brightness* spread and content.

This context can also give broad information about the colour or nature of lighting in a scene which can be a very destructive condition on many analysis tools.

Geometric context

The geometric context is controlled by the relationships of large and small areas of the image both to each other and in a more general sense to the image. Elements in this context are as follows:

- *Geometry* of areas in the scene, their relative size and shape. This would include:
  - *Fragmentation* of areas, or ratio of small elements to large.
  - *Relative size and shape* of areas
- *Topology* of areas in the scene, their connectivity or lack thereof.

For humans, it is not entirely necessary to know the exact specifications of shapes in order to perform fairly complex mental operations on them. Relative sizes and approximate shapes are generally all that is required.

Object context

Object context is dependent on relatively small collections of topologically close areas. It is from object context that such properties as shape from shading are derivable as well as more subtle cues.

Object context is made up of the following:

- *Luminosity.*
- *Lightness.*
- *Iridescence*
- *Lustre.*
- *Transparency.*
- *Chromatic illumination.*
- *Texture.*

Images to illustrate these contexts (after Birren) are to be found in appendix IV. Of these context elements texture is often the most analysed because it is the easiest to quantify using geometric or statistical methods.

6.5.2 Contexts and Frames

In some sense contexts and frames easily interact and a context is a controlling mechanism for how the filling of visual frames (or schema) are performed. This is because a context uses a complex, symbolic and relational data structure in much the same way as a frame does. Although originally Minsky was somewhat vague about the data derivation or selection mechanisms used for filling frame terminals, contexts allow real-time modifications of incoming data to be made in order to fill these terminals, even in extreme circumstances.

Contexts are not used simply as a preprocessing mechanism for frame filling although they can obviously be used as a broad brush look-ahead mechanism. Their data structure may be flexible in a way that frames may not, so they can learn what comprises a context in a probabilistic way. A fuzzy learning procedure for context agglomeration is discussed later in the chapter. Despite Minsky's assertion of doubt and those of the connectionists there is no real reason why symbolic operations cannot be carried out in parallel. A fine example of this is the simultaneous construction of context hypotheses that may be carried out. The only proviso is that the solution is not assumed to have been reached when any single hypothesis has been examined unless it shows overwhelmingly to be the probable solution. This idea fits very well with Edelman's ideas of 'neural-Darwinism' and naturally with Calvin's model of thought selection, both discussed in [61].

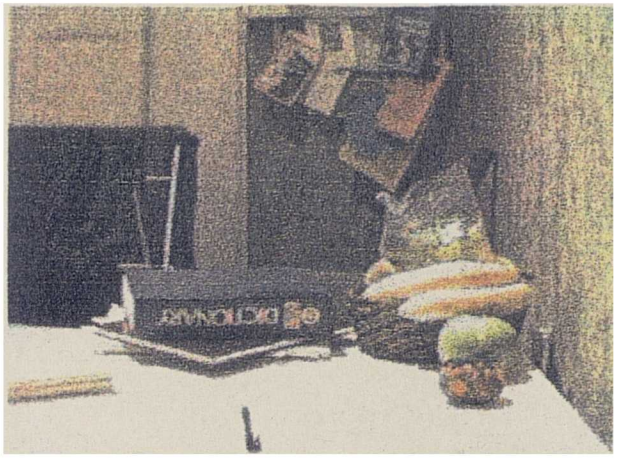


Figure 6.5: Example Scene

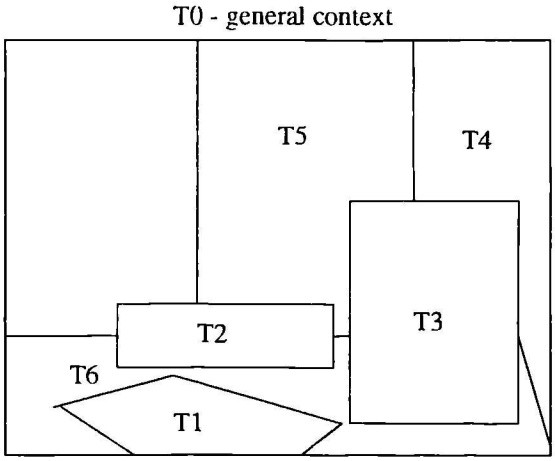


Figure 6.6: Example Scene Contexts

6.6 How Context Can Help to Solve Problems

There are two main ways in which contextual assessment can help to solve vision problems. One is in model invocation and the other is algorithm and parameter selection.

6.6.1 Model Invocation from Context

A very difficult problem in a general model based system is selecting the correct model and there have been some fairly complex symbolic methods of solving it, for example [153]. When considering a truly general vision emulation one must take into account that it must be able to deal with many sets of disparate data simultaneously. This is only possible if we consider a scene as consisting of different areas, each possessing its own contextual setting. If we take as our example the scene in fig. 6.5 together with the context diagram in fig.6.6 it can be seen that there is a breakdown into qualitatively different areas.

The overall scene is broken into essentially seven different contexts. Context $T0$ is the overall

(colour) context with its attendant lighting model as described in section 6.5.1. Some other contexts are :

T1: *Foreground and paper.* Local context, texture (i.e with writing.)

T2: *Book.* Local context, coloured , with writing.

T3: *Fruit.* Local lustre context, colours.

T4: *Side wall.* Local context with texture, shading and reflected light.

T5: *Leaflets.* Local context with texture and shading.

T6: *Foreground.* Reflective, affecting foreground objects.

It is worth noting that the treatment of writing as texture has been shown to produce some interesting results [164].

6.6.2 Algorithm Selection

Colour Constancy

Colour constancy is an issue that is very nearly taboo in the emulation of human vision. From a human vision perspective, the perceiver is not concerned with analysing the wavelength composition of light reflected from an objects surface but with exposing an object from its background. Totally different wavelength combinations can produce identical colours. The colours we see in objects are those that best set them off from their backgrounds under the prevailing light conditions [154]. Indeed, Armstrong goes as far as to say that:

> "Different combinations of wavelengths may be instances that fall under some general formula. Such a formula would have to be one that did not achieve its generality simply by the use of disjunctions to weld together artificially the diverse cases falling under the formula. Provided such faking were avoided, the formula could be as complicated as we please."

No such formula or approach to a formula has ever been stated [83]. So, a central problem for an objectivist approach would be to identify a link between physical or objective colours with the colours we perceive objects to have.

It may be that there is some level of projection from the viewer onto the perceiver scene and that judgments of various visual contexts provides the key to the de-convolution of what is viewed and what is later identified in the associative cortex. For example if one *knows* that it is a bright, sunny day from apriori knowledge, then one may well expect the fields nearby to be brighter than usual and possess a certain luminosity. The issue of how one knows that it is a sunny day would be the assessment of context, in this case the general colour context, which would then fine tune the other algorithmic processes.

6.7 Derivation a General Context

6.7.1 Physiological Motivation

What is required is an algorithmic method of producing general purpose symbolic model which is flexible, extendible and symbolic in nature while keeping all of the important analytical ideas of the syntax of chapter 3 and the idea of contexts in this chapter.

However, if one is to take for example the modeling process that humans are likely to employ then there are also specific limitations and abilities to be considered :

- *Data presentation.* Scenes are presented without preparation or processing and at high speed.
- *Stereotyping.* We possess the ability to perform classification, without guidance, into arbitrary groups and sub-groups. This is often considered as a test of intelligence.
- *Storage.* Undoubtably stereotyping is one method of space saving inside the limited capacities of the human brain. It is also an efficient method of data retrieval since only differences from some central stereotype need be saved.
- *Rule Formation.* Rules for stereotyping may be formed arbitrarily as may rules to connect stereotypes of different kinds.

We require a mechanism for turning examples of scenes of a given type into an abstract multi-facet symbolic model.

6.7.2 Context Assessment Scheme

In order to judge whether a specific situation fits a given schema one may apply the conceptual evaluation rule proposed by Tversky [118] and used widely by concept theorists [117]. This model maintains that the similarity of two concepts is based on some function of the attributes shared by the concepts, minus the attributes that are distinctive to both :

$$s(a,b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A) \tag{6.1}$$

where a and b are the two concepts, s is their similarity, A is the attribute set of object a and B is the attribute set of concept b. $A \cap B$ are the common attributes shared by the two concepts, $A - B$ are the attributes that are distinctive to a, $B - A$ are the attributes distinctive to b. Note that this formula predicts that as the number of common features increases and the number of distinctive features decreases the two objects a and b become more similar. The function f has the role of a weighting modifier to weight certain attributes according to their salience or importance and the parameters θ, α and β are used to reflect the relative importance of the common and distinctive attribute sets.

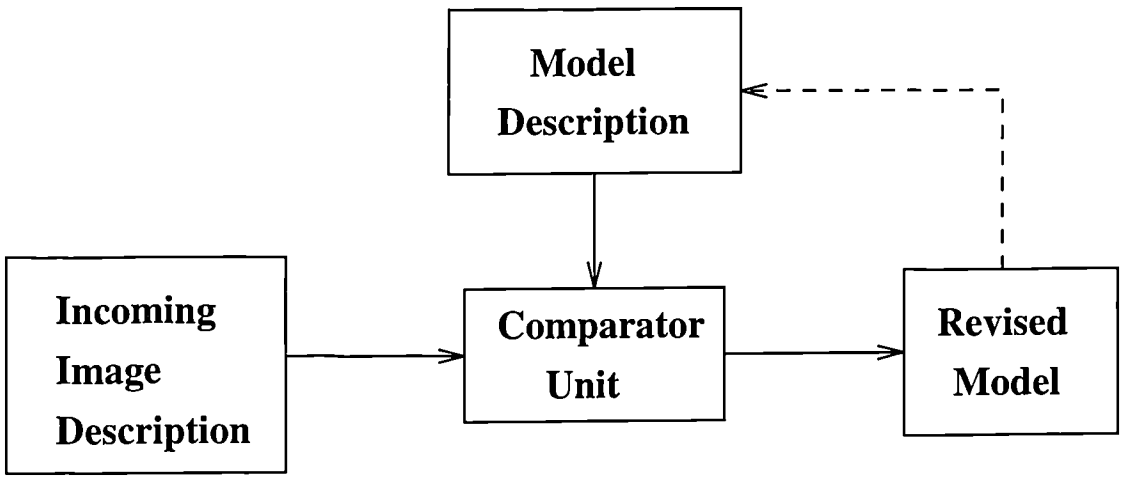


Figure 6.7: General Model Re-inforcement Correction Procedure

6.7.3 General Context Element Selection

Elements selected from the plane topology output from REDnet included in the general context were as follows:

- *Degree of fragmentation.* That is the ratio of object planes to lesser planes. Naturally, this is dependent upon the selection criterion for object plane selection.
- *Topology of object planes.* Their direct interconnectivity as well as derivable connectivity via other planes.
- *Plane relative positioning.* This is dependent upon the metric used for placement.
- *Connections of object planes to smaller planes,* which is related to degree of fragmentation.
- *Colours and internal mappings.* The hue and/or (R,G,B) designation of the plane is obviously an important factor for the model.

6.7.4 Selection Algorithm

The selection algorithm for model acquisition is guided by the idea of the Tversky similarity criterion. If the topological makeup of two images is similar then the common elements are considered as the starting point of a general model for that kind of scene. Distinctive features from all images should still be included but if these features are not included in subsequent presentations they gradually lose importance to the overall model.

Each of the object planes in the model is formed from features due to the presentation of previous images as shown in fig 6.8.

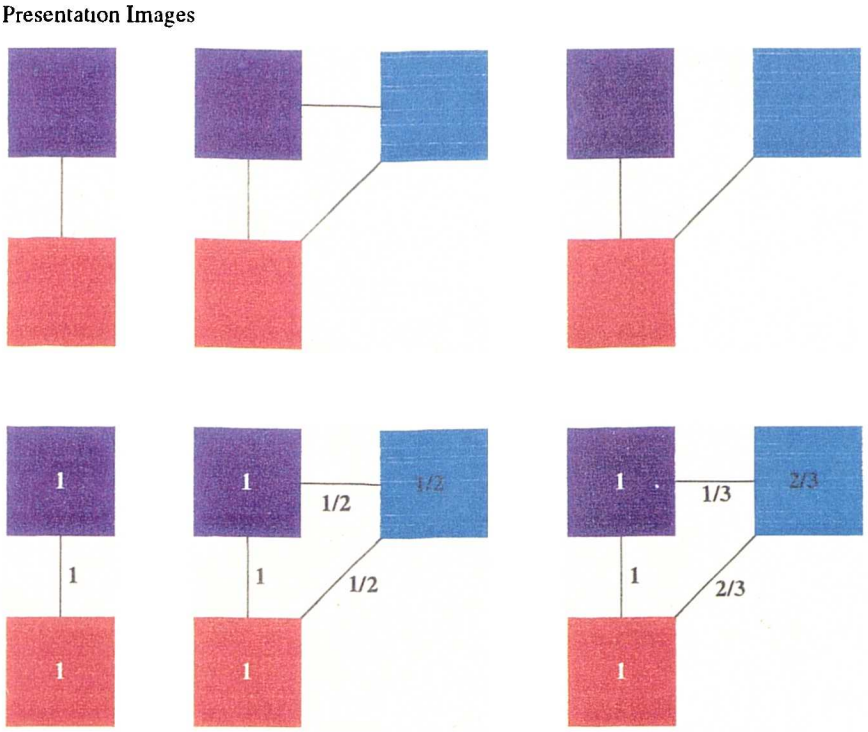


Figure 6.8: Presentation Images and the Development of a Simple General Model

6.7.5 The Comparator Function

Primarily, the best fit on a plane-by-plane basis was found by assessing the similarity function between the two planes .

$$s(a,b) = 1 - \prod_{i=0}^{d} \theta_i \frac{\sqrt{(c_i^a - c_i^b)^2}}{max(c_i^a, c_i^b)} \tag{6.2}$$

Where the function s is the normalized similarity function for the two planes a and b. The Euclidean distance between each of the d characteristics, c assessed and the highest value indicates the best fit to the model. The characteristic set may contain any of the elements of TDATA, as discussed in section 5.2.3. The weighting factor θ is also included in order to stress greater or lesser importance for each of the characteristic elements.

6.8 Summary

In this chapter, the importance of context as a *look-ahead* function as well as a paradigm for schema fitting. Some indications of both the biological inspiration and physiological plausibility for context assessment have also been discussed in this chapter.

Once an assessment has been made the results may be forwarded to other algorithms as parameters. This has the bonus that many of the algorithms will appear to become parameterless, as in the case of the human visual system. This will mean that a step has been taken toward the ultimate goal of

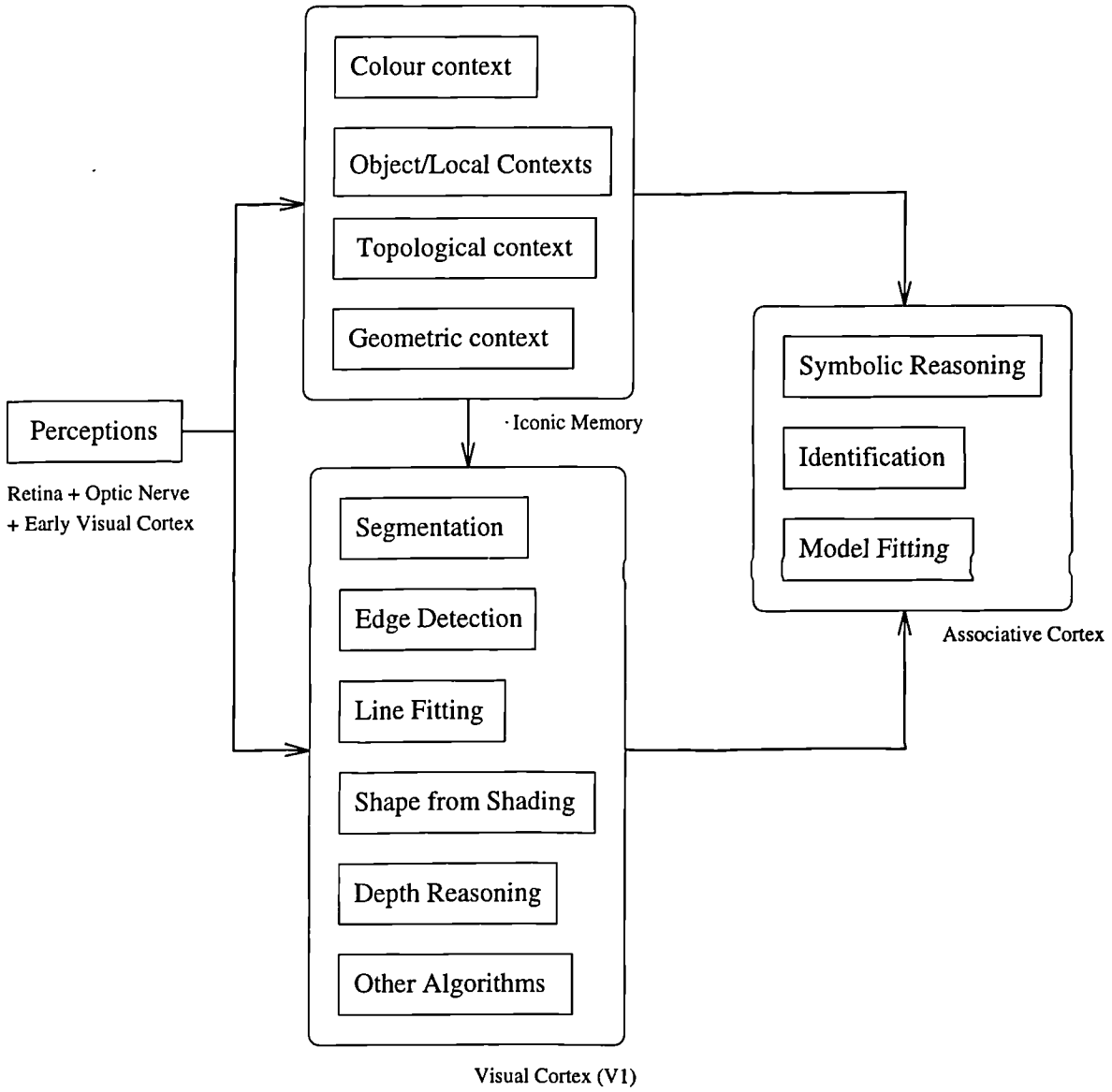


Figure 6.9: Visual Contexts in the General Vision Model

vision research - the actual emulation of human vision. The idea of how contexts fits into the general vision model is shown in fig.6.9.

One useful method of assessing context is a general context model. This model can give good indicators to a knowledge based system or model base as to the likely content of a scene as well as the likely structures to be present in that scene. These indicators include relative positioning, likely colour/hue content and their classifications within that scene. A training paradigm has been described that performs this kind of model stereotyping and several application have been given for such a paradigm.

In the next chapter, details of how this scheme may be applied to a database of natural scenes is described and extensions to the general context model are shown.

Chapter 7

Applications to a Natural Image Database

7.1 Application Overview

Natural scenes were chosen as an application space for the networks and ideas developed so far. This was primarily because it presented the possibility of naturally wide variation for input data both because of the organic nature of the scenes and the uncontrived nature of the light sources. For *conventional* methods there is simply no straightforward way of deriving either information on forms or general rules about the images presented. In addition, the selection of natural image exemplars is such that there would be reasonable distances between their possible hue ranges and so that they would represent the majority of possibilities in any natural scene. Details of how the images were sampled is given in Appendix II.

Exemplar images were first taken to form a hue-space database with which the cells in COLnet were to be trained. These images were taken with all of the frame occupied by the colour texture to be named. This data provides the inputs to the classification layer network used to classify plane data from the topology layer. The exemplar hue results are generalised in the space by neural networks with different architectures and generalisation methods. This generalisation then represents the colour bases of the general context model.

Segmentation of the images is performed by ICEnet as described in chapter 5 with the topology derivation performed by REDnet. The most important (object) planes then reported to COLnet where their possible name tag was determined and to the context agglomeration scheme where the color, topology and geometric context model is built.

The results in sections 7.4 and 7.5 show that a stereotyped general context model can be built automatically from a collection of images that is rich in information and can be used in further processing

and algorithm selection. Also, the model can be used to identify the *type* of scene that is presented probablistically by matching its context model with a fairly small collection of general models.

7.2 Exemplar Images Used for Training COLnet

In order to generate name-tags (symbol names) for the patches in the general context model, examples of different types of images were taken and their normalised colour properties were derived. There were around 100 of each image and their properties, together and classifications were used to train the labelling neural networks. The images fell roughly into three categories:

Clouds : Many different types of cloud were photographed under a wide range of lighting and weather conditions. Some images include small areas of sky and some "bleed-through" of sky due to their composition. By the very nature of clouds it was not possible to judge exact distances so a wide range was used.

Foliage : In this instance foliage is a general term for both grass and many different kinds of trees. These were photographed under a wide range of lighting conditions and in varying stages of growth and sample distance. This was done in order that fixation did not take place and skew the data sample towards one kind of foliage spectrum. Some overlap with both sky and cloud also took place.

Sky : Clear sky was sampled at many points from horizon to zenith and at many different times of the day as shown in fig 7.1. Each image is cloudless to the best possible extent although some overlap with the cloud images was inevitable.

It can be seen that the results for foliage are spread over a fairly wide range and partly mix with both sky and cloud. This is to be expected since the images of foliage occasionally include both of these. The group describing sky is perhaps the best defined of the three and intersects only with cloud.

7.3 Hue Spectrum Histograms

This section contains some representative example hue spectrum histograms. The histograms have been sampled at a 25×25 resolution which is representative of their 256 colour (8-bit) depth. The phi-theta axes are from 0 to $\frac{\pi}{2}$ and the height axis is in % of image pixels.

Cloud : Images shown in figs. 7.3 and fig. 7.4. The majority of points are in the monochrome (central) region, as would be expected since clouds are mostly white to dark grey. The colour variation across the entire cloud image data-base was quite large due to weather and light conditions.

Foliage : Images shown in figs. 7.5 and fig. 7.6. The foliage spectrum extends over the green part of the spectrum and has a broad base meaning that a range of greens are present in the image, extending into both brown and yellow. Colour variation in the entire foliage image database was surprisingly low. This is probably due to the light-energy collection efficiency of foliage pigmentation being relatively stable at set wavelengths.

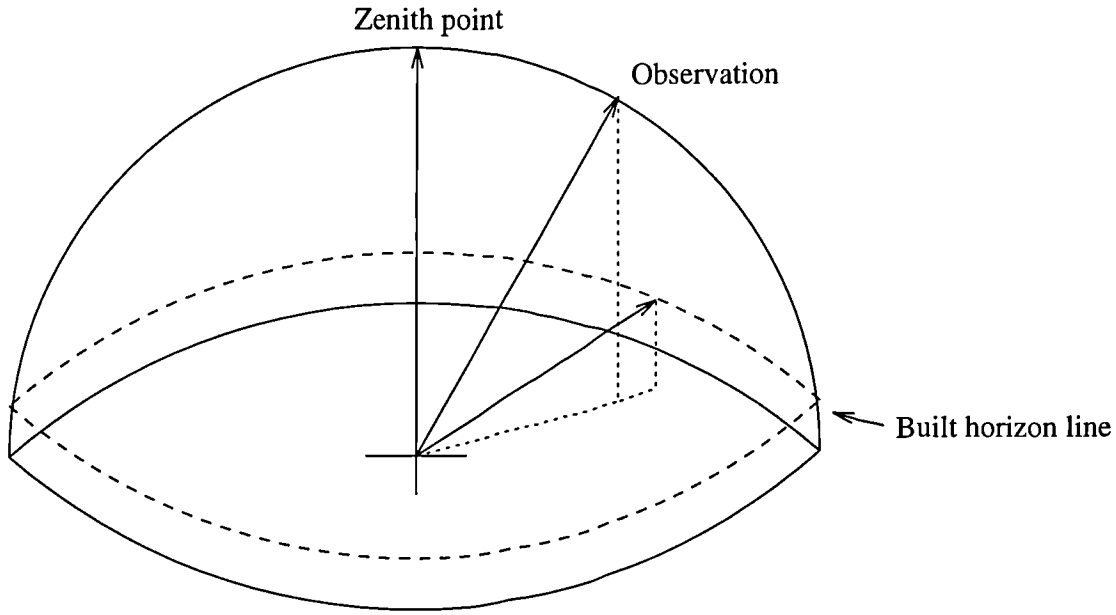


Figure 7.1: Sky Observations

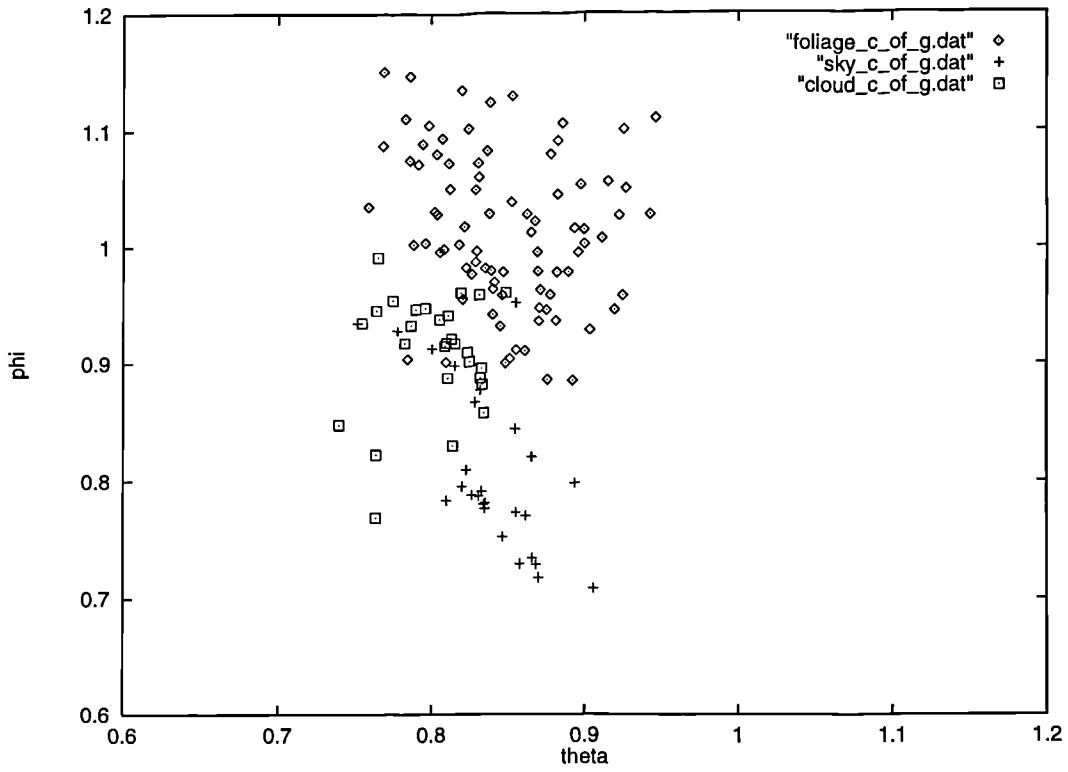


Figure 7.2: First Moment in 2-Dimensions

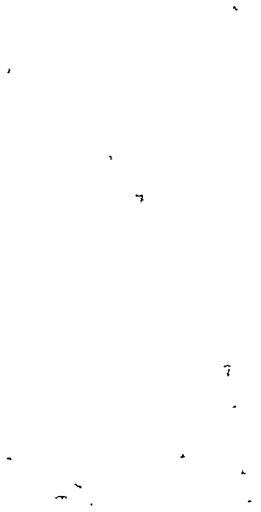


Figure 7.3: Cloud example

Sky : Image shown in figs. 7.7 and fig. 7.8. It seems obvious from the spectrum that the majority of pixels in the image are in the blue region. However, the question arises as to the proximity of the central peak to the centre of the spectrum. This is due to the low saturation [14] level in this particular image. Light from the sky tends to vary widely in its white content with times in the day and position.

7.3.1 Moments Data

Moments data was used in 2-D format, as shown in figure 7.2. The output data for training were the orthogonal bases: Foliage 100; Sky 010; Cloud 001. This basis gives equal distances between the classification locations in the decision space. If an image is to be used as an exemplar the first moment in each dimension of the spectrum-space is calculated. That is :

$$\phi = \frac{1}{N} \sum_{j=1}^{j=N} \phi_j . \rho_j \tag{7.1}$$

$$\theta = \frac{1}{N} \sum_{j=1}^{j=N} \theta_j . \rho_j \tag{7.2}$$

where N is the number of samples from the space and ρ, θ and ϕ are the usual hue directions and amplitude as described in [80].

7.3.2 Training Parameters and Algorithms

The following algorithms were used, with several parameter settings and at several different levels of iteration and several topologies quick-propagation; back-propagation with momentum. These algorithms are widely used and can be found in [7], [8] with further theory on optimisation and convergence.

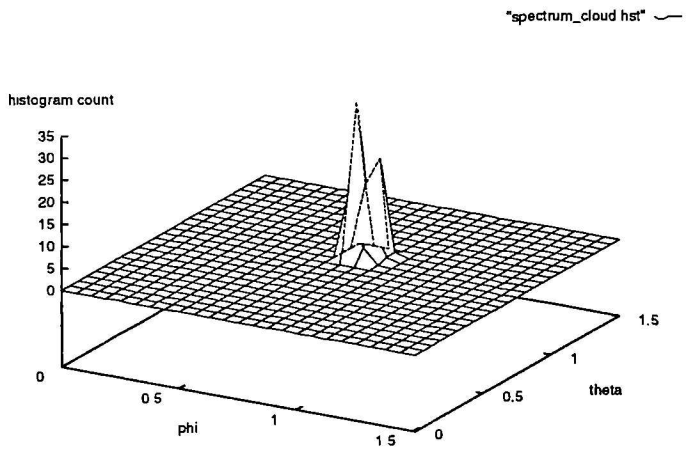


Figure 7.4: Example cloud hue spectrum

Figure 7.5: Foliage example

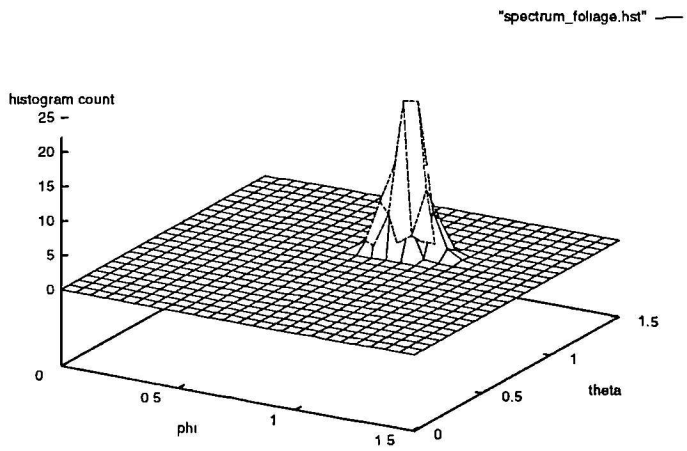


Figure 7.6: Example foliage hue spectrum

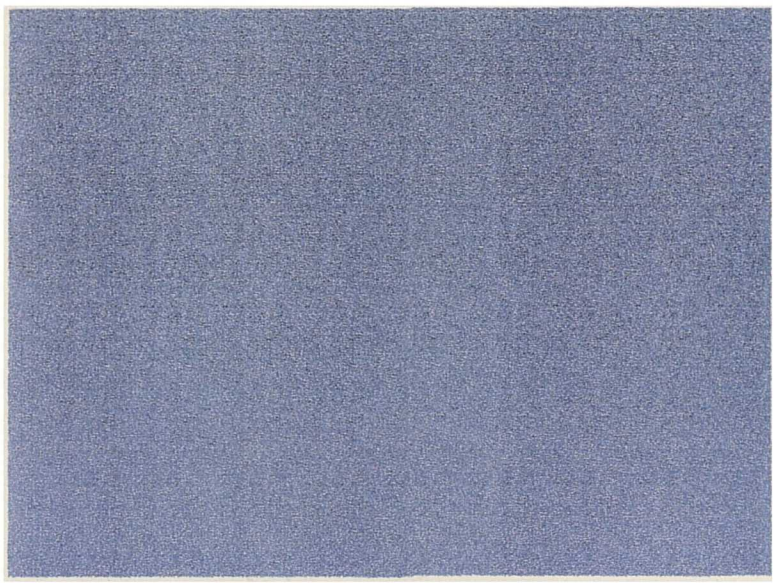


Figure 7.7: Sky example

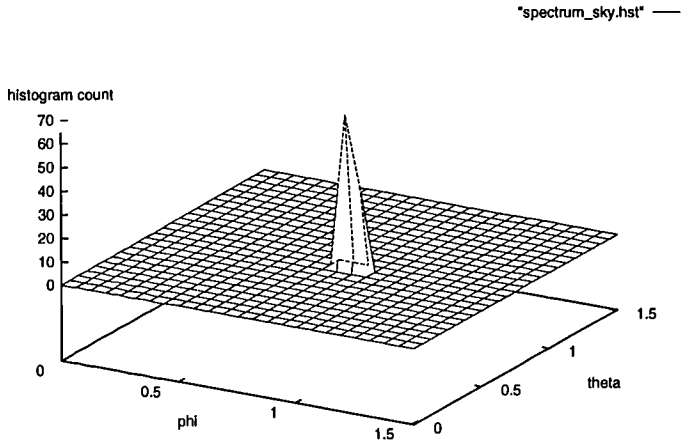


Figure 7.8: Example sky hue spectrum

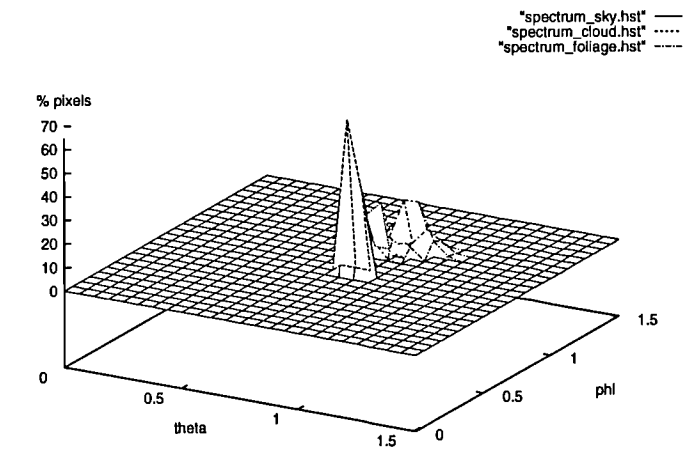


Figure 7.9: Hue spectra for all three examples

Each network was allowed to learn for a sequence of between 1000 and 5000 presentations of the input data (cohort iterations). This allowed for generalisation without the classic symptoms of over-specification.

Quick-propagation Results

By far the best results came from the quick-propagation algorithm, which is an optimised version of standard back-propagation.

The learning rule takes the standard back-propagation form:

$$\Delta w_{i,j}(t+1) = \eta \delta_j o_i + \mu \Delta w_{i,j}(t) - \nu w_{i,j}(t) \tag{7.3}$$

as discussed in [8].

In this test the best parameters were found to be as follows:

$\eta = 0.5$

$\mu = 1.75$

$\nu = 0.0001$

where ν, the weight decay term is obviously kept small. These parameters are not necessarily the optimal choice but the networks were seen to converge and provide good generalisation. A wide selection of parameters were tried with varying degrees of generalisation.

Quick-propagation also computes the gradient of the weight space in the direction of each weight, afterwards a direct step to the error minimum is attempted by:

$$\Delta(t+1)w_{i,j} = \frac{S(t+1)}{S(t) - S(t+1)} \Delta(t) w_{i,j} \tag{7.4}$$

where

$w_{i.j}$ is the weight between units i and j,

$\Delta(t+1)$ is the actual weight change,

$S(t+1)$ is the partial derivative of the error function by $w_{i,j}$

$S(t)$ is the partial derivative

Using a $2-4-3$ topology (fig. 7.11), the trained network had a predictive capability of 100% on the training data of 72 images and 94.4% on the remaining testing data. Precisely the same results were found with a $2-5-3$ topology (fig. 7.11), implying that $2-4-3$ is sufficiently specified for the classification.

Back-propagation with Momentum Results

In this test, the standard momentum learning rule was used :

$$\Delta w_{i,j}(t+1) = \eta \delta_j o_i + \mu \Delta w_{i,j}(t) \tag{7.5}$$

with the best parameters found to be :

$\eta = 0.6$

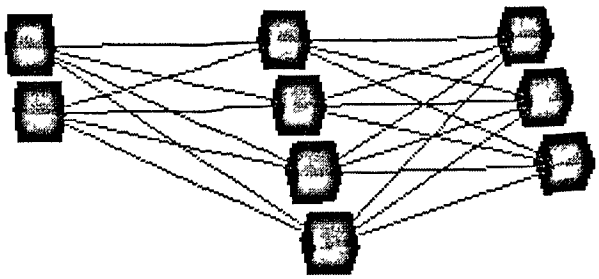


Figure 7.10: Feed Forward Neural Network with 2-4-3 Topology

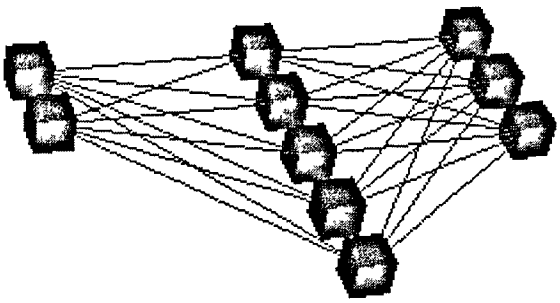


Figure 7.11: Feed Forward Neural Network with 2-5-3 Topology

$\mu = 0.3$

When using back-propagation with momentum and a $2-4-3$ topology, 100% was achieved on the training set with 91.5% on the testing set. A similar result was found with a $2-5-3$ topology. Once again, it seems that the best results come from a $2-4-3$ topology.

7.3.3 Discussion of COLnet training

It is inherent in neural network learning that over-specific learning can destroy the generalisation capabilities of the network so examples of this sort have been avoided and the best results only given. Similarly, overly specified networks with many degrees of freedom have been avoided since they are prone to very complex and contorted decision boundaries which, although are fine for training set tests, fare very badly with unseen (test) data. On close examination, the misclassified examples proved to be those which are just inside the decision boundaries of groups.

Once the COLnet is trained on exemplar colourings it can be used to generate labels for the context regions found in the general context model. In our case, these labels are fairly simple but there seems to be no reason why significantly more complex labelling (or name-tagging) cannot be generated using

Smoothing Level	Average No. Planes	Average No. Active Tweens	Av. Exemplars	Av. Tweens per Plane
9	744.9	11,236.4	4.6	15.1
7	727.6	11,570.1	4.6	15.9
5	797.0	13,943.9	4.5	17.5

Table 7.1: Results for Natural Images

other more robust characteristics of hue spectrum histograms.

7.4 Test Natural Images

Natural test images containing all three kinds of exemplars were sampled and processed by the three layers of the network, as shown in fig 7.12, as well as the context agglomerator. This provided a great deal of information about the scene, as described in chapter 5, as well a high level general context model.

Example results from each network layer are shown in figures 7.14 to 7.20.

7.4.1 Segmentation and Recolourings

Segmentation of the test images was performed with parameters set as follows:

- Smoothing was using a Gaussian kernel of sizes 5, 7, and 9 pixels. This provided some noise tolerance.
- Maximum number of exemplars per image, 30. This is set high enough that after palette reduction a representative number of exemplars still remain.
- Best yield for exemplars, 30%.
- Palette reduction heuristic on.
- Palette reduction heuristic similarity tolerance at 80%.
- Small plane contraction on (to remove noise) with a tolerance of 5 active nodes.

7.4.2 General Region Adjacency Graphs

Example region adjacency graphs derived from the example image are shown in figs. 7.17 to 7.19.

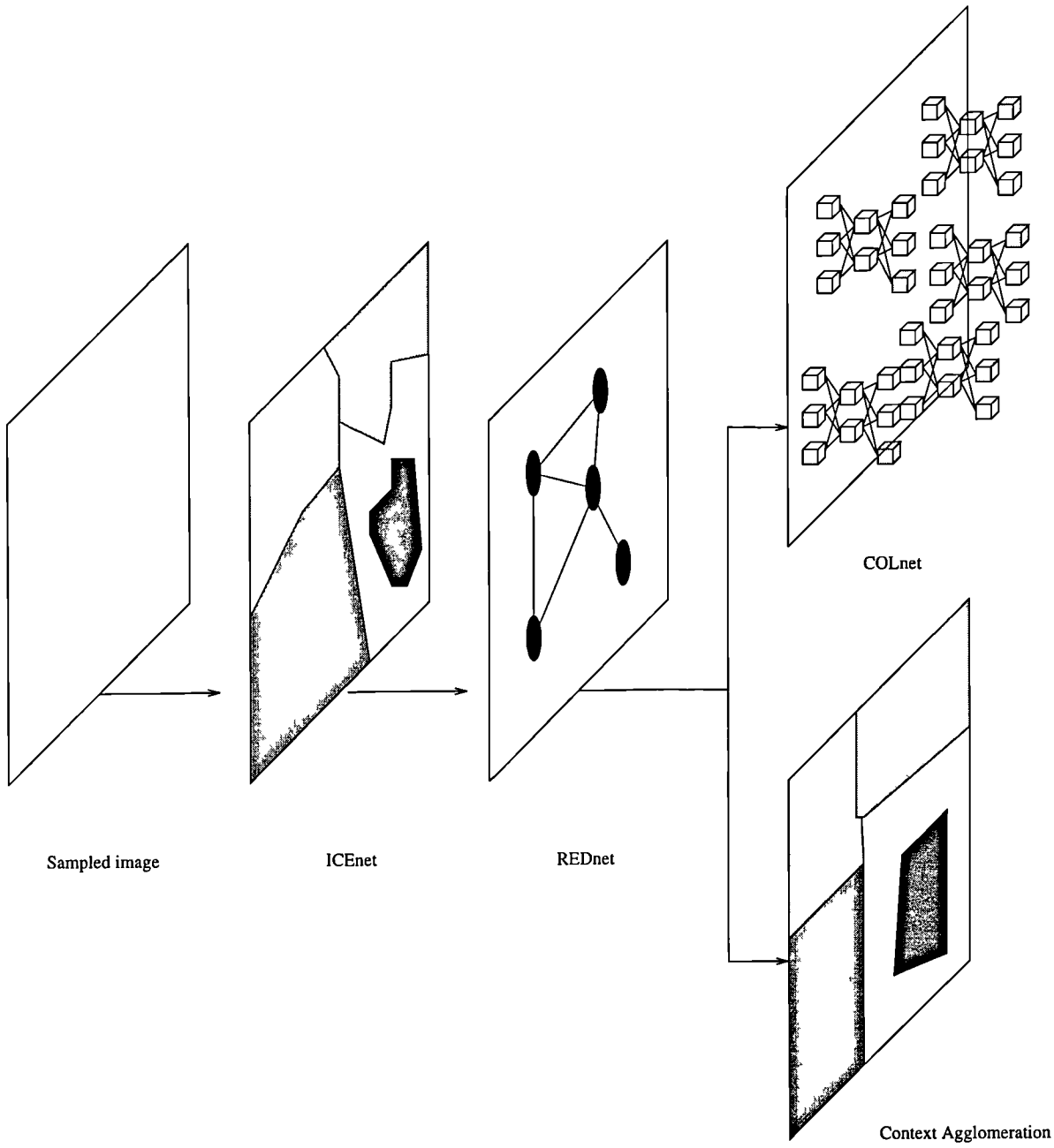


Figure 7.12: Full Processing System

Figure 7.13: Example Test Image

Figure 7.14: Smoothed Image

Figure 7.15: Segmentation

Figure 7.16: False Coloured Image

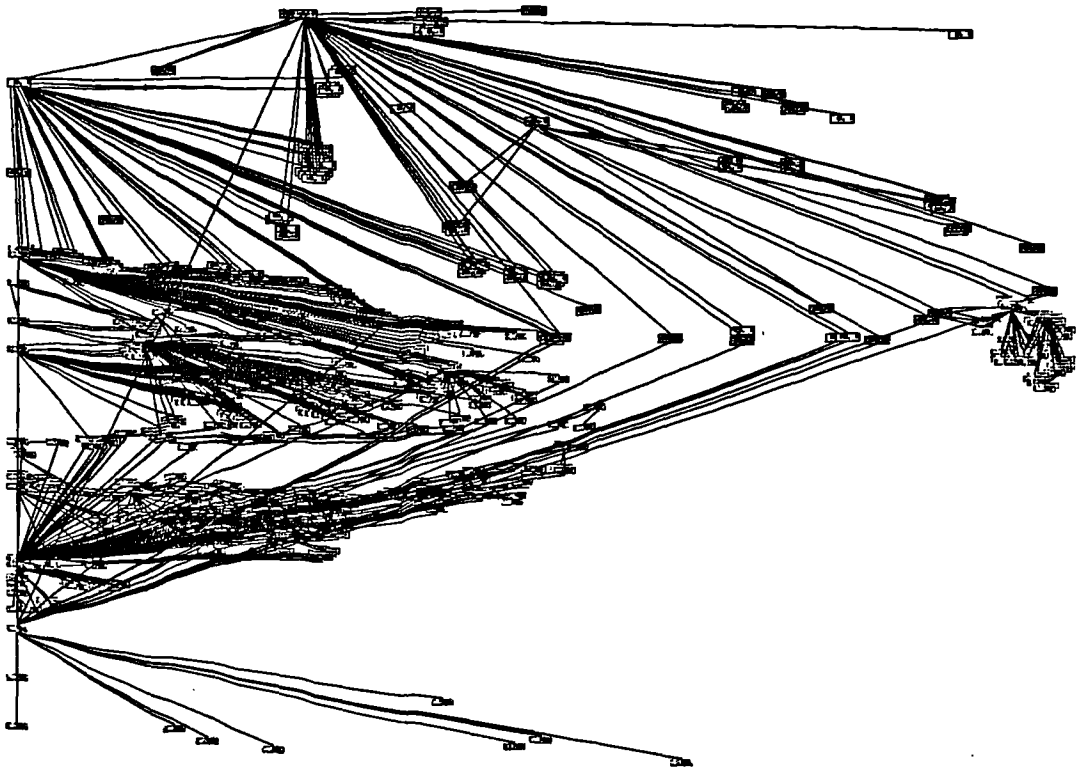


Figure 7.17: Example RAG Showing topological connections, colour and position

Colour, geometric and topological context

The context RAG for colour, geometric and topological context for the example image in fig. 7.16 is shown in fig. 7.17. It can be noted that without any specific region sizes in the graph is difficult to understand which regions are important. However, in the graph in fig. 7.18 relative areas are shown to highlight this. A graph of colour and position of planes is shown in fig. 7.19.

As an example of a derived property, a RAG showing saturation level, colour and area is shown in fig 7.20. Planes are ranked vertically with lowest saturation level at the top.

7.4.3 Language Outputs

Typical output from the REDnet layer is shown below. It is choked in such a way that any topology node of less than 5% of the total image area will not report its findings. This choke reduces output significantly. Note also that the nodes also analyse their area to deduce their own importance to the scene. For instance, any node representing a plane with area of greater than 50% of the image area considers itself to be the default background. Other descriptors are:

- area < 0.1% : *likely to be a speckle.*
- area >= 0.1% and area < 5.0% : *a small plane*
- area >= 5.0% and area < 20.0% : *a fairly large plane*
- area >= 20.0% and area < 50.0% : *a very large plane*

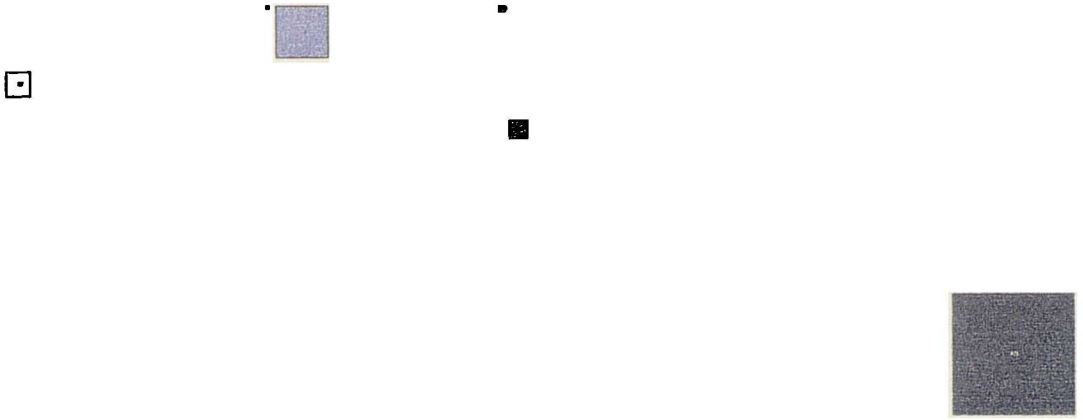


Figure 7.18: Example RAG showing colour, position and area

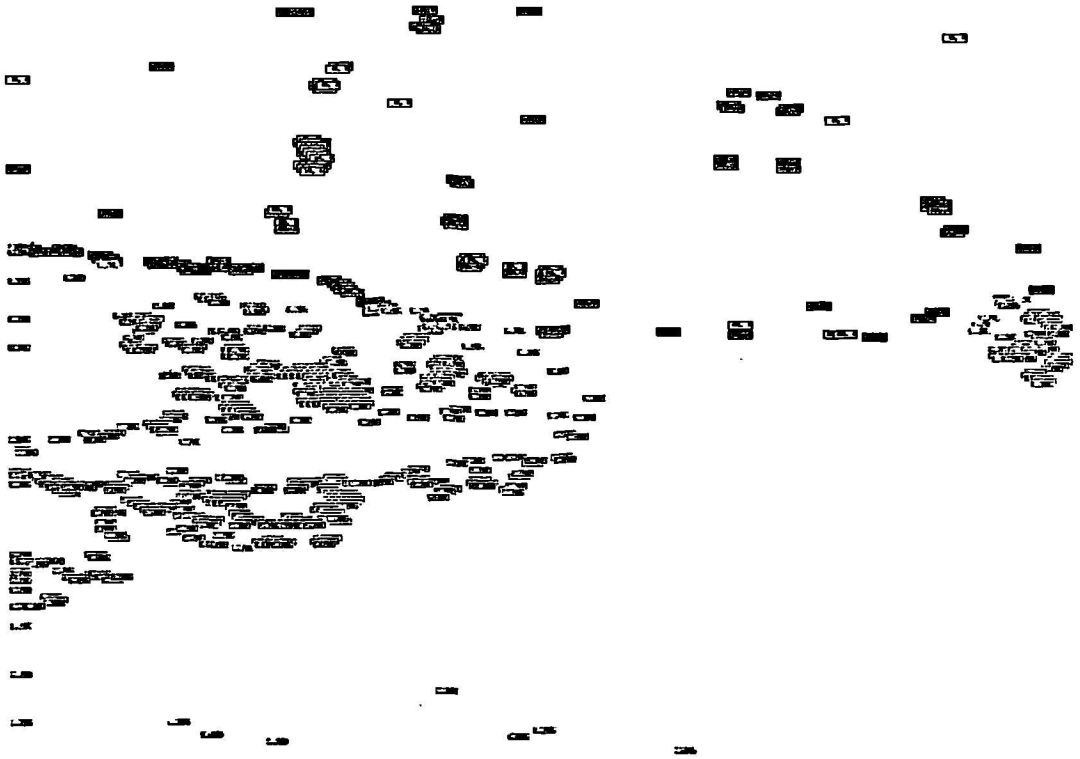


Figure 7.19: Example RAG showing colour and position

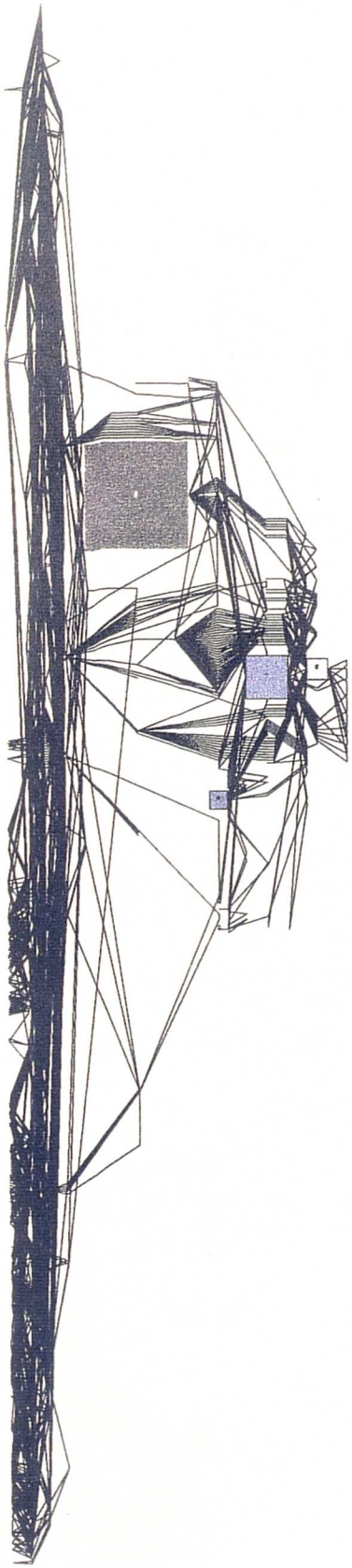


Figure 7.20: Saturation RAG

- area >= 50.0% :*a huge plane (default background)*

```
HEADER START

% Type 6 Description file
%
% Copyright C. Robertson 1997
%
% Original image      : natural_scene_80.pcx
% Analysis performed : 30/01/97
% Parameters
% Heuristics ON : Small plane reduction
% Output choke  : > 5% of images area, unordered
% Latex state   : OFF
% HTML state    : OFF

HEADER END

BODY START

================================================

Plane number   3
Colour palette value is   6
State is ON
Start X coord is   87, Y coord is    0
Plane area is 13305 units
which is 16.428365 perc. of image
Size analysis suggests its a fairly large plane
R=148, G = 156, B=252
Phi =  0.70641, Theta =  0.81171, Rho = 331.27632
Saturation is 0.336551
```

```
Intensity 2 is  2.180392

Perimiter length is 3482 units

Total in line set = 284

Potential depth is level   6

Line set :

 402  403  404  430  501  501  502  519  532   34   35   36   37

  38   39   40   93  112  115  117  122  126  128  132  137  138

 142  147  148  149  150  151  152  153  154  155  174  179  186

 197  204  209  217  225  229  233  238  242  245  251  270  301

 305  309  315  317  319  323  356  365  408  411  428  434  446

 449  486  525  527  542  566  950  951  988 1005 1023

Plane set :

   1    1    4    9   17   17   22   29   29   29   29   29   29

  29   29   29   30   51   54   55   60   63   64   65   67   67

  74   80   80   80   80   80   80   80   80   80   82   86   92

 103  110  113  120  127  130  132  137  141  143  148  166  197

 204  208  218  219  220  229  255  265  271  273  294  302  309

 311  330  354  355  372  388  442  475  490  501  501  501  501

 501  501  502  517  517  518  519  532


================================================

Plane number  29

Colour palette value is   5

State is ON

Start X coord is    0, Y coord is   21

Plane area is  6278 units

which is 8.165648 perc. of image

Size analysis suggests its a fairly large plane

R=236, G = 234, B=236

Phi =  0.95332, Theta =  0.78114, Rho = 407.61256

Saturation is 0.241475

Intensity 2 is  2.768627
```

```
Perimiter length is  945 units

Total in line set =  69

Potential depth is level   7

Line set :

  32   33   34   35   36   37   38   39   40   41   42   43   44

  45   46   47   49   50   51   52   54   56   59   60   61   62

  63   64   65   66   67   68   69   70   71   72   73   75   76

  81   82   83   85   86   87   88   89   92  333  334  335  336

 337  358  936  940

Plane set :

   3    3    3    3    3    3    3    3    3   53   59   62   95

  96   98   99  104  105  109  119  199  207  223  239  244  245

 250  251  252  258  262  263  264  269  272  276  277  279  287

 300  310  312  314  343  350  351  352  489  240  240  240  240

 240  257  488  490


================================================

Plane number  80

Colour palette value is   4

State is ON

Start X coord is  158, Y coord is   34

Plane area is  6025 units

which is 7.858698 perc. of image

Size analysis suggests its a fairly large plane

R=108, G = 102, B=180

Phi =  0.68998, Theta =  0.75683, Rho = 233.38380

Saturation is 0.482840

Intensity 2 is  1.529412

Perimiter length is  660 units

Total in line set =  27

Potential depth is level   5

Line set :
```

```
 147  148  149  150  151  152  153  154  155  157  158  159  162
 164  165  167  171  173
Plane set :
   3    3    3    3    3    3    3    3    3  135  144  167  203
 210  210  221  228  370

==================================================

Plane number 501
Colour palette value is   0
State is ON
Start X coord is    0, Y coord is  192
Plane area is 30523 units
which is 40.000525 perc. of image
Size analysis suggests its a very large plane
R=63, G =  98, B= 102
Phi =  1.01292, Theta =  0.73689, Rho = 171.89823
Saturation is 0.595517
Intensity 2 is  1.164706
Perimiter length is  789 units
Total in line set =  38
Potential depth is level   4
Line set :
 950  951  952  953  954  955  956  957  958  959  960  961  962
 963  964  965  966  967  968  969  970  971  972  973  975  976
 978  979  981  983  984  985
Plane set :
   3    3    3    3    3    3  249  249  351  372  374  488  489
 490  502  517  518  519  532  710  741 1360 1361 1362 1364 1364
1366 1367 1369 1371 1372 1372

BODY END
```

7.4.4 Application of COLnet

The trained networks were applied to the outputs from the REDnet topology nodes. A graph of the inputs to the trained COLnet with the inputs to it are shown in fig. 7.21. It can be seen that the spread of the input data is much wider than that of the training examples. This is not a problem since the trained network has *generalised* its training set in order to produce its decision boundary. Foliage inputs from the natural scenes seem to be much redder than those in the training set but to what extent this is an artifact of the sampling method is not clear.

Typical output, for the figure in 7.13, and detailed in the previous section, is given below:

```
HEADER START

% Type 7 Description file
%
% Copyright C. Robertson 1997
%
% Original image      : natural_scene_80.pcx
% Analysis performed : 30/01/97
% Parameters
% Heuristics ON : Small plane reduction
% Output choke  : > 5% of images area, unordered
% Latex state   : OFF
% HTML state    : OFF

HEADER END

BODY START
================================================
Plane number   3
Suggested name-tag: Sky

================================================
```

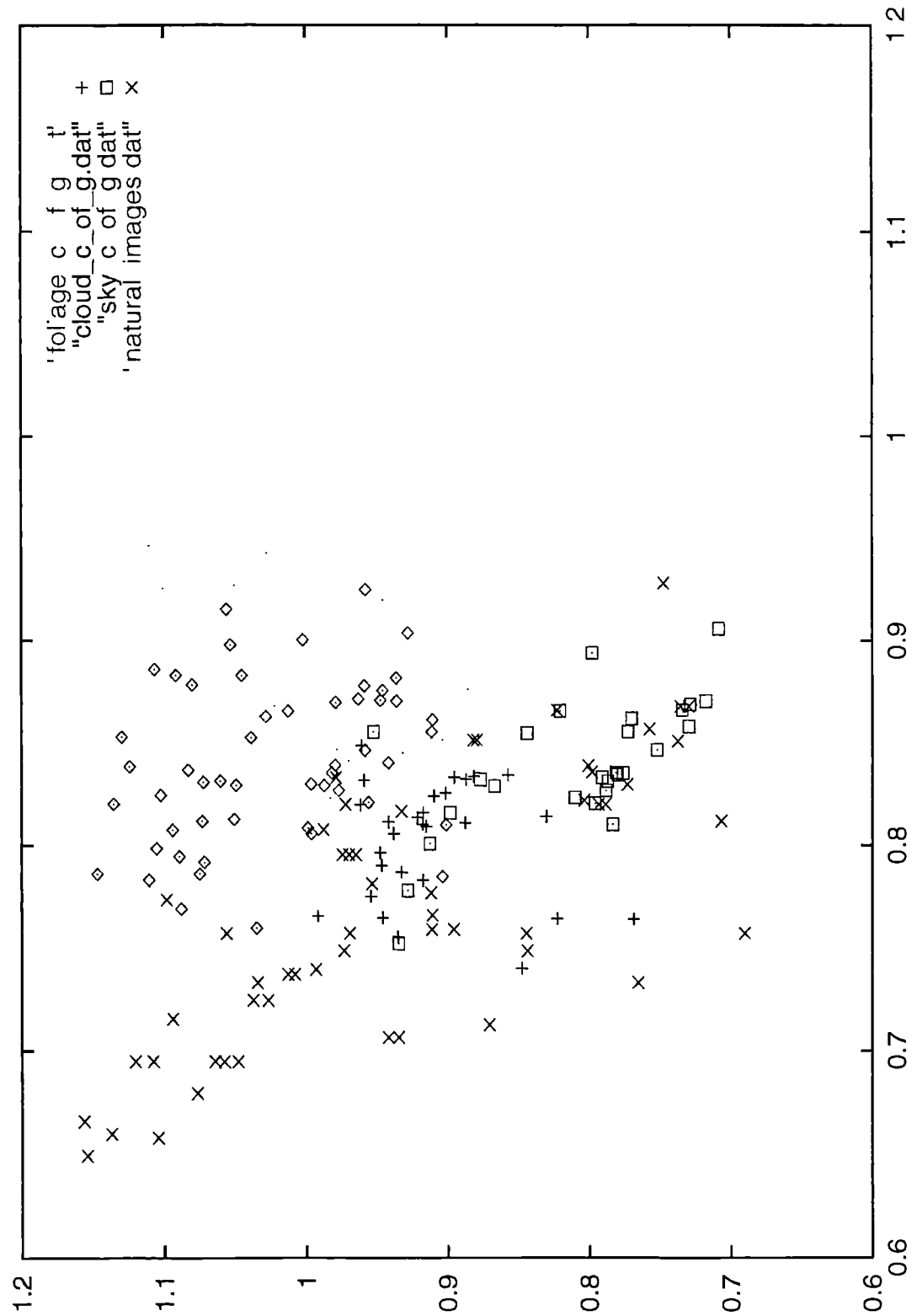


Figure 7.21: Input and Training Data for COLnet

```
Plane number  29

Suggested name-tag: Cloud


================================================

Plane number  80

Suggested name-tag: Sky


================================================

Plane number 501

Suggested name-tag: Foliage


BODY END
```

7.5 Context Model

Construction of the model consists of several steps, as detailed in section 6.7 :

- Hue segmentation, by ICEnet.
- Derivation of largest, most important nodes in the RAG, by REDnet.
- Derivation of direct or indirect links between these nodes.
- Derivation of general form of nodes including connections to types of less important nodes.
- Assessment of overall fit to previously stored model.
- Assessment of update required to stored model to increase generality.

This agglomeration was performed on all of the general images in the database in order to create the natural scene model. The resultant model position, colour and area RAG is shown in fig. 7.22. It can clearly be seen that the predominant structure is blue at the top with some patches of white/grey with smaller patches of brown/green at the bottom. This is not totally unexpected for a natural scene and as such seems to be representative of the whole database.

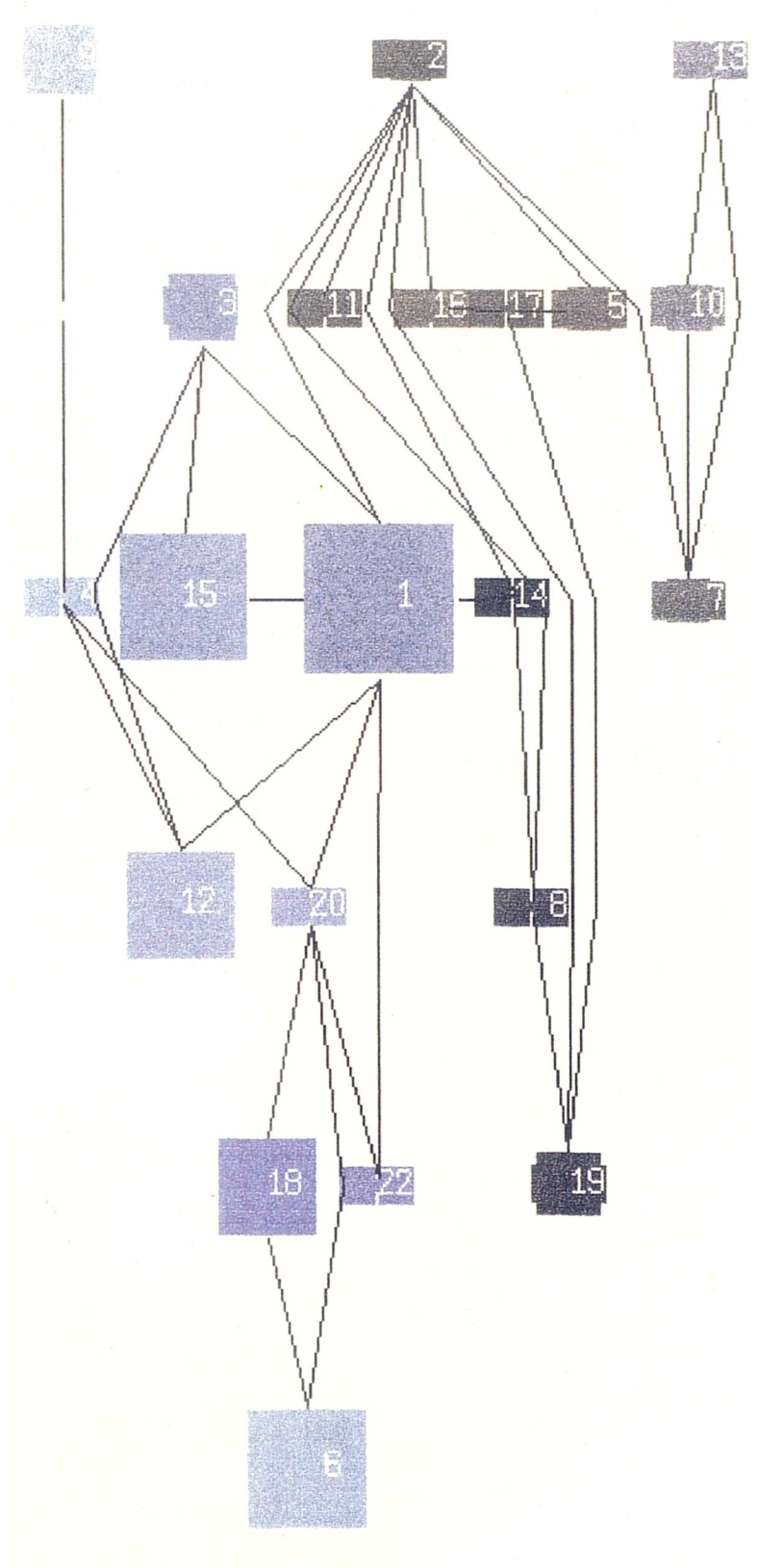


Figure 7.22: Context Model

7.6 Summary and Discussion

7.6.1 Application of COLnet

Under certain circumstances elements in natural images can be classified using the first moment of their hue-spectra in conjunction with an unsophisticated neural network. COLnet provides a straighforward classification and name-tagging part of the system that is still network based and can be implemented easily in parallel. As well as the first moment data descriptor there are many different ways of classifying the spectrum shape with statistical parameters. Two other shape classifiers were used for reinforcement in the neural network, the second and third moment (skew and kurtosis). This proved to have little effect on classification however. Neural algorithms were also applied directly to hue spectra histograms but the classifications resulting from this were due mainly to over-training (or under-generalising) due to the large number of degrees of freedom, therefore their predictive capabilities were not as strong as those of more simple networks. The main problem with the first moment dataset is the degree of overlap. Although this is predictable in the case of the sky-cloud boundary it can also be noticed that there is overlap in the cloud-foliage and sky-foliage interface. This may be due in part to the fact that none of the images were taken to be "canonical" and so contain certain elements from the others.

7.6.2 Application of ICEnet

ICEnet has been tested extensively both on human face data and on natural scene data and has shown that a network structure can provide very fast image segmentation at tolerable resolution and quality. Segmentation heuristics have been applied which mean that even a degree of noise tolerance is demonstrated. The network also has the added advantage that it is inherently based on the idea of a supergrid structure and is easily implementable in a parallel way. However, the main reason that ICEnet is employed is that it provides in an immediate way the low level primitives of the scene description langauge detailed in chapter 3. These primitives build up a symbolic representation of the scene that can easily be used in symbolic reasoning approaches to vision, such as context assessment and network-based reasoning approaches to vision.

7.6.3 Application of REDnet

The REDnet network has proved to be both a versatile and fast method of topology extraction when used to analyse the primitives provided by ICEnet. It is only one of many possible networks that may be used to analyse this output. Once again, the design emphasis was on the parallel qualities of the network.

7.6.4 Context Rule Agglomeration

The probablistic context agglomerator is the final part of the scheme and provides a simple but effective means of rule generation from a selection of syntax elements. As such it has generated a context model of natural scenes that demonstrates its usefulness. This is evidence that given a relatively small number of frames of examples it is possible to generate an evidential model based only on colour and topology that may be used to fine-tune the parameters of a large number of segmentation, recognition and expectation algorithms.

Chapter 8

Conclusions and Further Work

8.1 Conclusions on Modeling Human Vision

Modeling human vision is analogous to modeling human thought and it is not possible to easily separate the two. The question *How do human beings see ?* is well known to be an ill-posed problem (for example [83]). One may only generate models of approaches and test them against some efficiency metric, a process referred to by Turing [84] as a kind of model-child rearing. This, he contends, is split into two separate parts : the construction of the child program itself, the algorithm that operates upon the incoming data ; the education process or the selection mechanism for suitable data. Once this has been done, the efficiency of the model may be determined. It is how the suitability of the model may be assessed and a better model designed. The loop is repeated until convergence is achieved between the real system and the model behaviour.

This process is only slightly different in the case of neural, distributed or learning systems, with only the choice of algorithm changed. In the best learning systems data is fed to the system unprocessed and one no longer needs to choose its form or assess its worthiness. If the data is to be treated this way then the choice of algorithm is critical as it must be able to categorize and revise its entire data structure in the light of new data. That is why a general colour context model has been proposed and implemented in this thesis. It is the basis for a system of modeling that is flexible and able to incorporate many different processes, in the "symbol world sense", for symbolic extraction. It is not proposed, however, that this is a full model for human perceptual organisation since the physical

storage of such perceptions remains opaque and will probably remain so for some time to come.

What is needed for a model of perceptual organisation is a system of symbols which is scalable and has the ability to expand both in depth and breadth. That is a system which has rules which can be constantly expanded and which will generalise away data that is redundant through repetition while at the same time be able to accommodate novel and perhaps even contradictory data. Clearly then, this modeling and storage process goes beyond the remit of the simple image $\rightarrow$ processing function $\rightarrow$ image type processing and requires a hierarchy of symbols to describe images from the lowest to the highest level. The lowest level symbols are the form that the input data would take in the model-child regime and further processing represents the construction of higher level meta-symbols in order to generalise the data, or the child algorithm. The *syntactical language* described in this thesis is an attempt to perform this description and the construction of the meta-symbols is performed both by training and by the application of heuristics.

It is beyond doubt that human perceptions are self organising and self regulating with little redundancy. For this to be so there must be some mechanism of attribute selection and scene generalisation or stereotyping. This would enable an expectation based perception, as proposed by the Gestalt school (for example Katz, as discussed in Schilder [40]). Indeed, when a natural scene is viewed, it is unlikely that every leaf on a tree is examined or even perceived before reaching the conclusion that it is a tree. There must exist a "broad brush" approach for primary recognition of both the nature of the scene and elements in the scene itself. There is a great deal of evidence from disparate sources that this is the case including the puzzle of how people with reduced visual acuity can still recognise objects and how one can still recognise objects in an impressionistic painting at a level of detail so low that only vague planes exist. Excellent examples of this phenomenon are to be found in the early work of Kandinsky, see fig.8.1, Degas, Monet and many others. The model for this recognition is proposed here as a *general colour context model* with varying levels of detail (in local contexts) and as such contains both colour and topological parts. Dominant themes in the image form the lowest level of this model and further levels of detail may then be added at a later stage by compounding planes into object contexts using structural data, perhaps from a Marr-like filtering scheme.

It has been demonstrated here that it is possible to compute using the symbolic hierarchy at both the lowest and the highest level and generate rules with which to perform this kind of computation.

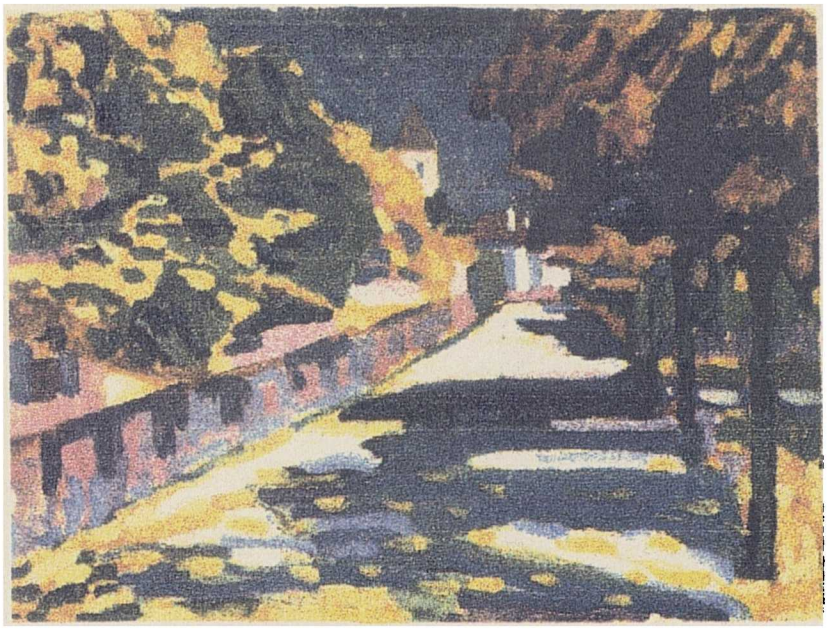


Figure 8.1: Kandinsky: Autumn in Bavaria

8.2 Achievements

This work has formulated an approach to modeling the colour functioning of the human visual system, from acquisition through segmentation to storage and basic analysis. A syntax has been developed and a hierarchical structure for symbolic representation presented. Work has been done to develop a parallel distributed segmentation network (ICEnet) with three complex post-processors which mimic the human parvocellular layers and the primary processing layers of the visual cortex as well as the instantaneous context agglomeration of the iconic store. All four systems, the segmenter ICEnet, topological reduction network REDnet, the colour processing network COLnet and the context algorithm are designed to perform their functions in very few processing steps and in parallel.

Post-processing heuristics are discussed together with some possible applications to object building and light/shading. For real application, a large database of natural images has been collected and analysed and rules for a general model of the database constructed.

8.3 Further Work

Further work on this subject falls into the following distinct areas:

- *Symbolic acquisition.* The sets of context grouping rules outlined in chapter 6 are currently being expanded and strengthened to include the local contexts shown in Appendix IV and a general expression of the syntax is being taken into the third dimension.

- *Hardware.* A field programmable gate array design is being produced to perform all three of the functions of the three parallel distributed networks. This will also allow for the fusing of stereo images, as discussed by Marr and Poggio [85] - in their case for structure.

- *Long term goals.* The generation of more sophisticated algorithms to perform both segmentation and post-processing are being examined. Also the generation of a complex symbolic storage system is now being addressed.

- *Use of context RAGs for data mining.* This is possible because of the generalisation capabilities of the storage mechanism. Examples of, say, pictures of humans may be stored and retrieved by simply building novelty onto a generalised model of the human face. This leads on to

- *Iconification.* Iconification of images may be achieved for data reduction by generalisation and use of the high level model.

- *Classification using RAGs.* This may be achieved by representation of RAGs as n-dimensional points for classification.

Epilogue

Historical Note on Artistic Representation

Classical art academies, such as those of the Italian Rennaisance, have directed students towards an algorithmic approach for the representation of colour [60]. This has expression in the methods of Hawthorne, as discussed in depth in [18], where areas of colour are iteratively refined in an attempt to perform an accurate representation. Each of these iterative passes is referred to as a statement.

The initial statement is generally in the form of perhaps three or four planes in some basic *representative* colours. These colours are assessed in an objective manner using a colour spot device with fixed aperture. The planes are then iteratively refined in successive statements using smaller and smaller colour spots until the full picture emerges. There are generally no more than half a dozen statements. By this time the colour spots are very small and photo-realistic effects are possible if one is diligent and accurate enough.

This naturally leads us to the question of what, in each statement, is a representative colour. I have proposed in chapter 6 that all scenes emanate from a limited set of absolute contexts. These contexts are then convolved with a series of transforming influences (or local contexts) that produce the viewed scene. There seems to be an internal mechanism for the objective assessment of the initial context which is performed in a very short time. However, in terms of artistic endeavour this context is naturally convolved with other factors, both physical and psychosocial in nature and it is these factors that affect the artist's final rendering of the scene. Gombrich [143], for instance, in his approach to art by way of psychology is fortified by ranks of cognitive psychologists and such eminents as Karl Popper. He says that painting is mimicry (a mimesis) of perception that is modified by individual schema such as social constructs and psychological constructs. Semiologists, however, emphasise the symbols rather

than the perceptions. The image is thus 'fine tuned' to the prevailing social and psychological trends of the time. Painters thoughts are thus basically scientific in nature, it is a continual process of refinement with typically scientific elements:

- Initial statement of problem.
- A trial solution to the problem based on perceptual memory.
- An experimental situation devised to test the strengths and weaknesses of the solution.

This is a continuous development or, as Gombrich says 'a gradual modification of traditional schematic conventions under the pressure of novel demands'. That is to say taking the current schema and modifying the individual elements to deal with new findings. Success is then judged in terms of whether the performed work is wholly representative of the artist's internal vision. Thus art is a raw (or context-*less*) rendering of scene into which many modifiers are convolved.

The final aim of any artist is a rendering of a scene purely in terms of light rather than in terms of objects. To the artist, objects *per se* are a false concept since local context light effects are all that exist.

References

1. Dana H. Ballard and Christopher M. Brown, *Computer Vision*, Prentice Hall Inc., New Jersey, 1982

2. D. L. Waltz, "Understanding Line Drawings of Scenes with Shadows" in *The Psychology of Computer Vision*, P. H. Winston (ed.), McGraw-Hill, New York, 1975, pp. 19-91

3. R. Ohlander, "Analysis of Natural Scenes", Computer Science Department Report (PhD thesis), Carnegie-Mellon University, Pittsburgh, 1975

4. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York, 1973

5. G. J. Agin and T. O. Binford, "Computer Description of Curved Objects", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol PAMI-3, No. 2, March 1981, pp.197-200

6. A. Guzman, "Computer Recognition of Three-Dimensional Objects in a Scene", MIT Report MAC-TR-59, December 1968

7. A. Guzman, "Decomposition of a Visual Scene into Three - Dimensional Bodies", *AFIPS Proceedings Fal Joint Conference*, Vol. 33, 1968

8. J. M. S. Prewitt, "Object Enhancement and Extraction", in *Optical and Electro-Optical Information Processing*, J. T. Tippett, et al (ed.s), MIT Press, Cambridge Mass., pp. 159-197

9. E. Wolff, "Wolf's Anatomy of the Eye and Orbit 7th Edition", Saunders, Philadelphia, 1976

10. Dartnell, Bowmaker and Mollon, *Proceedings of the Royal Society of London*, B220, pp. 115-130

11. Marks, Dobelle and MacNichol, "Retinal Pigmentation in Rhesus Monkeys", in *Science*, 143, pp. 1181-3

12. Y. Ohta, *Knowledge Based Interpretation of Outdoor Natural Colour Scenes*, Pitman, London, 1985

13. H. Derin and C. S. Won, "A Parallel Image Segmentation Algorithm Using Relaxation with Varying Neighbourhoods and Its Mapping to Array Processors", Computer Vision, Graphics and Image Processing, 1987, No.40, pp. 54-78

14. V. S. Nalwa and E. Pauchon, "Edgel Aggregation and Edge Description", Computer Vision, Graphics and Image Processing, 1987, No. 40, pp. 79-94

15. R. Booth, *Scene Analysis and 3-D Object Recognition Using Passive Vision*, PhD Thesis, University of Newcastle upon Tyne, 1988, Thesis No. L3616

16. C. Robertson and G. M. Megson, "Scene Analysis - A Brief Survey", University of Newcastle-upon-Tyne Technical Report Series, No. 480, March 1994.

17. T.H.Cormen, C.E.Leiserson and R.L.Rivest, *Introduction to Algorithms*, MIT Press, Cambridge Massachusetts, 1990.

18. A.Stern, *How to See Color and Paint It*, Watson Cuptill Publications, New York, 1988.

19. R. Narasimhan, "Picture Languages", in *Picture Language Machines*, ed. S. Kaneff, Academic Press, 1970.

20. M.B.Clowes, "Picture Syntax", in *Picture Language Machines*, ed. S. Kaneff, Academic Press, 1970.

21. S.B.Stanton, "Plane Regions: A Study in Graphical Communication", in *Picture Language Machines*, ed. S. Kaneff, Academic Press, 1970.

22. S.K.Reed, *Psychological Processes in Pattern Recognition*, Academic Press Inc., London, 1973. Ch III.

23. F.Birren, *Color Perception in Art*, Non Nostrand Reinhold Company, New York, 1976.

24. D. Marr, *Vision*, W.H.Freeman and Co., New York, 1982.

25. D.F. Rogers, *Procedural Elements for Computer Graphics*, McGraw-Hill Book Company, Singapore, 1985.

26. N. Chomsky, *Aspects of the Theory of Syntax*, MIT Press, Cambridge Mass., 1965.

27. N. Goodman, *Languages of Art*, Bobbs-Merrill, Indianapolis, 1968.

28. M. A. Hagen, "Introduction: What, then, are pictures ?", in M. A. Hagen, *The Perception of Pictures*, Vol.1, Academic Press, New York, 1980.

29. L.G.Roberts, "Machine Perception of Three-Dimensional Solids", *Optical and Electro-Optical Information Processing*, J.T.Tippett et al.(eds), MIT Press, Cambridge Mass., 1963, pp 159-197.

30. Sobel's Operator - No published source. However, it is associated with his name.

31. L. Davis, "A Survey of Edge Detection Techniques", *Computer Graphics and Image Processing 4 1975*, pp 248-270

32. E. Hildreth, "Edge Detection", MIT Artificial Intelligence Laboratory, AI Memo 858, 1985

33. D.H.Hubel and T.N.Weisel, "Functional Architecture of Macaque Monkey Visual Cortex", *Journal of Physiology* 195 (1968), pp 215-242

34. R.Kurzweil, *The Age of Intelligent Machines*, pp233-281, MIT Press, 1990.

35. M.H.Heuckel,"An Operator Which Locates Edges in Digitized Pictures", *Journal of ACM*, Vol.18, 1971, pp 113-115.

36. M. H. Heuckel,"A Local Visual Edge Detector Which Recognises Edges and Lines", *Journal of ACM*, Vol.20, 1973, pp 634-647.

37. M. James, *Pattern Recognition*, BSP Proffesional Books, London, 1987.

38. D.Marr and E.Hildreth, "Theory of Edge Detection", *Proceedings of the Royal Society of London, Series B, 200*, pp 269-294.

39. J. von Stauffenberg, "Klinische und anatomische Beitraege zur Kenntnis der aphasischen, agnostischen und apraktischen Symptome", Z. Neurol. Psychiat., 39:71-213, 1918.

40. P. Schilder, *Medical Psychology*, J. Wiley and Sons, New York, 1965.

41. J.Z.Young, *Philosophy and the Brain*, Oxford University Press, Oxford, 1988.

42. I. Newton, *Opticks, Or a Treatise of the Reflections, Refractions, Inflections and Colours of Light*, based on the fourth edition of 1730, Dover Publications, New York, 1952.

43. J. Locke, *An Essay Concerning Human Understanding*, ed. P. H. Nidditch, Oxford University Press, Oxford, 1975 (1690).

44. M. D'Zmura and P. Lennie, "Mechanisms of Color Constancy", *Journal of the Optical Society of America*, A3:1662-72, 1986.

45. H. Putnam, "Reductionism and the Nature of Psychology", *Mind Design*, MIT Press, Cambridge Mass., 1973.

46. G. Fauconnier, "Quantification, Roles and Domains", in *Meaning and Mental Representations*, ed. U. Eco, M. Santambrogio and P. Violi, Indiana University Press, 1988.

47. R. Jackendoff, "Conceptual Semantics", in *Meaning and Mental Representations*, ed. U. Eco, M. Santambrogio and P. Violi, Indiana University Press, 1988.

48. G. Lakoff, "Cognitive Semantics", in *Meaning and Mental Representations*, ed. U. Eco, M. Santambrogio and P. Violi, Indiana University Press, 1988.

49. S. D'O. Cotton, "Colour, Colour Spaces and the Human Visual System", Technical Report, Computer Science Department, University of Birmingham, 1995.

50. M. F. Fielding, *Fractals Everywhere*, Academic Press Inc., San Diego, 1988.

51. G. Korvin, *Fractal Models in the Earth Sciences*, Elsevier London, 1992.

52. D. L. Turcotte, *Fractals and Chaos in Geology and Geophysics*, Cambridge University Press, 1992.

53. B. B. Mandelbrot, *The Geometry of Nature*, Freeman, San Francisco, 1982.

54. H. Freeman, "Computer Processing of Line Drawing Images", in *Computer Surveys 6*, March 1974, pp 57-98.

55. G. Gallus and P. W. Neurath, "Improved Computer Chromosome Analysis Incorporating Preprocessing and Boundary Analysis", in *Physics in Medicine and Biology 15*, 1970, 435.

56. J. Ramesh, R. Kasturi and B.G.Schunck. "Machine Vision", McGraw-Hill, New York, 1995.

57. W. A. Perkins, "Representing Sentence Information", Paper 1468-85, SPIE Symposium, Orlando, FL, 1991.

58. R. H. McEachern, *Human and Machine Intelligence*, R and E Publishers, Saratoga, CA., 1993.

59. U. Eco, "On Truth - A Fiction", in *Meaning and Mental Representations*, ed. U. Eco, M. Santambrogio and P. Violi, Indiana University Press, 1988.

60. M. Podro, "Depiction and the Golden Calf", in *Visual Theory*, ed. N. Bryson, M. A. Holly and K. Moxey, Polity Press, Cambridge, United Kingdom, 1992.

61. W. H. Calvin, *The Cerebral Code - Thinking a Thought in the Mosaics of the Mind*, MIT Press, Cambridge Mass. and London England, 1996.

62. P. Fletcher and C. W. Patty, *Foundations of Higher Mathematics*, P.W.S Publishing Company, Boston, MA, 1996.

63. N.Stanoulov, "Preliminary Notes on a Functional Scheme of Human Thought", *Progress in Brain Research*, Vol. 2, ed. N.Weiner and J.P.Schade, Elsevier Publishing, London, 1963.

64. S.Goldman, *Information Theory*, Constable and Co., London, 1958.

65. J. Zeman, "Information in the Brain", *Progress in Brain Research*, Vol. 2, ed. N. Wiener and J.P.Schade, Elsevier Publishing, London, 1963.

66. M. Burton, "Good morning, Mr . . . er", *New Scientist*, 01 February 1992, Vol.133 No.1806 Page 39.

67. J. Davidoff and D. Concar, "Brain cells made for seeing: How do we visualise the world?", *New Scientist*, 10 April 1993, Vol.138 No.1868 Page 32.

68. M. Livingstone and D. Hubel, "Segregation of form, color, movement and depth: Anatomy, physiology and perception", *Science*, 240:740-749, 1988.

69. R. Shapley, E. Kaplan and R. Soodak, *Nature (London)*, **357**, 219 (1984).

70. M. Minkowski, Archives of Neurological Psychiatry, 6, 201, 1920.

71. O. D. Creutzfeldt, "The Neurophysiological correlates of colour induction, colour and brightness contrast", *Progess in Brain Research*, Vol. 95, "The Visually Responsive Neuron: From Basic Neurophysiology to Behaviour", Elsevier, London, 1993.

72. S. Zeki, "The Representation of Colours in the Cerebral Cortex", *Nature, 284*, 412-418, 1980.

73. S. Zeki , "Colour Coding in the Cerebral Cortex : The Reaction of Cells in the Monkey Visual Cortex to Wavelengths and Colours", *Neuroscience, 9*, 741-765, 1983.

74. D. I. Perrett, E. T. Rolls and W. Caan, "Visual Neurones Responsive to Faces in the Monkey Temporal Cortex", *Experimental Brain Research, 47*, 329-342.

75. K. Kendrick, "Through a sheep's eye: Sheep use their visual sense to recognise food, friends and foes.", New Scientist, 12 May 1990, Vol.126 No.1716.

76. W. Rall, "Functional Aspects of Neuronal Geometry", in *Neurons without Impulses*, Society for Experimental Biology Seminar Series 6, Cambridge University Press, 1981.

77. J.D Foley, A van Dam et al., *Computer Graphics - Principles and Practice*, Addison Wesley, 1996

78. S. Haykin, *Neural Networks - A Comprehensive Foundation*, Macmillan, 1994.

79. J.L. McClelland, D.E. Rumelhart and the PDP Reseach Group, *Parallel Distributed Processing Vol. I & II*, MIT Press, Eighth Printing 1988.

80. C. Robertson and G. M. Megson, "Neural Network Analysis of Hue Spectra from Natural Images", in *Proceedings of the 3rd International Conference on Artificial Neural Networks and Genetic Algorithms 1997*, University of East Anglia, Springer-Verlag, 1997.

81. A. Church, "A Note on the Entscheidungsproblem", *Journal of Symbolic Logic 1*, 1936.

82. J. J. Gibson, *The Ecological Approach to Visual Perception*, Houghton Mifflin, Boston, 1979.

83. E. Thompson, *Colour Vision : A Study of Cognitive Science and the Philosophy of Perception*, Routledge, London and New York, 1995.

84. A. M. Turing, "Computing Machinery and Intelligence", *Mind*, October 1950, **59**:433-460.

85. D. Marr and T. Poggio, "A Computational Theory of Human Stereo Vision", Science, *194*, 283-287.

86. A. M. Waxman, M. Siebent, R. Cunninghman, J Wu, "Neural Analog Diffusion Enhancement Layer and Early Visual Processing", SPIE Cambridge Symposium on Optical and Optoelectronic Engineering, paper 1001-137, November 1988.

87. H. J. Caulfield, "ART Application to Coherent Optical Pattern Recognition", SPIE Cambridge Symposium on Optical and Optoelectronic Engineering, paper 1001-106, November 1988.

88. G.W.Cottrell, "Principal Component Analysis of Images via Back Propagation.", SPIE Cambridge Symposium on Optical and Optoelectronic Engineering, paper 1001-105, November 1988.

89. M. R. Sayeh and J. Y. Han, "Pattern Recognition Using a Neural Network", SPIE Cambridge Conference on Intelligent Robots and Computer Vision, paper 848-45, November 1987.

90. D. H. Ballard and S. R. Sload, "Experience with the Generalised Hough Transform", Proceedings of the 5th International Conference on Pattern Recognition, pp74-179, 1980.

91. P. V. C. Hough, "Method and Means for Recognising Complex Patterns", U.S. Patent 3,069,654:1962

92. S. R. Deans, "Hough Transform from the Radon Transform", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 3 No. 3, March 1981.

93. D. H. Ballard, "Generalising the Hough Transform to Detect Arbitrary Shapes", Pattern Recognition, pp111-222, Volume 13, 1981.

94. I. Rock, *The Logic of Perception*, MIT Press, Cambridge, 1983.

95. A. Rosenfeld and A. C. Kak, *Digital Picture Processing*, Academic Press, New York, second edition, 1982. Two volumes.

96. R. M. Haralick and L. G. Shapiro, " Image Segmentation Techniques", *Computer Vision, Graphics and Image Processing, 29*(1):100-132, 1985.

97. Y. Yakimovsky, "Boundary and Object Detection in Real World Images.", *Journal o fthe Association of Computing Machinery*, 23:599-618, 1976.

98. P. J. Besl and R. C. Jain, "Segmentation through Variable-Order Surface Fitting, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239-256, February 1992.

99. G. Healey, "Using Colour for Geometry Insensitive Segmentation", *Journal of the Optical Society of America*, 6(6):920-937, 1989.

100. R. Ohlander, K. Price and D. R. Reddy, "Picture Segmentation Using a Recursive Region Splitting Method", *Computer Graphics and Image Processing*, 8(3):313-355, 1978.

101. A. M. Nazif and M. D. Levine, "Low Level Image Segmentation: An Expert System", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(5):555-577, 1984.

102. J. R. Beveridge et al., "Segmenting Images Using Localized Histograms and Region Merging", *International Journal of Computer Vision*, 2(3):311-347, 1989.

103. T. Matsuyama, "Expert Systems fro Image Processing: Knowledge-Based Composition of Image Analysis Processes", *Computer Vision, Graphics and Image Processing*, 48:22-49, 1989.

104. V. S. S. Hwang, L. S. Davis and T. Matsuyama,"Hypothesis Integration in Image Understanding Systems", *Computer Vision, Graphics and Image Processing*, 36:321-371,1986.

105. F. F. Sabins, *Remote Sensing: Principles and Interpretation*, W. H. Freeman, New York, second edition, 1987.

106. E. De Micheli, B. Caprile, P. Ottonello and V. Torre, "Localization and Noise in Edge Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(10):1106-1117, Octber 1989.

107. I. E. Abdou and W. K. Pratt, "Quantitative Design and Evaluation of Enhancement / Thresholding Edge Detectors", *Proceedings of the IEEE*, 67(5):753-763, May 1979.

108. A. L. Yuille and T. Poggio, "Fingerprint Theorems for Zero Crossings", *Journal of the Optical Society of America*, 2: 683-692, May 1985.

109. M. J. Quinn, *Parallel Computing: Theory and Practice*, McGraw-Hill, New York, 1994.

110. R. Lachman, J. L. Lachman and E. C. Butterfield, *Cognitive Psychology and Information Processing*, Erlbaum Associates, New Jersey, 1979.

111. H. A. Simon and C. A. Kaplan, "Foundations of Cognitive Science", in *Foundations of Cognitive Science*, ed. M. I. Posner, MIT Press, Cambridge Mass., 1989.

112. G. W. Humphreys and M. J. Riddoch, "Visual Processing in Normality and Pathology : Implications for Rehabilitation", in *Cognitive Neuropsychology and Cognitive Rehabilitation*, ed. G. W. Humphreys and M. J. Riddoch, Erlbaum Associates, 1994.

113. K. Koffka, *Principles of Gestalt Psychology*, Harcourt Brace, New York, 1935.

114. F. C. Bartlett, *Remembering: A Study in Experimental and Social Psychology*, Cambridge University Press, Cambridge, 1932.

115. E. Kant, *Critique of Pure Reason (2nd Edition)*, MacMillan, London, 1963, originally published in 1787.

116. R. C. Schank and R. P. Abelson, *Scripts, Plans, Goals and Understanding*, Lawrance Erlbaum and Associates, Hillsdale, New Jersey, 1977.

117. E. E. Smith, "Concepts and Though", in *The Psychology of Human Thought*, eds.R. J. Sternberg and E. E. Smith, Cambridge University Press, 1988.

118. A. Tversky, "Features of Similarity", in *Psychological Review, 84*, 327-352.

119. C. D. Brody, "A Model of the Feedback to the Lateral Geniculate Nucleus", *Advances in Neural Information Processing Systems*, ed S. J. Hanson, J. D. Cowan and C. L. Giles, Morgan Kaufmann Publishers Inc., San Mateo, California, 1993.

120. A. Zell, N. Mache, T. Sommer and T. Korb, "Recent Developments of the SNNS Neural Network Simulator", in *SPIE COnference on the Applications of Artificial Neural Networks, Universit.* at Stuttgart, April 1991.

121. W. E. Skaggs, B. L. McNaughton, K. M. Gothard, E. J. Markus, "An Information-Theoretic Approach to Deciphering the Hippocampal Code", in *Advances in Neural Information Processing Systems*, ed S. J. Hanson, J. D. Cowan and C. L. Giles, Morgan Kaufmann Publishers Inc., San Mateo, California, 1993.

122. N. Burgess, J. O'Keefe and M. Recce, "Using Hippocampal 'Place Cells' for Navigation, Exploiting Phase Coding", in *Advances in Neural Information Processing Systems*, ed S. J. Hanson, J. D. Cowan and C. L. Giles, Morgan Kaufmann Publishers Inc., San Mateo, California, 1993.

123. C. A. Barnes, B. L. McNaughton, S. J. Y. Mizumori, B. W. Leonard and L. H. Lin, "Comparison of Spatial and Temporal Characteristics of Neuronal Activity in Sequential Stages of Hippocampal Processing", *Progress in Brain Research*, **83**, pp 287-300, 1990.

124. G. J. Goodhill, "Topoography and Ocular Dominance with Positive Correlations", in *Advances in Neural Information Processing Systems*, ed S. J. Hanson, J. D. Cowan and C. L. Giles, Morgan Kaufmann Publishers Inc., San Mateo, California, 1993.

125. S. B. Udin and J. W. Fawcett, "Formation of Topographic Maps", in *Annual Review of Neuroscience*, **11**, pp. 243-264.

126. J. S. Bruner, J. J. Goodnow and G. A. Ausin, *A Study of Thinking*, John Wiley, New York, 1956.

127. M. Gsehwind and V. Salapura, "Space Efficient Neural Network Implementation", Proceedings of the Second ACM International Workshop on Floating Point Gate Arrays, Berkley, California, 1994.

128. A. Sloman, "On Designing a Visual System (Towards a Gibsonian Computational Model of Vision)", in *Journal of Experimental and Theoretical A. I.*, 1,4 289-337, 1989.

129. L. P. Beaudoin and A. Sloman, "A Study of Motive Processing and Attention", in *Prospects in Artificial Intelligence*, eds. A. Sloman, D. Hogg, G. Humphreys, D. Partridge, A. Ramsay, I. O. S. Press, Amsterdam, pp. 229-238.

130. R. N. Haber and M. Hershenson, *The Psychology of Visual Perception*, Holt, Rheinhart and Wilson, London, 1973.

131. R. Watt, *Visual Processing : Computational, Psychophysical and Cognitive Research*, Lawrence Erlbaum Associates, Hove and London, 1988.

132. A. R. Luria, *The Mind of a Mnemonist*, Basic Books, New York, 1968.

133. P. Gouras, "Colour Coding in the Primate Retinogeniculate System", in *Central and Peripheral Mechanisms of Colour Vision*, ed. D. Ottoson and S. Zeki, Proceeding of an International Symposium Held at Stockholm, Macmillan, 1984.

134. K. Nakayama, "The Iconic Bottleneck and the Tenuous Link between Early Visual Processing and Perception", in *Vision : Coding and Efficiency*, ed. Colin Blakemore, Cambridge University Press, Cambridge, 1990.

135. M. Mishkin and T. Appenzeller, "The Anatomy of Memory", *Scientific American*, **256**, pp.80-89.

136. S. Zeki, "Colour Pathways and Hierarchies in the Cerebral Cortex", in *Central and Peripheral Mechanisms of Colour Vision*, ed. D. Ottoson and S. Zeki, Proceeding of an International Symposium Held at Stockholm, Macmillan, 1984.

137. K. Rayner, "Eye Movements in Reading and Information Processing", in *Psychology Bulletin*, **85**(3), pp. 618-660.

138. G. Legge, D. Pelli, G. S. Rubin and M. M .Schleske, "Psychophysics of Reading - I. Normal Vision", in *Vision Research*, **25**, pp. 239-252.

139. C. E. Shannon, "A Mathematical Theory of Communication", *Bell System Technical Journal, 27*, pp. 379-423; pp. 623-656, 1948.

140. R. B. Fisher, "Extracting Second Order Topographic Features from Range Data", *Proceedings of the 1990 Machine Vision Association Conference*, pp 241-246, Oxford, 1990.

141. J. R. Parker, "Algorithms for Image Processing and Computer Vision", J. W. Wiley and Sons Inc. Computer Publishing, New York, 1997.

142. G. Sander, *XVCG, Visualisation of Compiler Graphs*, User Documentation V1.30, Universitat des Saarlandes, Saarbruken, Germany.

143. E. H. Gombrich, *Art and Illusion : A Study in the Psychology of Pictorial Representation*, Princeton University Press, 1961.

144. J. A. Fodor, *The Modularity of Mind*, MIT Press, Cambridge Mass., Tenth Reprint, 1996.

145. R. Descartes, *L'homme*, 1664, Selections translated by E. Clarke and C. O'Malley in *The Human Brain and Spinal Cord*, University of California Press, Berkley, 1968.

146. R. Descartes, Meditations Metaphysique, 1667, Selections translated by E. Clarke and C. O'Malley in *The Human Brain and Spinal Cord*, University of California Press, Berkley, 1968.

147. K. Oatley, Perceptions and Representations, Methuen and Co., Great Britain, 1978.

148. M. Minsky, "A Framework for Representing Knowledge", in *The Psychology of Computer Vision*, ed. P. H. Winston, McGraw-Hill Book Company, New York, 1975.

149. M. W. Eysenck and M. T. Keane, *Cognitive Psychology*, Ch. 6, Psychology Press, Erlbaum, UK, 1996.

150. G. Sperling, "The Information Available in Brief Visual Presentations", in Psychological Monographs, **74**, no. 498, 1-29, 1960.

151. R. N. Haber, "The Impending Demise of the Iconic: A Critique of the Concept of Iconic Storage in Visual Information Processing", in *Behavioural and Brain Sciences*, **6**, 1-11, 1983.

152. A. Cussins, "The Connectionist Construction of Concepts", in The Philosophy of Artificial Intelligence, ed M. Boden, Oxford University Press, Oxford, UK, 1990.

153. R. B. Fisher, *From Surfaces to Objects*, John Wiley and Son, UK, 1989, Ch.8.

154. P. Gouras and E. Zrenner"Color Vision: a review from a neurophysiological perspective", in *Progress in Sensory Physiology 1*, 1:139-79, 1981.

155. J. M. Von Wright, "Selection in Visual Immediate Memory", *Quarterly Journal of Experimental Psychology, 20*, pp. 62-68, 1970.

156. V. Bruce and P. R. Green, *Visual Perception: Physiology, Psychology and Ecology*, Ch. 7, Lawrence Erlbaum Associates, London, 1992.

157. M. T. Turvey, "Contrasting Orientations to the Processing of Visual Information", in *Psychological Review, 84*, pp.67-89, 1977.

158. L. Wittgenstein, *Tractatus Logico-Philosphicus*, trans. D.F.Pears and B.F.McGuiness, Routledge and Kegan Paul, 1961, first German edition 1921.

159. A. J. Ayer, *Language, Truth and Logic*, Dover Publications, New York, 1936.

160. A. J. Ayer, *Logical Positivism*, MacMillan Publishing, New York, 1959.

161. R. B. Fisher, "Is Computer Vision still AI", A.I. Magazine, Vol. 15, No.2, pp 21–27, 1994.

162. A. Rosenfeld, *Picture Languages: Formal Models for Picture Recognition*, Academic Press Inc, 1979.

163. D. H. Hubel, *Eye, Brain and Vision*, Scientific American Series, W. H. Freeman and Co., New York, 1988

164. G.S.Peake and T. N. Tan, "Script and Language Identification from Document Images", Proceedings of the British Machine Vision Conference, Vol.2, 1997.

Appendix I - Syntax and Symbols

Notation

$\underline{u_i}$	Pixel vector
A	Matrix of pixel vectors
$\underline{p_i}$	Exemplar palette vector
P	Segmentation set of palette vectors
Q	Segmentation matrix
$q_{i,j}$	Segmentation element of Q
$\underline{t_n}$	Tween number n
$l^n_{i,j}$	Line n, between partitions i and j
D^n	Plane number n
L^n	Line set n

Syntax

Descriptors

$D^n \odot D^m$	Plane n contains plane m
$D^n \ominus D^m$	Plane n borders plane m and vice versa

Functions

$\| \underline{p}_n \|$	Brightness of palette colour n
$\| l^n \|$	Length of line l_n
$\| L^n \|$	Number of line segments in line set n
$\underline{p}_n + \underline{p}_m$	Concatenation of palette colour m and colour n
$D^n + D^m$	Concatenation of plane n and plane m
$L^n + L^m$	Concatenation of line set n and line set m
$P(D^n)$	Perimeter length of plane n

Special Symbols

W	Line set of the viewport edges
$\underline{t}_0$	empty tween
l^0	empty line segment
D^0	empty plane
$\underline{p}_0$	empty colour

Appendix II - Sampling Equipment and Software

Equipment

The commercially available digital camera used was constructed around a 0.5cm CCD with 250,000 pixels arranged in a 320 x 240 matrix with 3 samples per pixel. The recording system was 24-bit and included hardware JPEG encoding. Exposure was based on a TTL centre-point photographic element for light metering giving an exposure range of +5EV to +18EV with a fuzzy logic element which could adjust by −2EV and +2EV. Manual exposure was also possible but was not used in this series of tests. Shutter speed was equivalent to $\frac{1}{4000}$ sec. Images were uploaded to the computing platform via a 19200 baud serial link. Images once sampled were colour reduced from 24-bit colour to 8-bit colour using the median cut algorithm where the (R,G,B) colour cube is iteratively bisected, as described in [5]. The camera could be used in both self-contained (portable and self powered) or tethered mode.

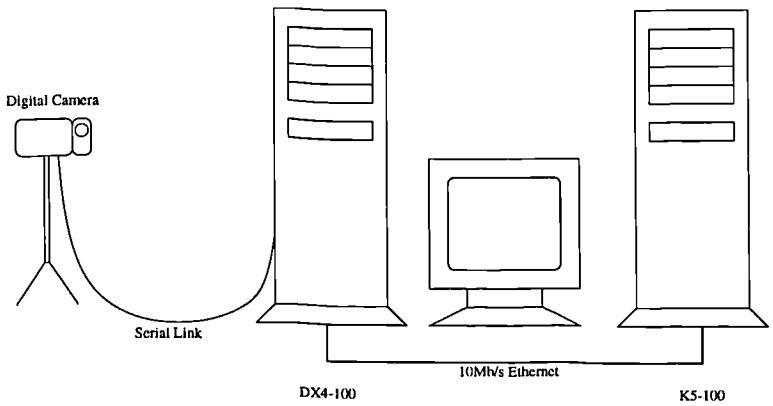


Figure 8.2: Equipment

The computing platforms consisted of an AMD 486-DX4 running at 100MHz with 16Mb of main system memory and an Pentium class AMD-K5 running at 75MHz with 32Mb main system memory and 512Kb cache. Both machines shared the same file space (8.6 Gb) via a 10Mb/s ethernet NFS. All software prototyping was done using the Linux freeware operating system and the freeware Gnu programming environment (g++, gcc, gdb, xxgdb etc.).

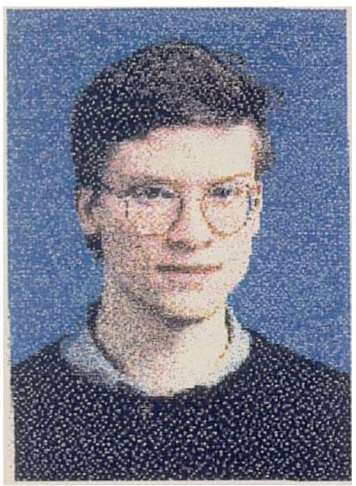
Coding Methods

All software used was written in C++ employing an object-oriented approach. The objects, or specialised data holders, were designed to be both flexible and hierarchical and expressed the underlying set-theoretic structure of the approach. On occasion, network training was cross-checked using the SNNS neural network simulator from the University of Stuttgart [120] which provided excellent flexible output and was particularly useful for benchmarking.

Appendix III - Example Processed Images

Processed Images from ICEnet

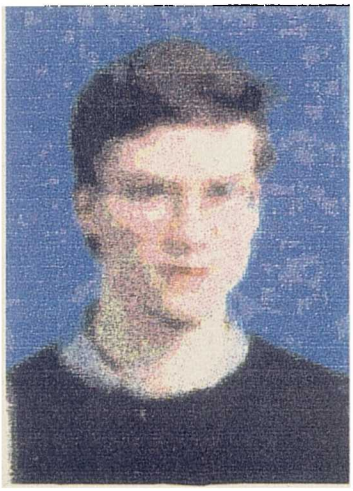
Segmented Image BGB



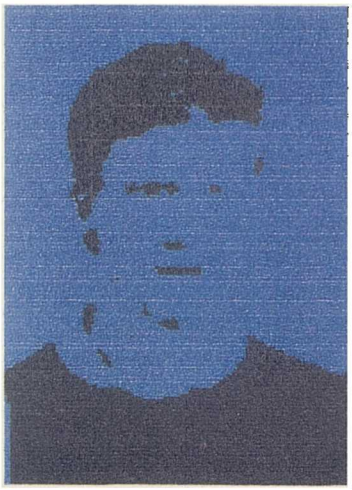
Original Image

Smoothing level 5

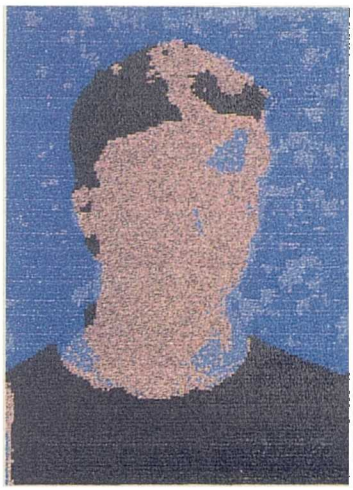
Segmented images which have been smoothed at level 5 then segmented with varying numbers of exemplars, as discussed in the text.



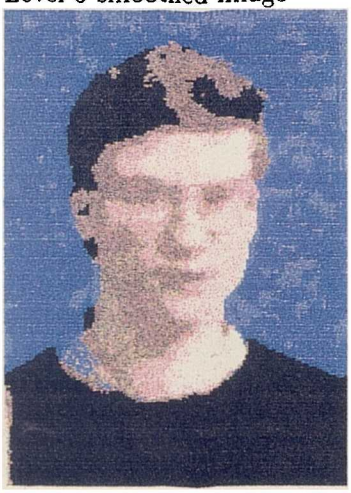
Level 5 smoothed image



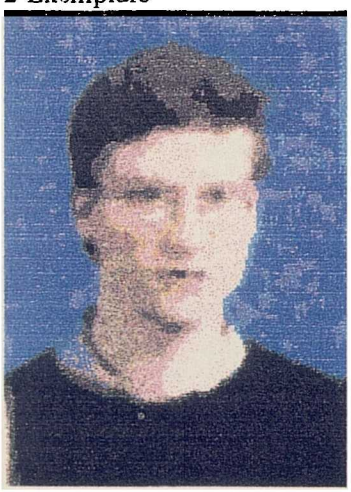
2 Exemplars



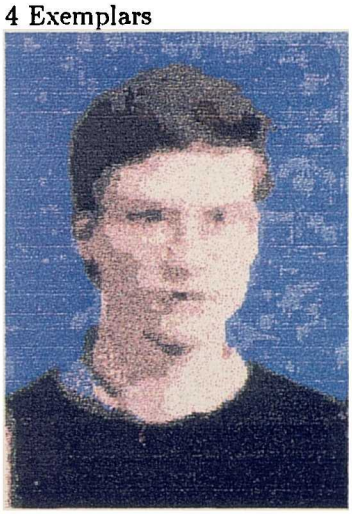
4 Exemplars



8 Exemplars



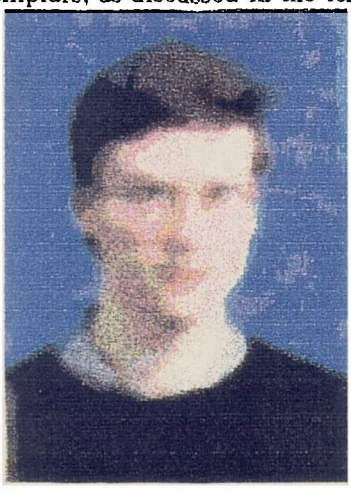
16 Exemplars



50% Exemplar coverage

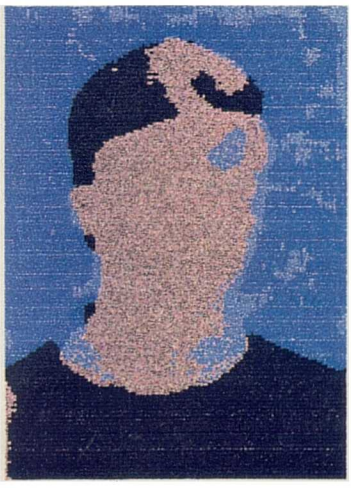
Smoothing level 7

Segmented images which have been smoothed at level 7 then segmented with varying numbers of exemplars, as discussed in the text.

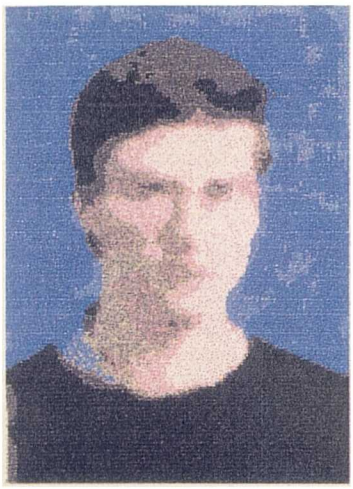


Level 7 smoothed image

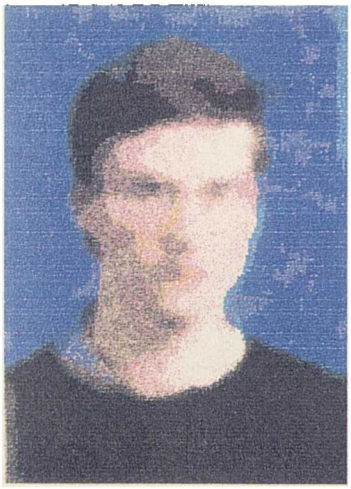
2 Exemplars



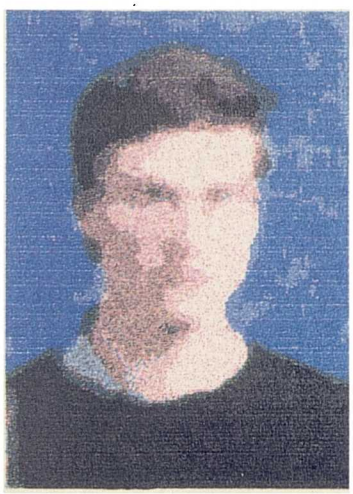
4 Exemplars



8 Exemplars



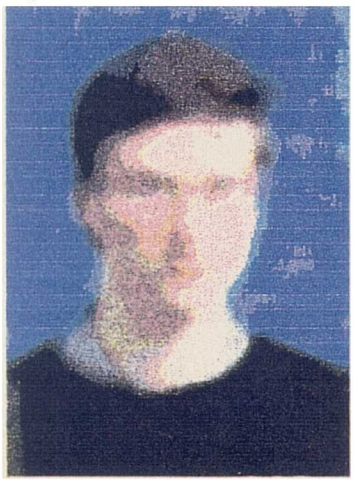
16 Exemplars



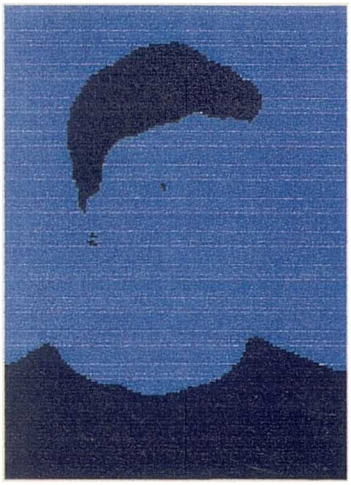
50% Exemplar coverage

Smoothing level 9

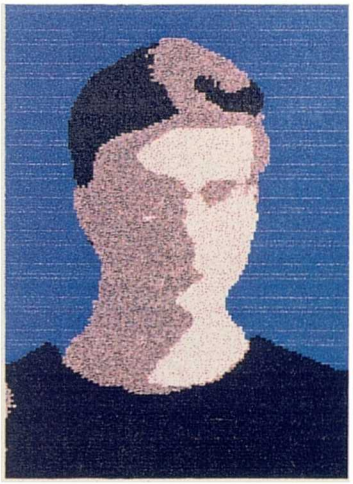
Segmented images which have been smoothed at level 9 then segmented with varying numbers of exemplars, as discussed in the text.



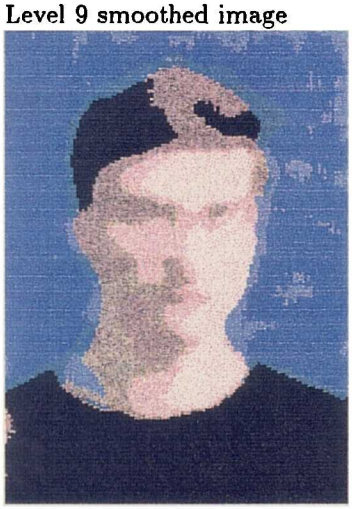
Level 9 smoothed image



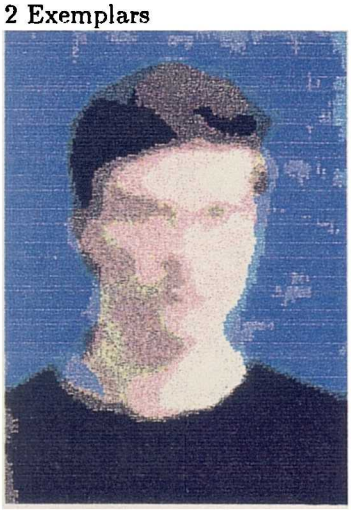
2 Exemplars



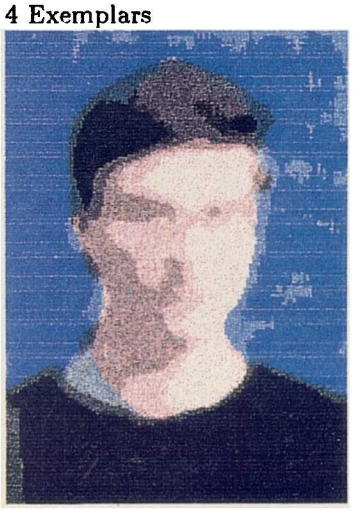
4 Exemplars



8 Exemplars



16 Exemplars

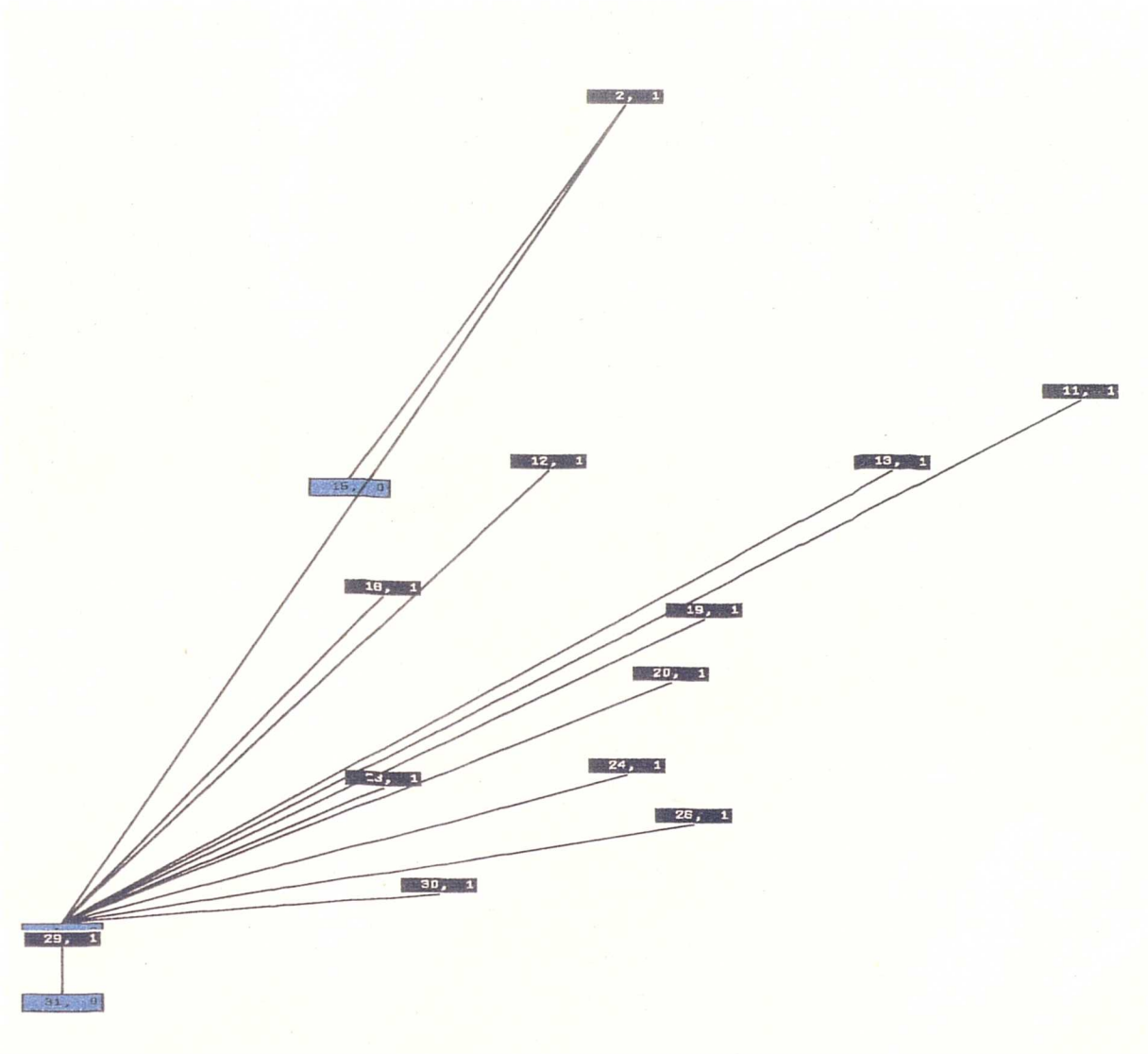


50% Exemplar coverage

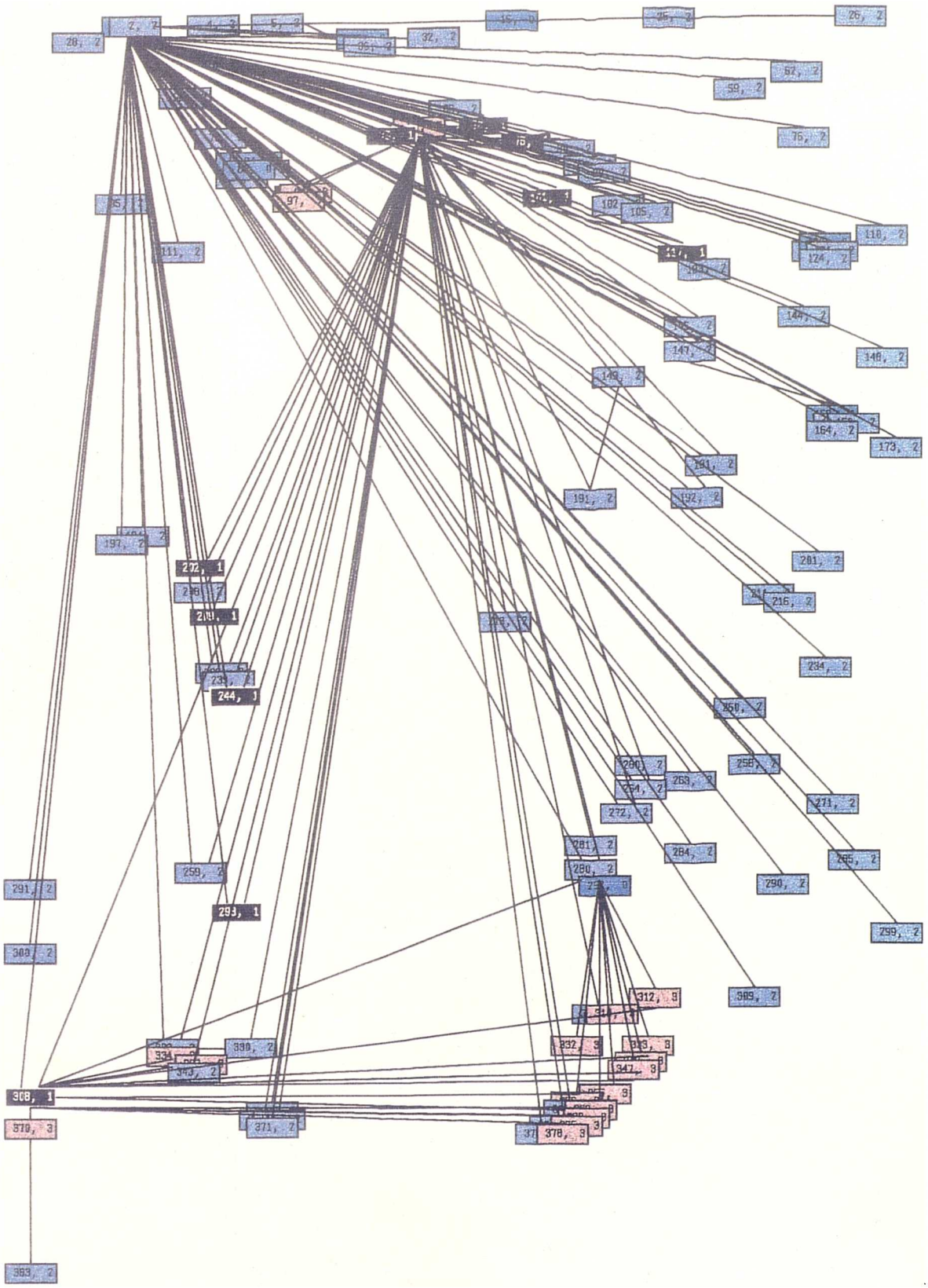
Processed Images from REDnet

Smoothing level 5

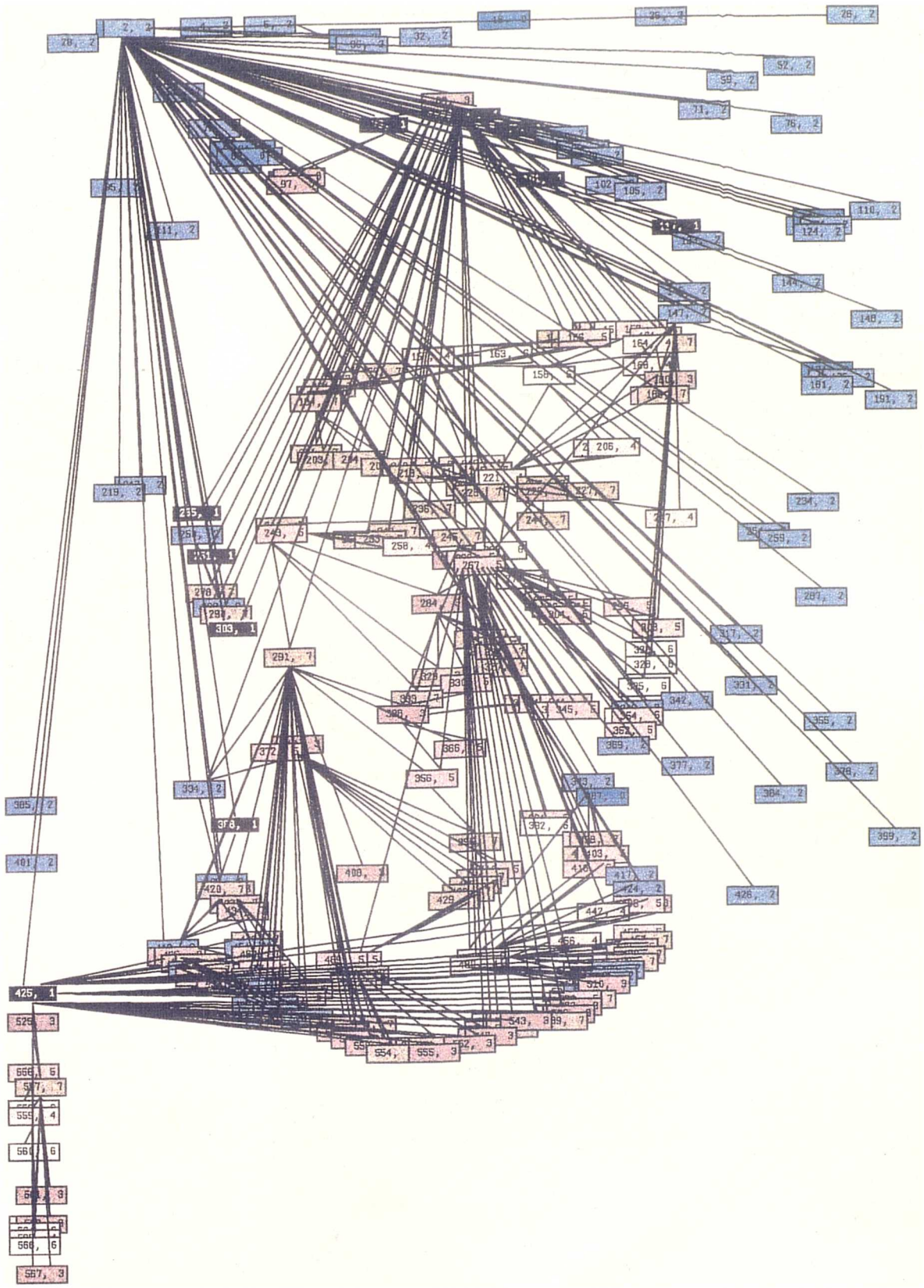
Segmented images which have been smoothed at level 5 then segmented with varying numbers of exemplars, as discussed in the text.



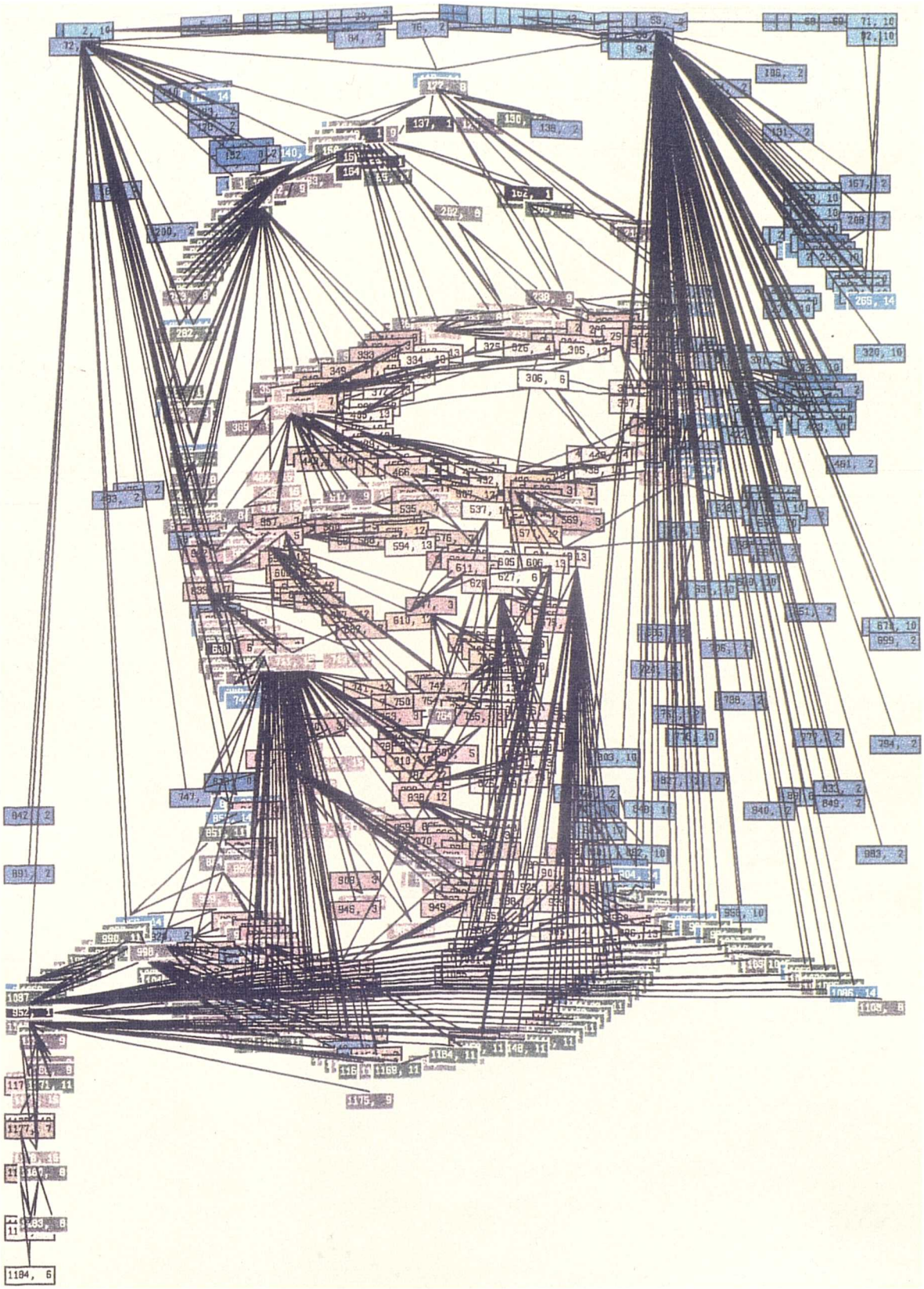
2 Exemplars


26, 2
59, 2
67, 2
76, 2
110, 2
124, 2
140, 2
148, 2
164, 2
173, 2
201, 2
216, 2
234, 2
350, 2
258, 2
271, 2
285, 2
290, 2
299, 2
309, 2
312, 9
330, 2
371, 2
378, 3
383, 2
272, 1
293, 1
308, 1

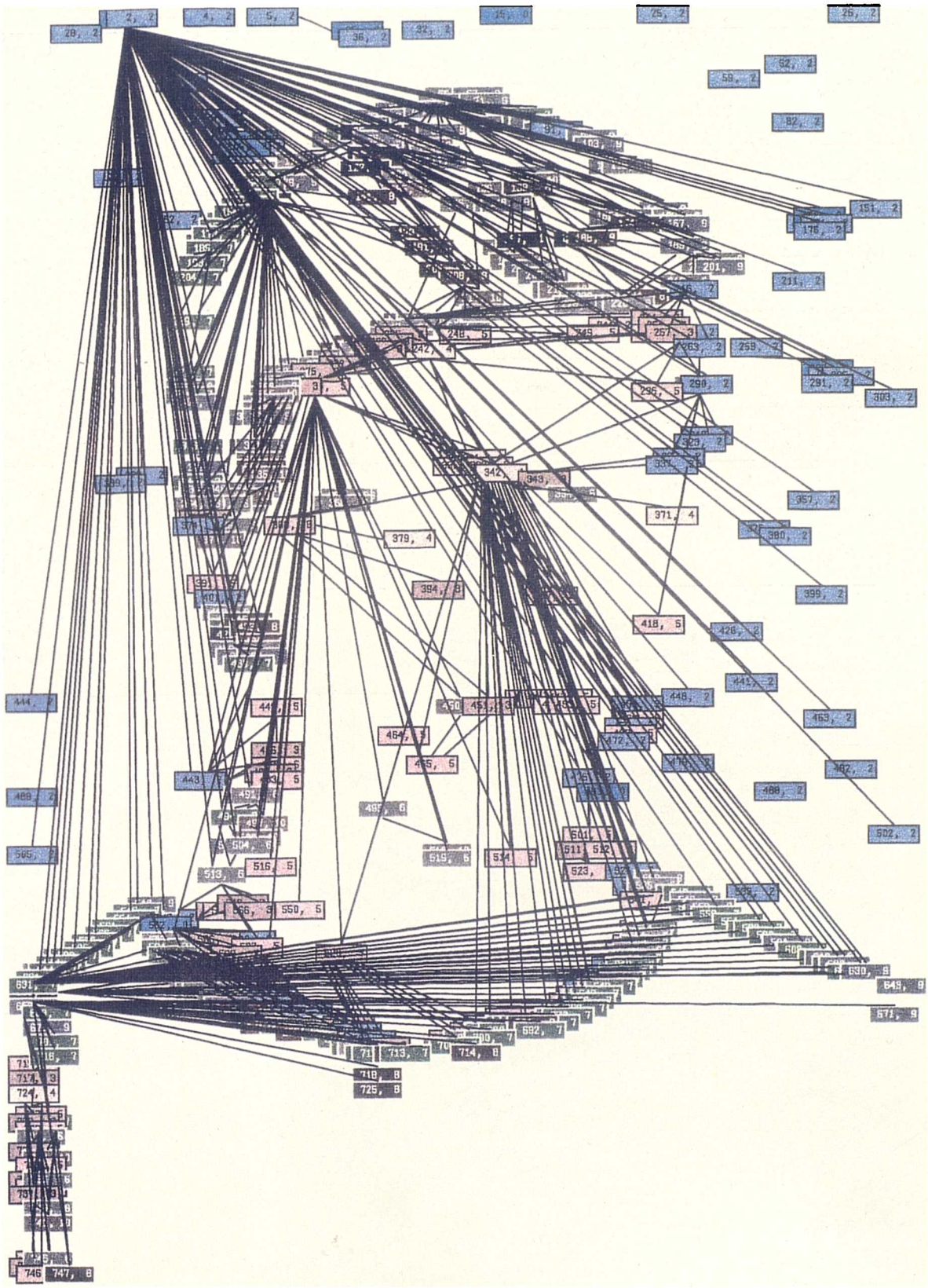
4 Exemplars



8 Exemplars



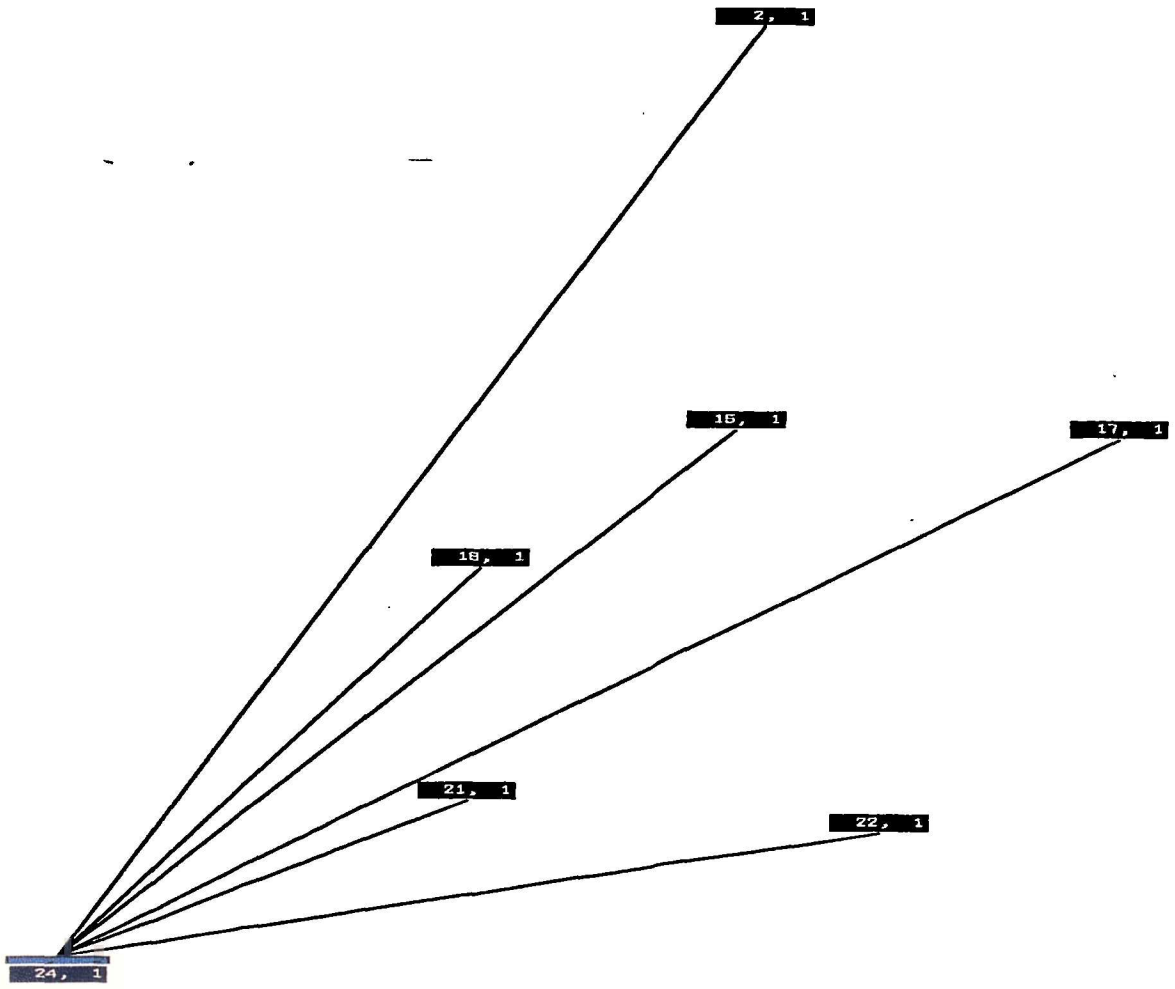
16 Exemplars



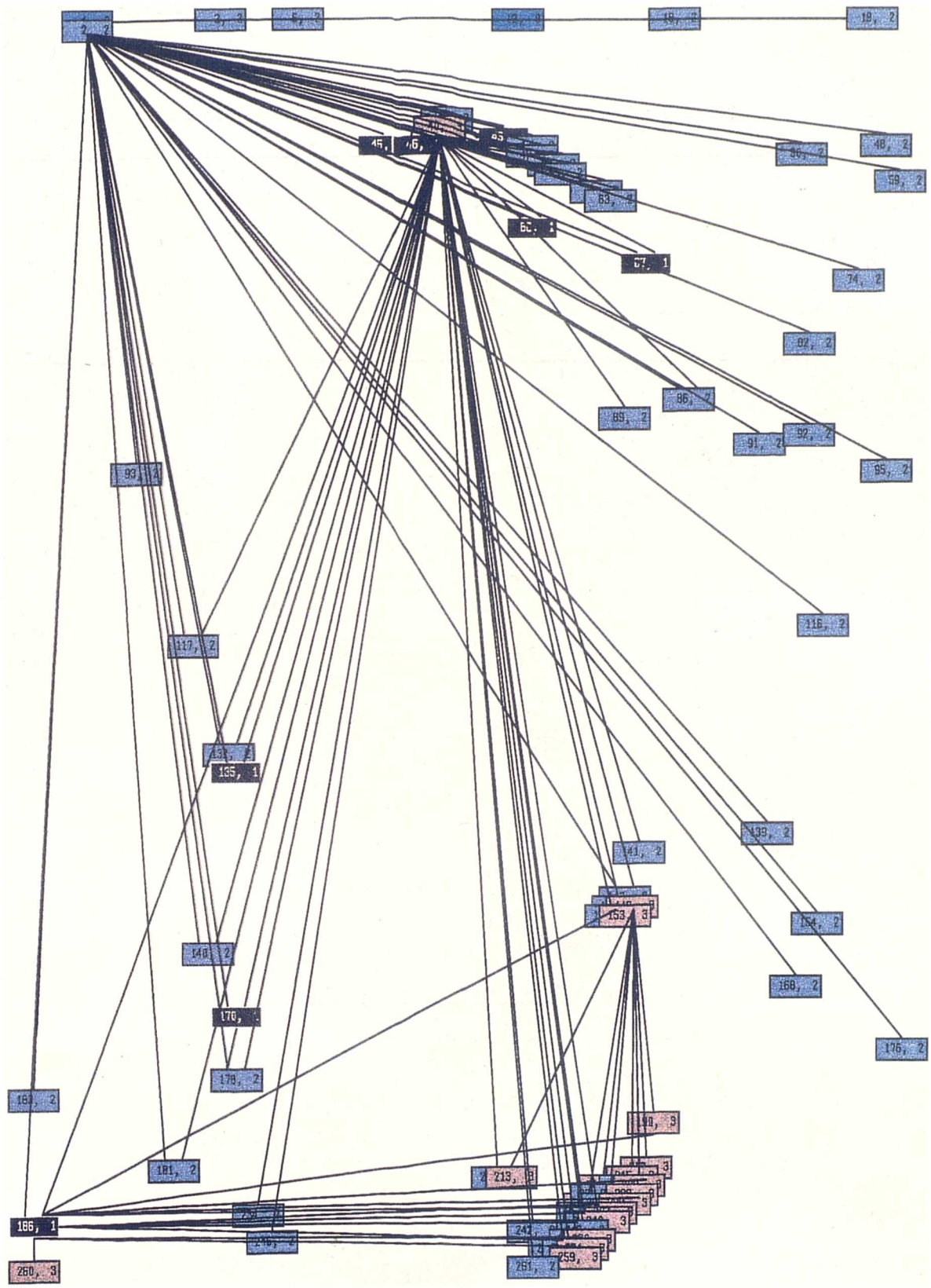
50% Exemplar coverage

Smoothing level 7

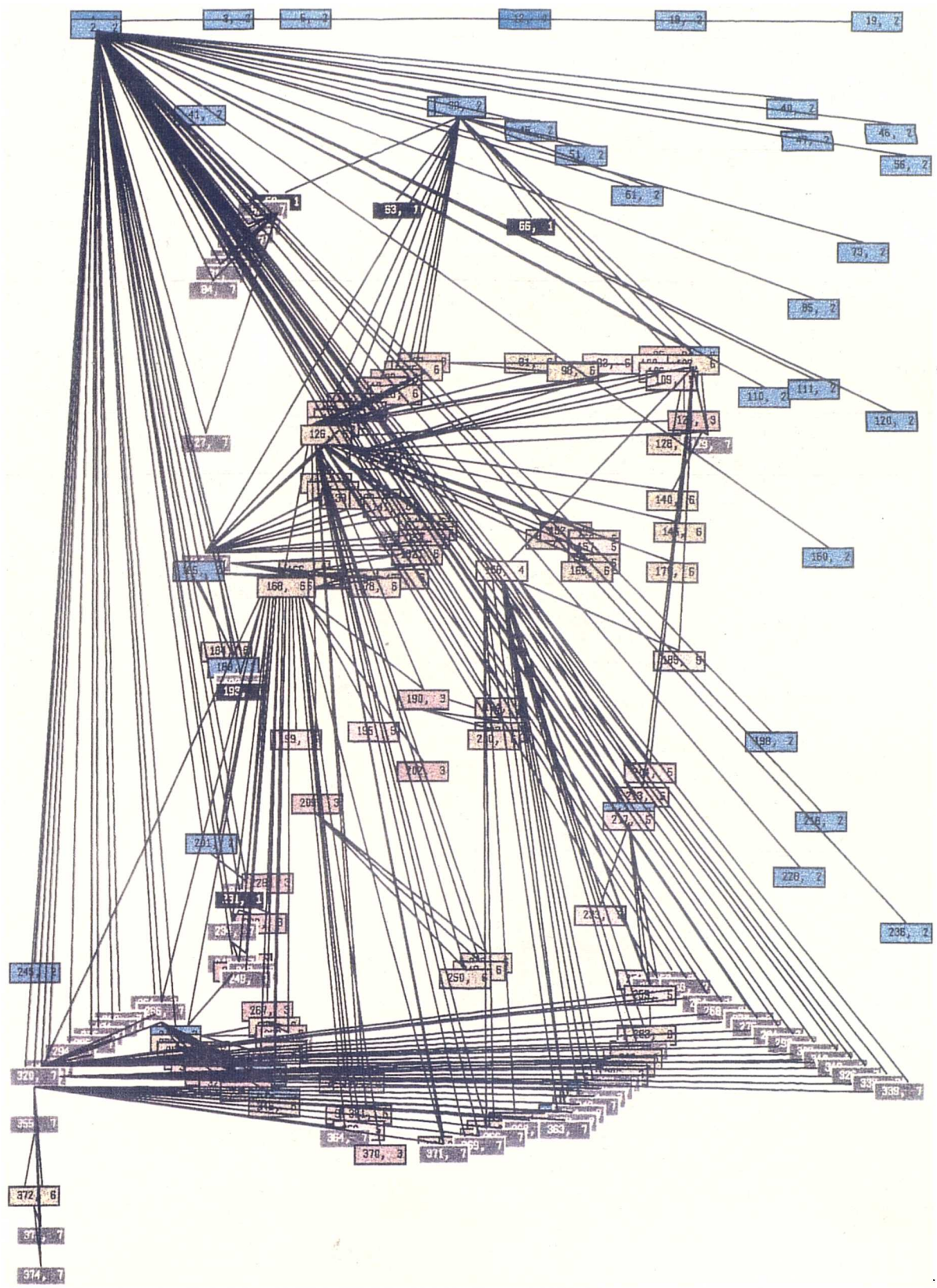
Segmented images which have been smoothed at level 7 then segmented with varying numbers of exemplars, as discussed in the text.



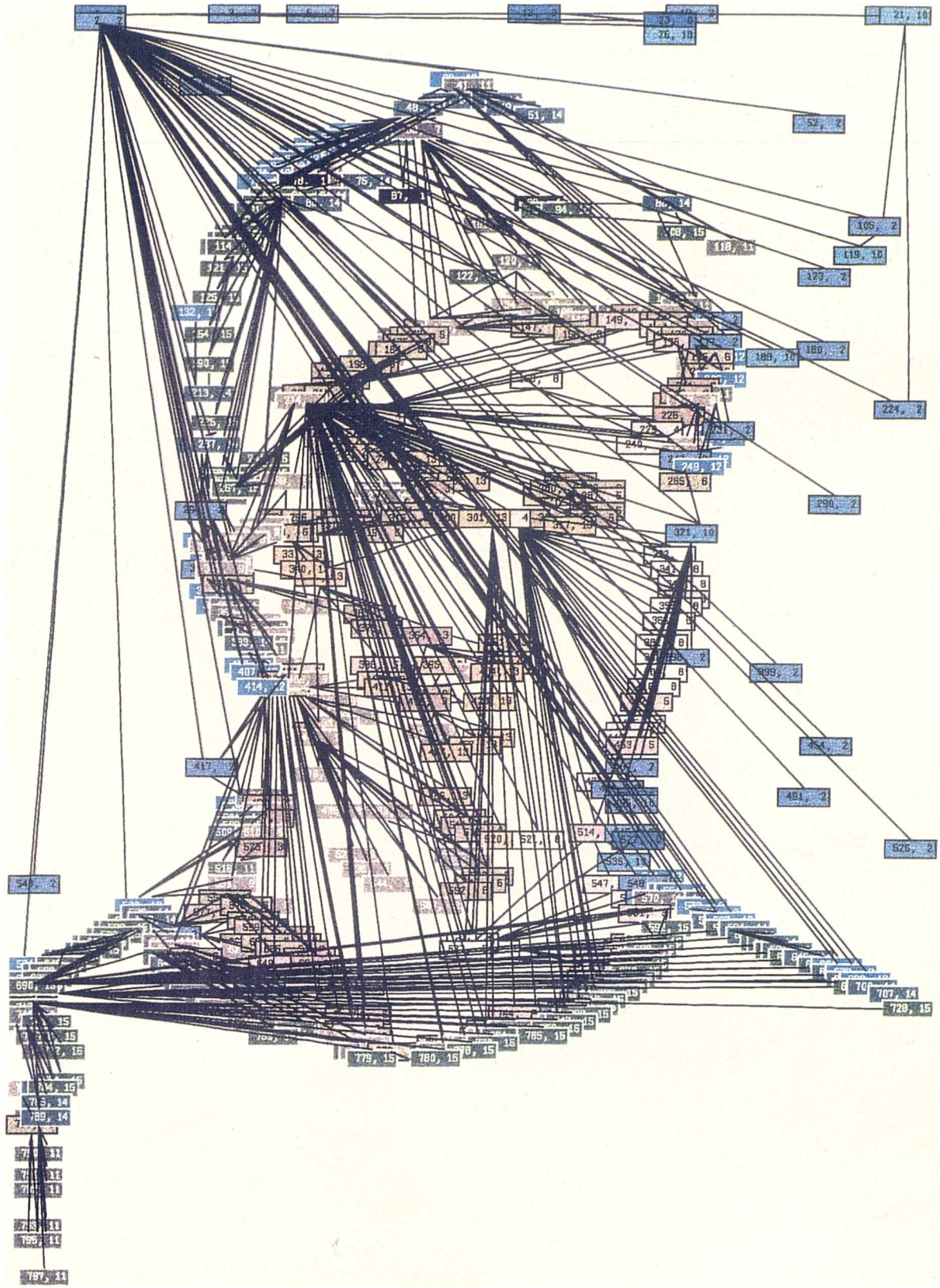
2 Exemplars


19, 2
48, 2
59, 2
74, 2
92, 2
86, 2
85, 2
91, 2
95, 2
116, 2
117, 2
135, 1
139, 2
141, 2
154, 2
168, 2
176, 2
178, 2
181, 2
186, 1
260, 3

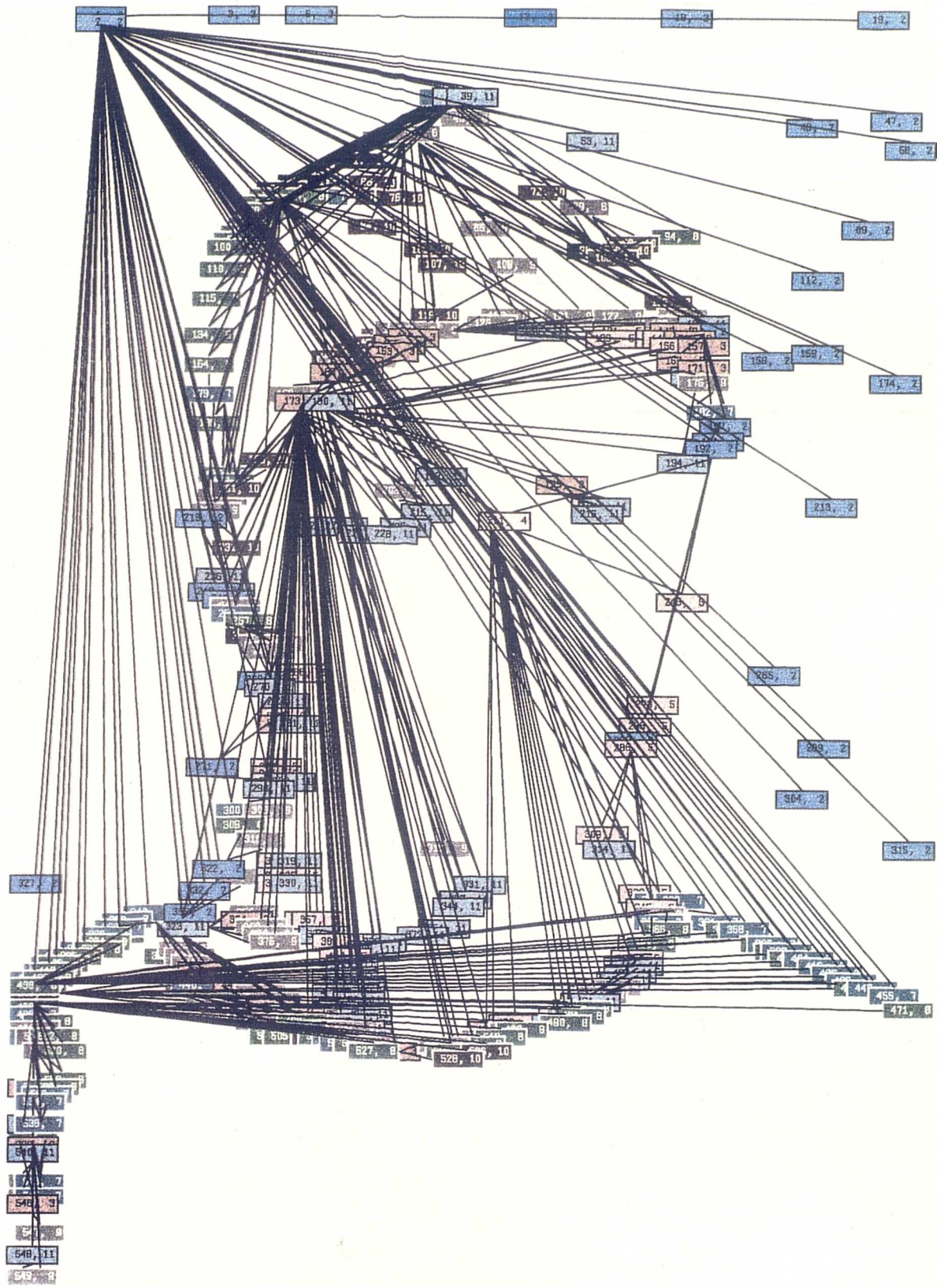
4 Exemplars



8 Exemplars

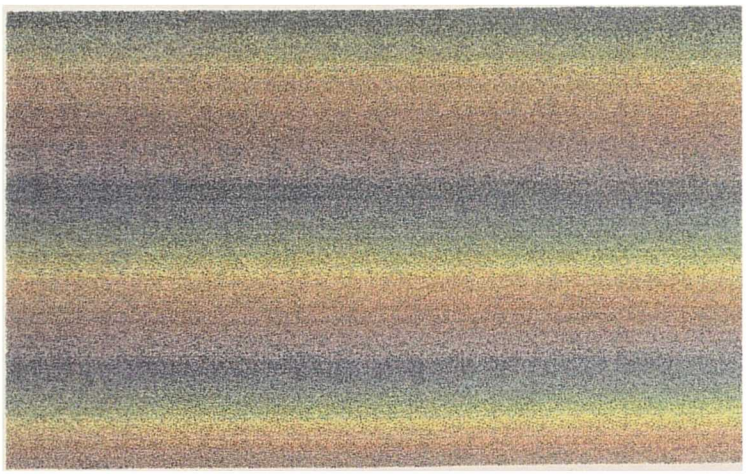


16 Exemplars

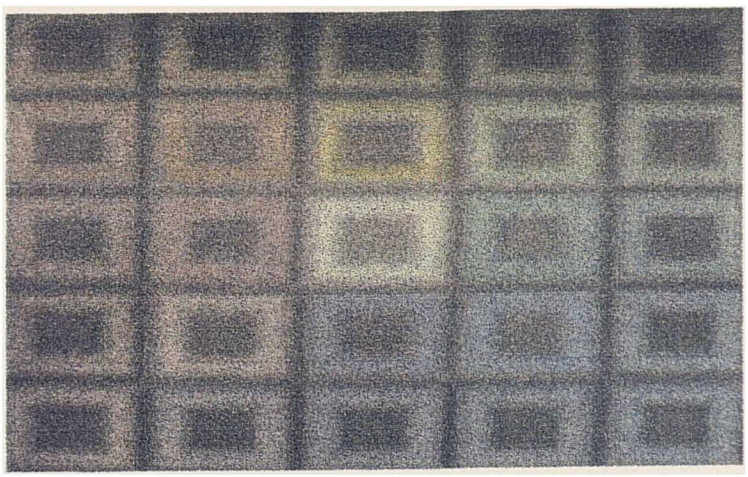


50% Exemplar coverage

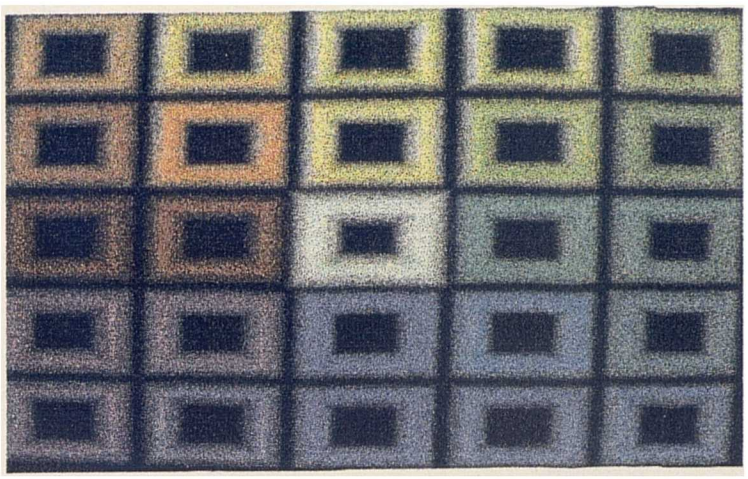
Appendix IV - Colour Contexts



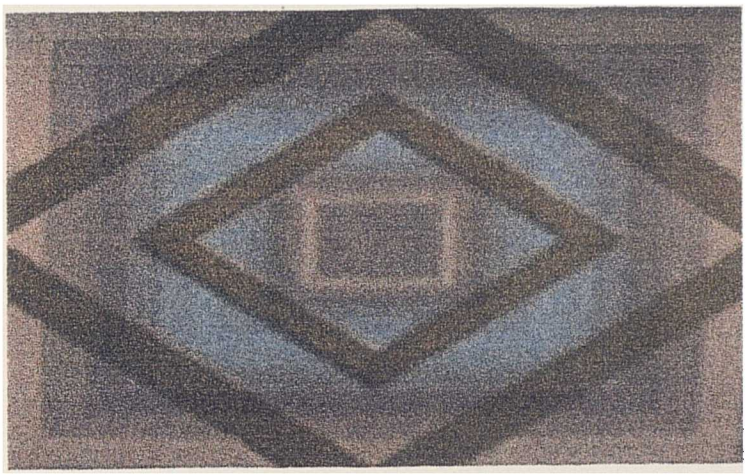
Luminosity



Iridescence



Lustre



Transparency

Chromatic Illumination

Texture