ASPECTS OF MULTI-RESOLUTIONAL FOVEAL IMAGES FOR ROBOT VISION

Fong Aik Meng

A Thesis submitted for the
Degree of Doctor of Philosophy of the University of London
and the
Diploma of Membership of Imperial College

October 1989

Department of Electrical Engineering
Imperial College of Science, Technology and Medicine
University of London

ABSTRACT

This thesis is concerned with the investigation of some issues and problems of *real-time or near real-time* robot vision and studies if an *appropriate image representation* can serve as one approach to solve some of these problems.

The approach adopted here is based on a *multiresolutional foveal (MF) image* derived from some pertinent characteristics of the human retina image-acquisition system at the fovea and at the para/perifovea known as *Multiresolutional Foveal Image Representation or Multiresolutional Foveal Images.*

The computational formulation and mathematical analysis of the MF image representation is pursued in further detail.

A suggestion is put forward using Linear Time-Invariant analysis that such an MF image provides an explanation for the source of the *independent narrowly-tuned spatial channels in the Human Vision System*, widely observed in experimental psychovisual studies.

The techniques of the Multiresolutional Foveal Image representation are applied to three widely-researched issues of robot/computer vision, namely *Semantic Object Recognition, Occluded Object Recognition* and *Hierarchical Image Transmission.*

Computation and experimental results in this investigation show that the methods of MF Images do provide viable alternative solutions to some of the problems of robot vision.

Pages

ABSTRACT 2
CONTENTS 3
LIST OF IMAGES. 7
LIST OF ABBREVIATIONS. 8
STATEMENT OF ORIGINALITY. 9
ACKNOWLEDGEMENTS. 11

CHAPTER 1
IMAGE REPRESENTATIONS AND ROBOT VISION
1.1 The Nature of the Problem and Some Definitions. 12
1.2 Real-Time Robot Vision: Difficulties and Problems. 15
1.2.2 Problems in the Hardware. 16
1.2.3 Problems in the Computational Approach. 16
1.2.4 Limitations of Image Representations. 17
1.3 Objectives and Outline of the Thesis. 18
1.3.1 Primary Objectives of the Thesis. 18
1.3.2 Related Work. 19
1.3.3 The Multiresolutional Fovea Image. 19
1.4 Organisation of the Chapters. 19

CHAPTER 2
ROBOT VISION: REVIEW OF APPROACHES AND TECHNIQUES
2.2 Some Approaches to Robot Vision. 24
2.2.1 Recovering Intrinsic Scene Characteristics from Images. 25
2.2.2 Computational Theory of Marr. 27
2.2.3 Cellullar Array and Multiresolutional Approaches. 29
2.2.4 The Logical-Node Neural-like approaches of Aleksander. 30
2.2.5 Multiresolutional Approaches. 31
2.2.6 Syntactic and Structural Approach of Fu. 32
2.3 Model-based Approaches 33
2.3.1 The Necessity of the Model-based Approach in Industrial Robot Vision. 34
2.4 Current trends in Robot Vision. 34
2.5 Summary. 35

CHAPTER 3
COMPUTER REPRESENTATION OF IMAGES FOR OBJECT RECOGNITION
--HIERARCHICAL AND NON-HIERARCHICAL--
3.1 Relationship between Image Representation and Computer Data Structures. 37
3.2 Hierarchial and Non-Hierarchical Image Representations. 37
3.3 Non-Hierarchical Representations. 38
3.3.1 Non-Hierarchical Representations of Segmented Images. 38
3.3.2 Non-Hierarchical Representations of Unsegmented Images. 41
3.4 Hierarchical Representation of Images. 42
3.4.1 The Canonical Image Pyramid. 43
3.4.1.1 The Basic M-Pyramid. 44
3.4.1.2 The Basic T-Pyramid. 44
3.4.2 Computation in Pyramids. 45
3.4.3 Pyramidal Image Applications. 45
3.4.4 Pyramidal Images as Efficient Image Codes. 46
3.5 Other Hierarchical Image Representations. 47
3.6 Summary. 47

CHAPTER 4
THE MULTIRESOLUTIONAL FOVEAL IMAGE REPRESENTATATION.
4.1 The Human Visual System. 48
4.2 Relationship Between Multiresolutional Foveal Images and the HVS. 50
4.2.1 Physiological and Neural Parallels. 50
4.2.2 Psychological Studies of Human Object Perception. 53
4.3 Spatial Frequency Theory of HVS. 54
4.4 Structure of the Multiresolutional Foveal Images. 55
4.4.1 The Global Multiresolutional Foveal (GMF) Image. 55
4.4.1.1 Choice of GMF Image Dimensions. 58
4.4.1.2 Shape of GMF Images. 59
4.4.1.3 Resolution Reduction in GMF Images. 59
4.5 Comparisons with Pyramidal Images. 60
4.6 The Optimal Generating Kernel for GMF Images. 67
4.7 Summary. 69

CHAPTER 5
OPTIMAL FILTERING FOR ROBOT VISION USING MULTIRESOLUTIONAL FOVEAL IMAGES

5.1 Relationship Between the Multiresolutional Images and Spatial Bandpass Filters in Robot Vision. 70
5.2 Optimal Filtering in Machine Vision and Image Understanding. 71
5.3 Filtering in Multiresolutional Images. 73
5.4 Simplified Analogies with Receptive Fields and Neuronal Inhibitory Conduction. 76
5.4.1 Receptive Fields near the Fovea and the Generation of High-Pass Filters. 78
5.4.2 Bandpass Channels at the Perifovea. 78

CHAPTER 6
PERIFOVEA-GUIDED SEMANTIC-LINE TRACKING
6.1 The Problem of Defining a Semantic Line. 82
6.2 Related Work in Line Following. 84
6.3 The Ideal Semantic-line Tracking and the Perifovea-Guided Semantic Line Tracking. 86
6.4 Perifovea-Guided Semantic Line Tracking Algorithm. 89
6.5 Formulating Perifovea Support. 92
6.6 Experimental Results. 95
6.7 Limitations in the Perifovea-guided Algorithm. 100
6.8 Summary. 100

CHAPTER 7
MULTIRESOLUTIONAL INTRINSIC IMAGES FOR OCCLUDED OBJECT RECOGNITION
7.1 Occluded and Unoccluded Object Recognition. 101
7.2 The Problem of Occluded Object Recognition. 102
7.2.1 Description of the Experimental Scene of Occluded Objects. 102
7.2.2 Minimum Set of Intrinsic Images. 105
7.3 Survey of Techniques for Occluded Object Recognition. 106
7.4 Adapting Intrinsic Images to Detect Occlusion. 107
7.5 Local Multiresolutional Foveal Images (LMF) as Intrinsic Images. 110
7.6 Algorithms. 110
7.7 Experimental Results. 113
7.8 Limitations of the LMF Intrinsic Image Approach. 116
7.9 Comparisons with Other Methods for Occluded Object Recognition. 117
7.9.1 The multiresolutional Pyramidal Method. 117

7.9.2 The Local Feature Focus Method. 117
7.9.3 Some Comparisons. 118
7.10 Summary. 119

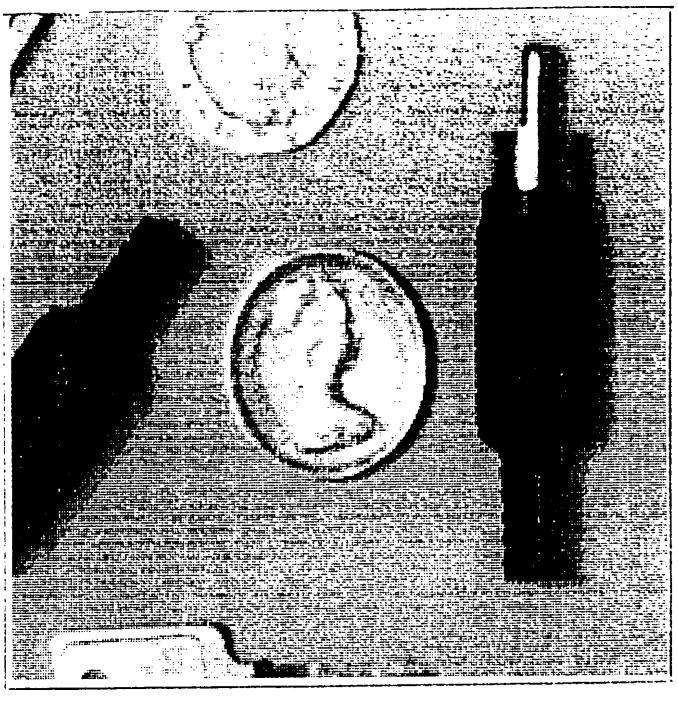
CHAPTER 8
The Multiresolutional Foveal Image in a Hierarchical Gross Coarse-Information-First Progressive Transmission Scheme.
8.1 Brief Survey of Image Coding Methods 121
8.2 Multiresolutional Images and Progressive Transmission Codes. 122
8.3 Multiresolutional Foveal Images for Progressive Transmission. 123
8.3.1 System Description for Progressive Transmission. 124
8.3.2 Decorrelation in the Bandpass Images. 130
8.3.3 Quantisation Advantage From Decreased Dynamic Range of Bandpass Images. 135
8.4 Experimental Results. 136
8.5 Comparison with Related Progressive Transmission Schemes 140
8.5.1 Comparisons with the Gaussian Pyramid. 140
8.5.2 Comparisons with the Laplacian Pyramid. 140
8.6 Summary. 141

CHAPTER 9
CONCLUSIONS
9.1 A Suggestion for the Hardware Implementation of Multiresolutional Images. 145
9.2 Suggestions for Future Research. 150

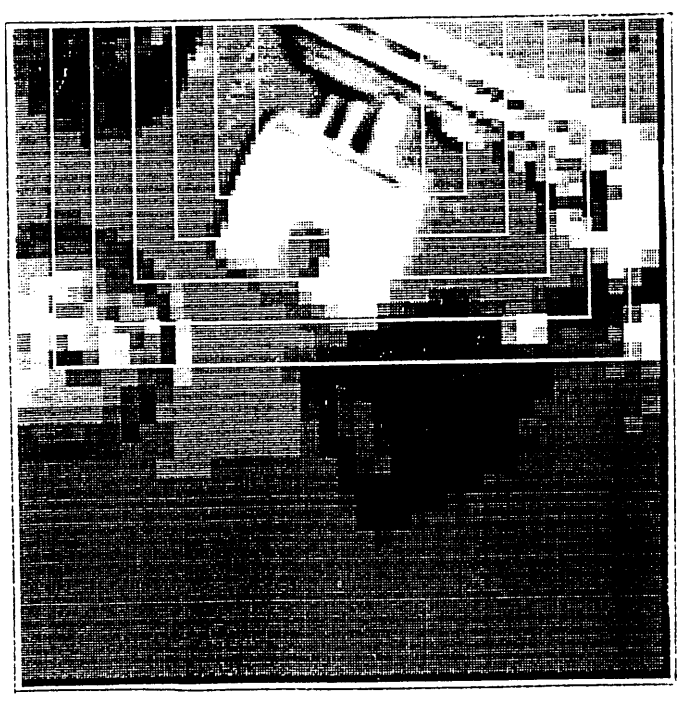
REFERENCES: 153 - 163

IMAGES USED IN THE THESIS.

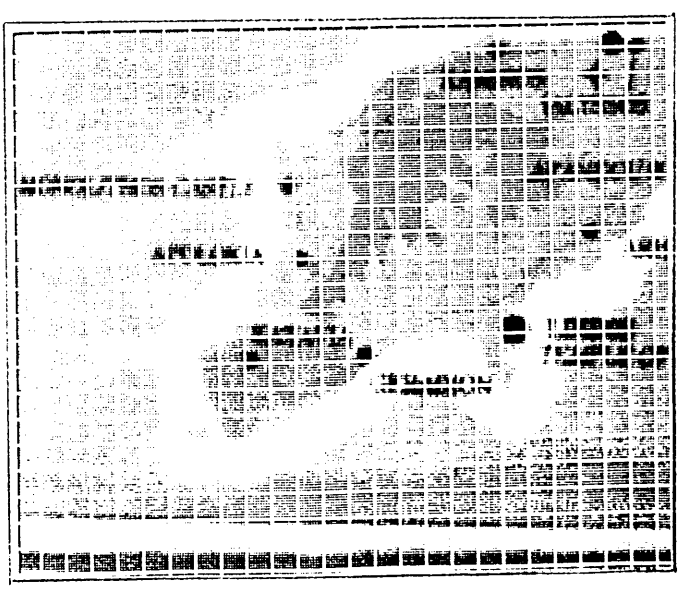
Gentl Image



Fcoins Image



Fkey Image



Rhodes Image

LIST OF ABBREVIATIONS

MF	Multiresolutional Foveal Image
GMF	Global Multiresolutional Foveal Image
LMF	Local Multiresolutional Foveal Image
LMSE	Least Mean Squared Error
DSP	Digital Signal Processing
∇	Gradient of Intensity-discontinuity of an Image
Δ	Magnitude or modulus of the Gradient
SIMD	Single Instruction Multiple Data
MIMD	Multiple Instruction Multiple Data
WISARD	Wilson Igor Stonham Adaptive Recognition Device
PLN	Probabilistic Logic Node
HVS	Human Visual System
LGN	Lateral Geniculate Nucleus
LPF	Low-pass Filter
BPF	Band-pass Filter
ZOH	Zero-order Hold
DOG	Difference of Gaussian
LTI	Linear Time-Invariant
PT	Progressive Transmission (of Images)

Statement of Originality.

The following paragraphs describe some aspects of the thesis that are original contributions.

In developing an alternative image representation for robot vision, the properties of the human visual system was the motivating factor. This led to the formulation of the Multiresolutional Foveal Image representation, (MF image) based on the foveal and perifoveal characteristics of the human eye.

This MF image has subimages of gradually decreasing resolution, extending from the image centre. The degraded resolution-factors of the perifoveal regions are represented as equivalent low-passed filtered images of different cut-off frequencies. These contrast with the multiresolutional pyramidal image representations in which each image layer has only one resolution factor.

The multiresolutional forvea-centred image was used as a composite mask for the generation of the semantic boundary of an object of interest. This mask is called the local MF image and is defined by a multiresolutional set of pixels, obtained from descriptions of both unthresholded edge-orientations and edge-strengths, thus providing a richer and more robust, weighted boundary detection scheme.

The study of the characteristics of this Multiresolutional image as a set of cascaded filters led to the search for a possible correspondence with optimal filters for robot vision, as proposed by others. By combining the properties of the Cornsweet Model for backward neuronal inhibitory interaction and the properties of visual receptive fields, it was found that spatial bandpass filters do result from this MF image representation. It was found that such bandpass filters in the MF image representation had bandpass-filter variables that make them arguably a suitable candidate for a class of optimal filters for robot vision.

The adaptation of the Multiresolutional Image Representation as a basis for an alternative image coding scheme referred to, by various authors as 'gross coarse-information first' coding, further showed that it does possess properties of compact image storage, simple reconstruction and interactive transmission for visual images, which are often required in a robot vision applications.

Two suggestions on the possible methods to generate Multiresolutional Foveal Images in near real-time were proposed. The first method is based on the properties of the optical defocussing of visual images in a specially structured CCD array inside the vision camera. The second method is more conventional, using modified electronic timebase circuitry, together with pipeline, hardware, analogue low-pass filters as generators of the Multiresolutional Fovea images.

ACKNOWLEDGEMENTS

I am grateful to my Supervisor Professor I. Aleksander who has provided me with the right conditions and helped me along with beneficial discussions in the course of this study.

I must acknowledge the motivations that I have been able to draw from the work of Professor Aleksander in his novel approaches to Artificial Vision. This has given me the courage to look at some aspects of the Multiresolutional Foveal Images.

I must thank my colleagues A. Redgers, C. Myers, T. Ludemir, N. Panayotis and A. Fulcher for the many discussions on neural systems. For the oportunity to use the fine image-processing workstations and associated hardware, I owe a word of thanks to the Signal Processing Group under Prof. A. G. Constantinides.

Finally I must thank the Economic Development Board of Singapore and the Divisional Director, Mr. C.T. Lin for this research scholarship and financial assistance to participate in two international conferences relating to robot vision.

CHAPTER 1

Image Representations and Robot Vision.

Image Representation and Robot Vision can be very much interrelated. Robot Vision must inevitably use an appropriate representation to perform its computational tasks. An image representation therefore constitutes the data structure or knowledge representation of the image to be computed upon. It not only specifies the way in which image entities at its various levels are to be encoded and described, it also directly or indirectly implicates the hardware and computational efficiencies of Robot Vision. These ideas will be expanded upon in the definitions that follow.

1.1 The Nature of the Problem and some Definitions.

Robot Vision, sometimes also loosely referred to as **computer vision or machine vision** is the application of **image processing** techniques for robotic tasks [Horn86]. Rosenfeld [Rosen87] makes a distinction between image processing and robot vision in defining image processing as a transformation of one image to another image with some desired characteristics (e.g smoothed) without a need to provide any description, whereas robot vision transforms an image with the aim of generating a description of the image to the user (in this case the robot). In the broadest sense, robot vision may be defined as the automated process of extracting, characterising and interpreting real-world, 3-D image information and may be subdivided into the following areas:

a)image sensing and acquisition, b)preprocessing, c)segmentation, d)description, e)recognition and f)interpretation. [Fu,Gonzales,Lee87].

These tasks can be grouped as low, mid, and high-level processes and a diagrammatic interpretation of these tasks is illustrated in fig 1.1 where some details of these processes commonly used in present-day robot vision are also described. Robot Vision would be more powerful if it could readily use some of the richer real-world image information like motion,stereo perception, colour etc. This may be possible in the near future with faster and parallel systems without the limitations of **data I/O** and other bottle-necks but generally, current **real-time passive** industrial robot vision systems have to be understandably and practically restricted to process images that are *2-D planar and bilevel* with a high-contrast and a easily-segmentable background [Rosen87].

The techniques used in robotic vision are also relatively multi-disciplinary in nature and cover a broad area, ranging from image and signal processing, (image transformation, encoding and transmission), pattern recognition (feature extraction, statistical and syntactic), geometric modelling (model fitting), cognitive processing and Artificial Intelligence techniques (searches, algorithmic graphs, goals and plans). It is therefore not too surprising that

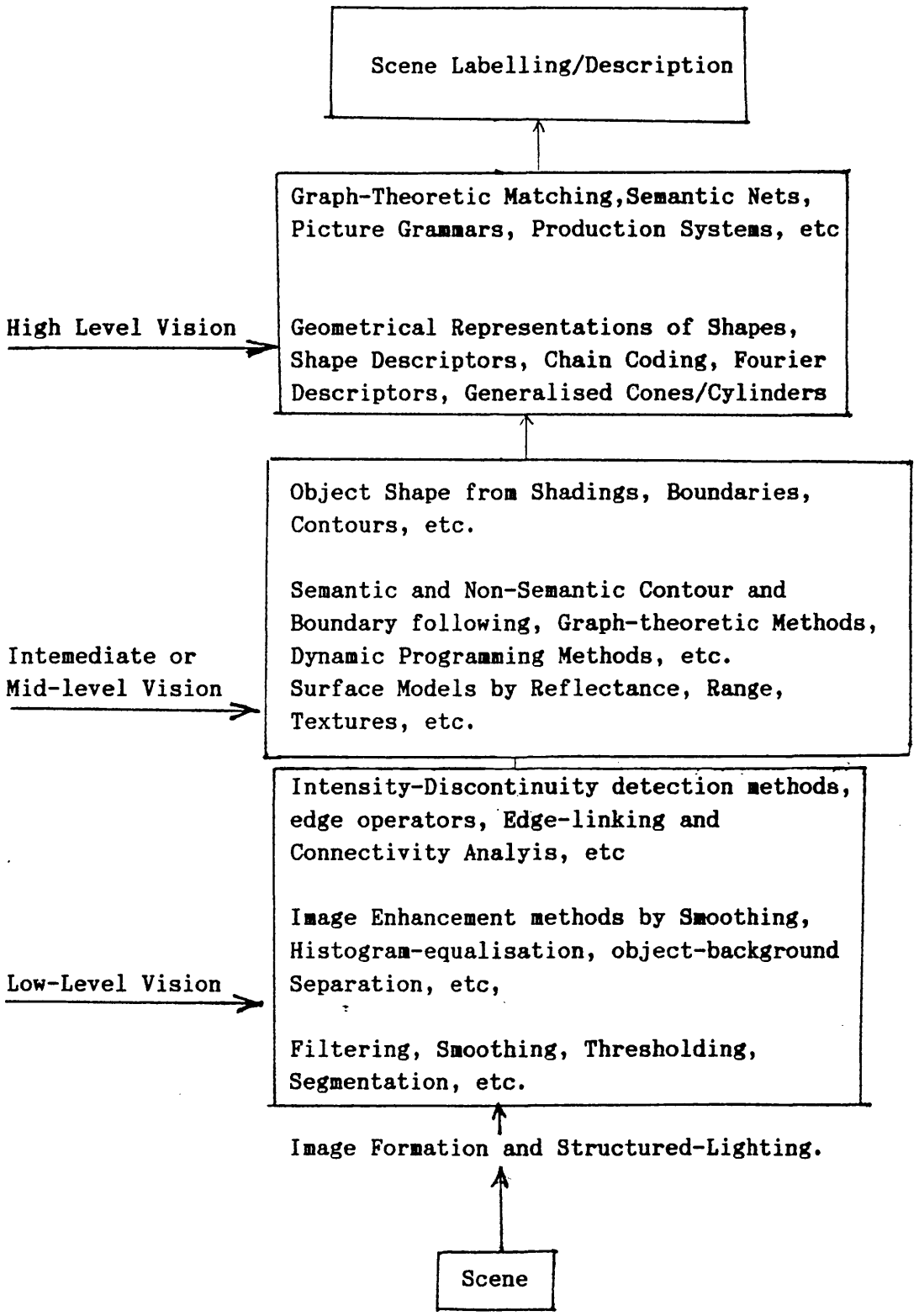


Fig. 1.1 Illustration of Various Processing Levels of monocular Static Robot Vision.

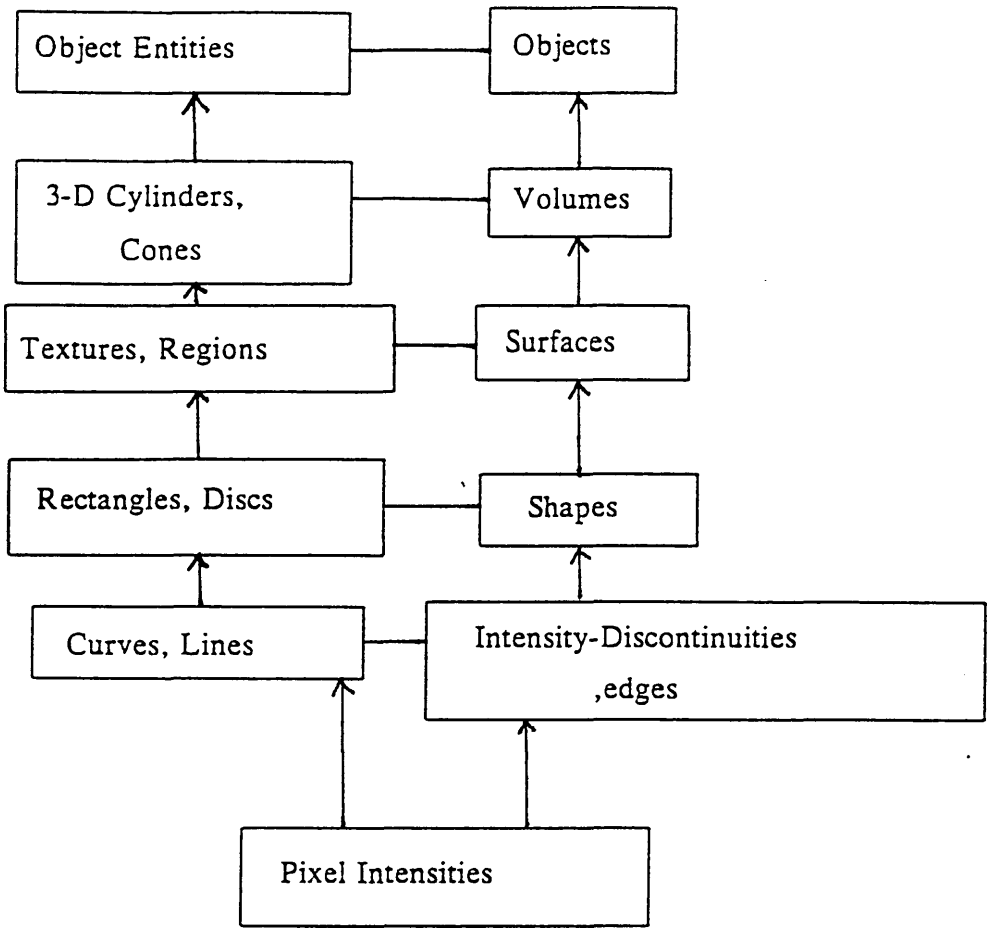


Fig 1.2 Image Representations at different levels

it can be considered as one of the more difficult technological challenges today.

Image Representation.

A **representation** is a *formal system for making explicit certain entities or types of information, and a specification of how this has to be done.* [Marr82]. Representation plays a crucial role in determining the computational complexity of an information-processing problem. In enabling the computer to process the images, a suitable representation of the visual image must be employed that will encode the visual scene such that the essential knowledge and *structural relationships* in the scene are maintained. Such a knowledge representation of the image can be described as the **image representation.** Image representation is not a strict definition as it also implies a *range of representations* (depending on which processing level as in para 1 above) whose appropriateness as a good **representation** depends on the computational level in question. These representations are shown in figure 1.2 which emphasises the various representations due to the different image processing levels and the *nature* of the image. It will be noticed from the figure that the choice of an image representation at any lower level will implicate the computational complexities (serial/parallel operations), data structures (graphs, nodes, trees, lists etc) and algorithms at the higher levels. and therefore a judicious choice of the right representation at the right level will be quite a deciding factor on how good is the solution.

1.2 Real-Time Robot Vision: Difficulties and Problems

It may be necessary to emphasis that one of the aims of Robot Vision is its ultimate capability to execute its tasks in real-time or near real- time. This really means that solutions should be obtained within a time suitable for the task at hand, or within an equivalent time if performed by man (e.g parts inspections in a assembly line). The human model of real-time vision processing is sometimes employed as a standard (e.g. geometrically invariant object recognition) since it is considered the most superior for such applications. This creates a desire to incorporate features of human-like visual processing for machine vision as is also pursued in this thesis. The difficulties and problems in real-time robot vision are not only due to the enormous amount of image information that needs to be processed but how best to make inferences from the information that is available using the existing hardware for image acquisition and information processing, which, at a non-laboratory environment, is likely to be a standard Von-Neuman machine with conventional *Television Standard front-end systems*. These difficulties and problems may classified into 4 main areas in paragraphs 1.2.1 to 1.2.3.

1.2.1 A case of a Ill-posed Problem.

The task of visual perception is acknowledged to be intractable in a straightforward manner. Even the subproblem of polyhedral labelling have been shown to be *NP-complete* [Kirousis and Papadimitriou85]. This is notwithstanding the issue of the ill- posedness of the problem. It is quite widely accepted that many of the problems in Computer Vision are **ill-posed**. This means that the problem domain contains uncertainties that cannot be resolved by the problem solver. One way of easily describing this ill- posedness is to consider a dimensionality inference task in Computer Vision where rich 3-D or even 2 1/2-D descriptions are expected from a essentially 2-D image frame without sufficient inputs from higher level knowledge and subjected to the vagaries of optical physics and global assumptions (e.g shape from shading). Current trends in **Active Vision and Multisensor Fusion** [AAAI87] are indicative of ways in which these ill-posed problems are proposed to be solved.

1.2.2 Problems in the hardware.

It is difficult to pin-point one single area in hardware that would be considered a crucial impediment to Robot Vision. Whereas in earlier years,the image acquisition front-end was considered in need of improvement in terms of spatial resolution, quantisation errors and speed of frame- capture, advances have been rapidly achieved in recent years where spatial resolution in excess of the human visual acuity and real- time colour frame-grabbers are not uncommon. It seems that at the present, the problem relating to Computer Vision hardware is that of *how best to use the enormous image information* that is available, in a way that is faster and more Gestalt-like, than at present. While novel computer architectures like cellular arrays can solve certain classes of computer vision problem [Uhr72], the author feels that a more economical and imminent solution perhaps lies in the wider use of hardware systolic **or** pipeline circuitry, special arithmetic processors and other digital Signal Processing Chips, together with processors capable of both serial and parallel processing (e.g. Transputers), as far as passive real-time Industrial Robot Vision is concerned.

1.2.3 Problems in the Computational Approach.

The multi-disciplinary nature of Robot Vision requires a study of a possibly unified approach in the Computation. Although in earlier years, the human perception model was suspected to be a detractor to how best Robot Vision should be approached , there is now more consensus that the human perception model is a more realistic model for Robot Vision. Workers like Marr[Marr82],Hanson and Riseman [Hanson80], Uhr [Uhr72],

Macclelland, [Rumel/Macc86], Aleksander [Aleks87] have, in their various ways, incorporated features of human visual perception for robot vision. Although it is felt that a single **Universal Computational approach** seems more unlikely now as the problems in Robot Vision become more user-specific,with new technologies being explored (e.g. active illumination, active viewer and multi-sensor fusion) an understanding of the fundamental difficulties is still very desirable.

1.2.4 Limitations of Image Representations.

Following the classification of Rosenfeld [Rosen82] and concentrating on the hierarchical characteristics of the representation, it is convenient to classify image representations into two categories:-

a) **Hierarchical or non-traditional** image representation and

b) **Non-hierarchical or traditional** image representation.

The traditional image representation is defined as those image representations that treat each pixel in the image as an entity of equal importance or weightage as in uniform fine-resolution images of conventional robot vision. The non-traditional image is characterised by the variation in importance of the pixels within the image as in *pyramidal or hierarchical* image data structures. Each of the representations above has a particular suitability for certain robot vision task. As an example, the pyramidal representation is useful where hierarchical coarse to fine searches are required but has the weakness of being unable to handle image pixels with intermediate fine-resolution as each pyramid layer has a finite logarithmic-order resolution. The traditional image representation is not as elegant when used in performing data abstractions in a more global window, but has the ability to perform global uniform-resolution, low-level searches (e.g Hough searches). Other methods of image representation using *transform Methods* (e.g Fast Fourier, Discrete Cosines, Hardamard, Haar etc) though generally used with uniformly weighted coefficients can similarly be used in a hierarchical manner by ordering these coefficients. A clear distinction between hierarchical and non-hierarchical representations is not as easily defined since it is always possible to have a means of creating a hierarchy in the computational structure in order to make good use of such *hierarchically structured data*. However for speed and other desirable properties in computation, it should be more desirable that such hierarchically data should be derived in a *natural* and as *early* as possible. Further a desirable image representation should also *isomorphically map* [Tsotos87] one domain to the other so that the representation maintains the features of the domain as much as possible throughout the hierarchy. This can be explained by noting that if the knowledge of spatial relationships in a scene is necessary for vision, then it should be desirable that spatial information be directly and easily extracted from the representations.

A suitable image representation that has the optimal combination of the good features of the

various representations, especially for robot vision would be a very desirable goal. It is therefore believed that the Multiresolution foveal Image representation in this thesis fulfils some of the requirements of a desirable image representation for Robot Vision, in that it possesses the following essential characteristics:- [Fong(INNS-90), Fong(ICARCV-90)]

a. There is a *hierarchy* in the structure from the various graded-resolution images.

b. The hierarchy is *naturally ordered* through the use of *smooth transitions* in the multiresolutional images.

c. *Emergent* properties of such arrangements of hierarchical processing can arise from interactions between the hierarchies. As an example, the primitive hierarchically-ordered low-pass images can generate spatial bandpass channels, such 'independent narrowly-tuned channels' being a reason for optimal filtering in the human visual system [Sullivan82]. These aspects will be modelled in chapter 5.

1.3 Objectives and Outline of the Thesis.

In the following paragraphs the main objectives and the outline of the thesis shall be described. In essence, it hopes to explain why the Multiresolution image representation with the finest resolution in the central part of the image called the **Multiresolution Fovea-Centred Image Representation** or **MF image** for short, possesses features making it worthy of further investigation as a suitable image representation, especially for Robot Vision. It has been strongly motivated by the arguments of experimental psychologists [Sull82],[Campbell and Robson68] that the optimal vision processing for *image understanding* is a set of *spatial vision channels* and the well-accepted knowledge of the structure of the vertebrate retina with recent findings of the visual neuronal processing in the aspects of receptive fields and inhibitory actions [Cornsweet70].

1.3.1 Primary Objectives of the Thesis.

The primary objectives of the thesis shall be to investigate if there exists an alternative image representation that has desirable features for robot vision tasks. Such a image representation shall be formulated, analysed and used in two applications for robot vision tasks of a) semantic object-segmentation using **perifovea-guided line-tracking**, and b) occluded object recognition using local MF images that provide *Intrinsic images* in the context of Barrow and Tennenbaum [Barrow and Tennenbaum78]. The multiresolutional structure of the retina shall be modelled as as set of cascaded low-pass filtered images with properties of spatial bandpass,emerging from receptive fields and inhibitory actions. A study will also be made on the properties of the MF image representation as a compact and lossless image code, for which any desirable image representation should also concurrently possess. Constant reminders shall be made to the need for real-time performance and engineering implementability of the solutions.

1.3.2 Related Work.

Most Multiresolutional models point to the earlier work of Kelly who used a multiresolutional image to obtain the outline of a human head by the *planning* approach of Minsky [Mins61]. Although Kelly's image was a thresholded binary image of 32 by 32 pixels, it was shown to perform considerably better than current approaches of its day. *Recognition Cones* were developed by Uhr [Uhr72] as a visual computation model emulating the hierarchical processing in human visual perception. Similar pyramidal or cone systems were developed by Hanson and Riseman [Hanson80], Tanimoto [Tani79], Klinger [Klinger76] and Levine [Levine85] and Burt [Burt86]. Although there are some differences in the their various approaches, they are essentially an application of a parallel set of operations (generally called transforms) on the raw image at various layers with each layer processing the transforms generated by the previous layer, until the highest layer is reached. They could appear as operations on quad-trees, octrees, regular recursive decompositions or just plain pyramidal graph operations but the essence is a fast hierarchical computation on the nodes of a graph. Further related extensions were made for specific computer vision tasks using these hierarchical images for image segmentation [Trevor and Wohn88], model-based object recognition [Neveu and Chin86], Gaussian Pyramids [Tani80], Laplacian Pyramids [Burt86].

The work in this thesis is most related to the Gaussian and Laplacian pyramids, though with the difference that the emphasis is towards robot vision tasks rather than on general parallel image transformations (e.g image smoothing, enhancements and non-semantic edge detection etc), for which much of the pyramidal systems have been directed towards.

1.3.3 The Multiresolutional Fovea Image.

In the most simple way, we can say that the MF image is an attempt to have an image that has, as much as possible, the properties of image-acquisition through the graded resolutions of the vertebrate retina. This MF image is relatively crude and primitive, having ignored descriptions of colour, motion, stereo perception etc. Two formulations of the MF image shall be studied, namely the **Global MF image** (GMF) which attempts to simulate the image in the *pre-attentive* phase of object perception and the **Local MF image** (LMF) which simulates the *local focused-attentional* phase of object perception [Marr82].

1.4 Organisation of the chapters.

Chapter 1 introduces the definitions and a statement of the problems of robotic vision in general pointing out the multi-disciplinary nature and the ill-posedness of robot vision problems. It describes the strong link between good image representation and robot vision in establishing sound computational approaches and introduces the Multiresolutional Foveal Image Representation as a

representation for some robot vision tasks.

Chapter 2 is a review of the approaches and techniques available to solve a selection of the robot vision tasks. It discusses the concepts and approaches of some prominent workers in this field of Computer Vision, these approaches being selected with regard to their seemingly differing solutions towards general Robot vision problems. Briefly, these are:-

a. The importance of a *Structural Description* of an image through the creation of intrinsic images of Barrow and Tennenbaum.

b. The need for a *Computational Theory of Vision* using some aspects of human perception as a model.

c. The approaches of faster solutions to image processing by the use of cellular array processors[Duf85]and the pyramidal approaches of Uhr and Tanimoto.

d. The solutions to some specific problems in real-time robot vision as recognition devices, through the use of some equivalent massively parallel random access memory computations, as in the approaches of Aleksander [Aleks79] and further possible extensions in Probabilistic Logic Nodes (PLN).

e. The approaches of K.S. Fu using *syntactic methods* Fu[Fu74]

f. The industrial methods of robot vision as simpler model-based searches which have tested successes in practice [Perkins78].

Chapter 3 surveys the hierarchical and the non-hierarchical image representations that have been proposed for computer vision problems. The dichotomy follows that of Rosenfeld in that multiresolutional and pyramidal image representations are considered hierarchical.

Chapter 4 introduces the details of the Multiresolution Foveal image. It describes the structure of the image, the motivations, the biological and psychological parallels in using such a representation for robot vision tasks. The choice of a suitable generating kernel for the perifoveal areas is discussed. It also shows the computational advantages in using the multiresolutional foveal images.

Chapter 5 provides a mathematical analysis of the filters that generate the multiresolutional images. Using the backward inhibition model of Cornsweet [Cornsweet70], these non-recursive filters are modelled as a cascade of low-pass filters of progressively increasing cut-off frequencies. The inclusion of the the interactions of the *receptive fields* and *backward inhibition* thus producing changes in the **configurations** of these filters, shows that separate spatial filtering channels can be produced.

Chapter 6 uses the LMF image in an application to obtain a semantic edge of an object of interest that otherwise is not segmentable using non-semantic boundary tracing methods. It uses a set of variables to decide if an intensity-dicontinuity should be considered a valid edge point and subsequently

uses these confirmed edge-points in line-following. Line-following is performed by matching semantic-line descriptions. The use of perifovea-support reduces the possibility of false searches.

Chapter 7 is another application on the possible use of the MF image for a robotic vision task of object recognition for objects that are partially occluded. This follows the *Intrinsic Images* approach of Barrow and Tennenbaum, except that the intrinsic images are not generated for the whole of the image but as foveal fixations by the LMF images. It contrasts with the more conventional approach of matching the object boundaries/shapes, in that the grey level information is used to infer some basic 2.5 D information from a minimum set of intrinsic images.

Chapter 8 studies the MF image as a possible image structure for progressive transmission in a robot vision environment where still images may need to be shared amongst users. The multiresolutional foveal image progressive transmission scheme is compared with that of the Gaussian and Laplacian pyramidal schemes.

Chapter 9 provides the conclusions and a summary of the main points of the thesis. It lists out the major advantages and limitations of the MF image representation and describes some areas for further work.

CHAPTER 2

ROBOT VISION: REVIEW OF APPROACHES AND TECHNIQUES

In a real-world scene, objects are characterised by many properties. These properties include 3-dimensional object descriptions, surfaces, shadows, occlusions, relative motion of entities in the scene, textures, colours etc. To-date, the majority of robot vision systems have mainly used the grey intensity image with very little use of the other information. The reasons are well justified considering the vast amount of visual data that have to be processed. However, in so doing, it has also imposed quite insurmountable problems for the more widespread and general use of robot vision. Statistics on robot vision applications since 1982 [Wallace88] show that the following are the 3 main applications of Industrial Robot Vision:

a. Visual Inspection: Analysis of image dimensions or other figures of merit (pcb defects, dimensional inspections,surface finish, burrs in plastic extrusions etc.).

b. Object Recognition: Recognition of industrial parts.

c. Visual Control: Use of the visual information to control manipulators, and robots.

According to Meyers [Meyers85] world-wide applications in visual inspection and object recognition constituted 70 percent of all robot vision applications in the year 1985.

The robot vision tasks in this thesis is mainly biased towards the restrictive task of object recognition. However an additional difficulty is introduced in these investigations in that the image has a *visually complex scene* background, where visually complex here means that the objects are not easily segmentable from the background using simple histogrammed or other global, non-semantic descriptions. Such an example of a scene is shown in fig 2.1a which shows an example of multimodal grey level distributions of the image. This is contrasted with fig 2.1b of a scene with sufficiently high contrasted background that relatively simple global, non- semantic descriptions will be able to obtain the desired object from the other objects in the scene.

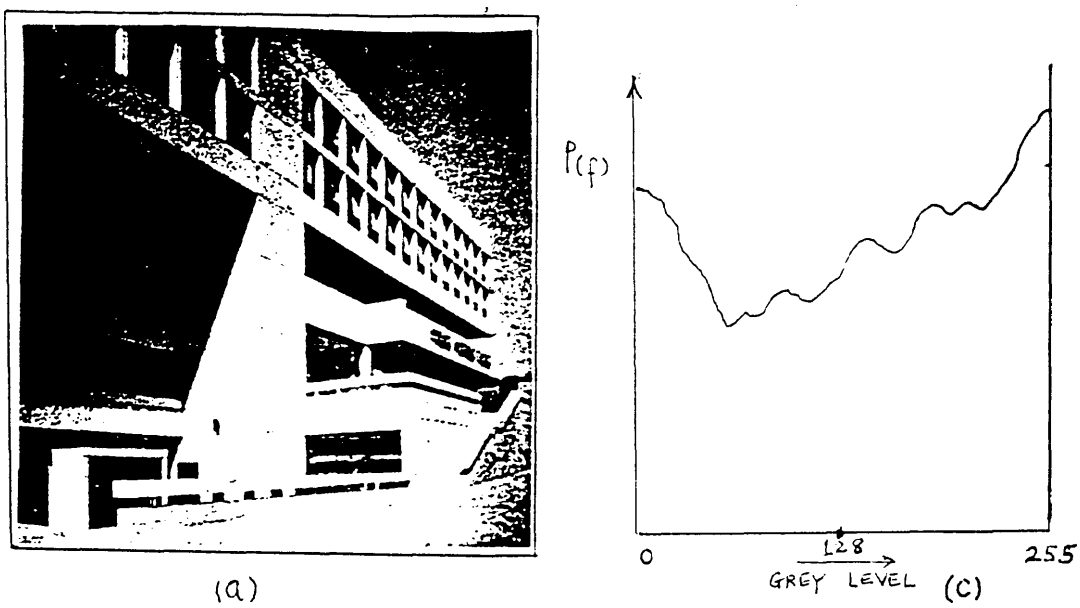
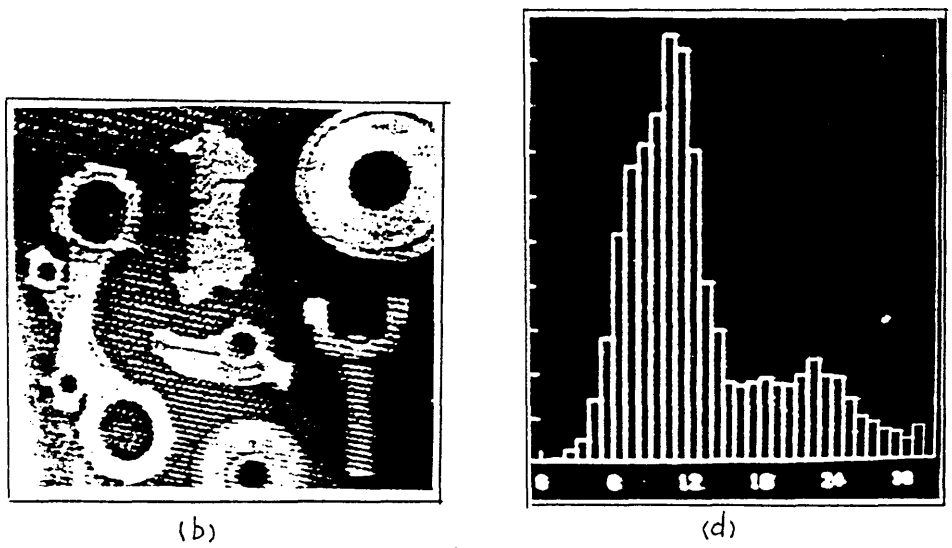


Fig. 2.1a Example of "Visually-Complex" Scene and
Fig. 2.1c Multimodal Histogram of Visually Complex Scene.



Source: W.A. Perkins, "Area Segmentation of Images Using Edge Points," IEEE, PAMI-2, Number 1, January 1980. © *1980, IEEE. Used by permission.*

Fig. 2.1b Example of "High-Contrast, Pre-engineered" Scene and Fig.2.1d Unimodal Histogram of High-Contrast Scene.

Fig 2.1b is usually obtained with some form of *structured lighting*. The corresponding grey-level histograms in fig 2.1c and 2.1d will show that in the case of fig 2.1c, some simpler histogrammed-equalised thresholding method will not be able to segment the object from the background, whereas in fig 2.1d, the histogram shows an obvious modal distribution so that global methods of thresholding will readily yield the object from the background.

It may be not too inappropriate to further emphasise that a visually complex scene provides a big challenge to pattern recognition and can sometimes be a central issue in scene analysis. This is summarised in this tautological argument below:

> "In a Visually Complex Scene, it is necessary to isolate the object of interest from its background or from other objects so that the isolated object can be identified or recognised. In order to isolate the object from the background it is necessary to recognise the object so that segmentation can proceed.(which was the original task at hand)."

This shows an example that robot vision in a visually complex scene is a non-trivial task.

2.2 Some Approaches to Robot Vision.

As has been previously mentioned, amongst the major categories of usage of Robot Vision, object recognition in a visually complex scene is the major concern of this thesis. Although the main approaches

described in the following paragraphs, do not specifically address the issue of robot object recognition, they do address the wider problem of *robot vision*, the understanding of which do offer a generic solution to many robot vision problems of which object recognition is one.

The approaches of Barrow and Tennenbaum, Marr and Aleksander share a common theme of using some analogues of the Human Visual System (HVS) in robot vision. In Barrow, the approach to robot vision emphasises the need to obtain more information than just the boundary of shape information through the derivation of *intrinsic images*. Marr, amongst other exhortations, stressed the need to emulate some aspects of the the HVS in obtaining better 'structural information' from the scene and model the objects in the scene using a set of consistent *generalised cones/cylinders* representation. Aleksander's approach is to make full use of **engineering-implementable** networks of Random Access Memories as networks that could be *adapted*, thus providing a desirable property of *generalisation capability* in robot vision problems, posed as pattern recognition tasks.

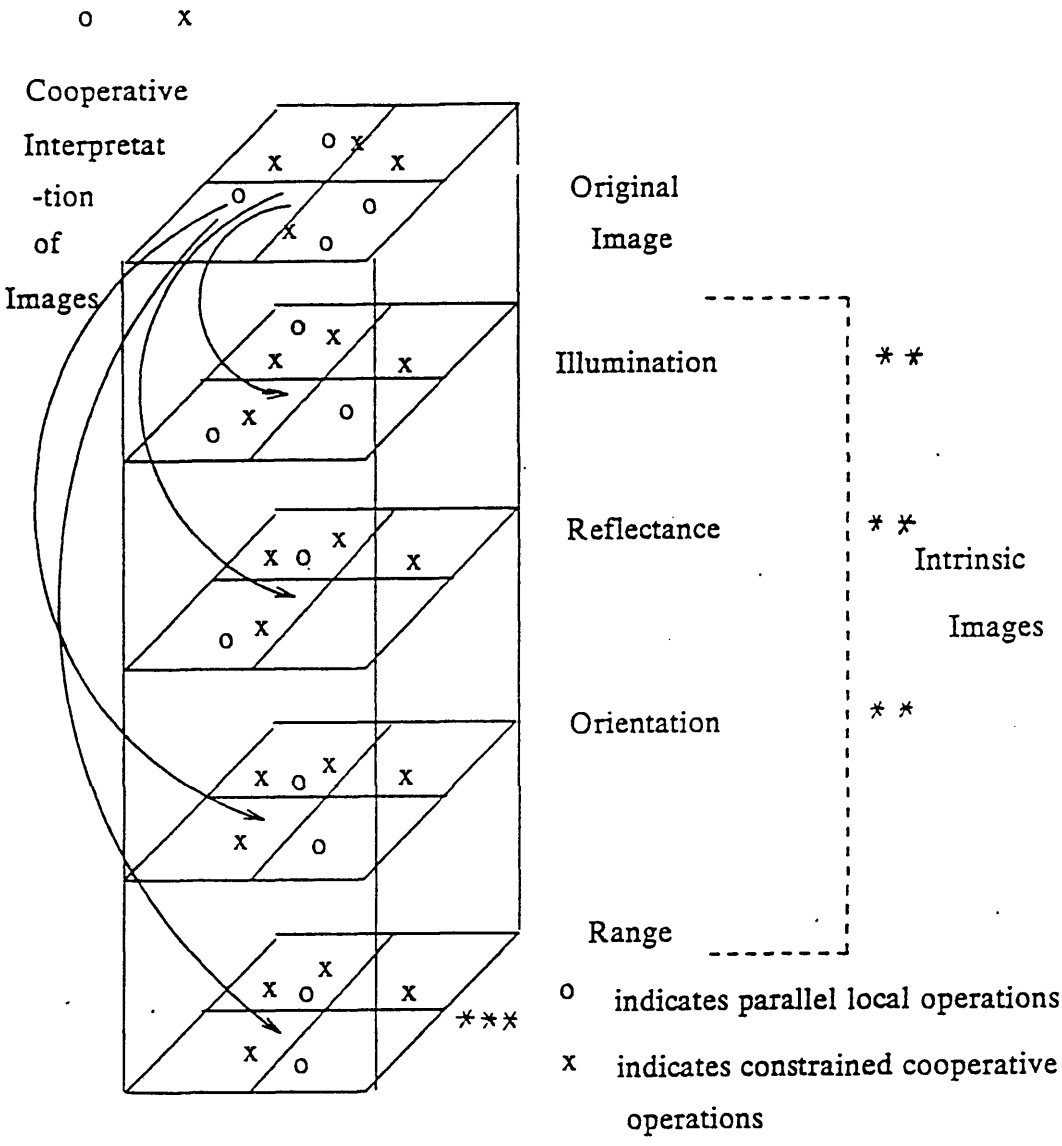
The approaches of cellullar arrays for image domain. These have been found to provide speed advantages of $O_{(n)}$ compared to the more conventional serial computations. Though such cellullar image operations are not primarily aimed at robot vision, its ability to perform fast image transformation and other lower- level preprocessing tasks like image enhancements, smoothing and other global operations could be employed as a precursor to robot vision. The Syntactic and Structural Approaches of [Fu 74] makes use of the established *Language Theory* as a computational procedure to resolve the image *primitives* and structures that have been already derived.

2.2.1 Recovering Intrinsic Scene Characteristics from Images.

The approach of Barrow and Tennenbaum [Barrow and Tennenbm78] is based on the premise that an image should not just be described by intensities but that these intensities should be used to provide *intrinsic information*. These intrinsic information should be generated for every point of the scene and may be descriptions of range, reflectance, incident illumination, surface orientations etc. to be used later by higher level analysis. Although not specifically stated by Barrow, the process of obtaining these intrinsic images need not preclude the use of other *active methods* like laser ranging etc., its primary aim being to recover as much of the 3-D information as possible of the scene. It is only through the application of such intrinsic images that there can exist a more unique interpretation of the scene [Barrow78]. As an example, a shadowed area is characterised by a region having intensities proportional to the reflectance and constant over its surface. Hence in the context of intrinsic images, an instance of an edge can is defined as a point corresponding to the extreme boundary surface or to the boundary of a cast shadow. Using a set of such definitions of intrinsic image features, the *scene understanding* process can be summarised in the following steps:

a. For the whole of the scene, obtain the intrinsic image features.

b. Find the step discontinuities in the input image (e.g. edge points).



** and *** shows existence of non-obvious model-driven computation

Fig 2.2.1. A cooperative model for the generation of Intrinsic Images as suggested by Barrow/Tennenbaum

c. Interpret the intrinsic feature of these discontinuities according to the set of intrinsic features defined by their *constancy or tangency* characteristics.

d. Hypothesise on the values to be assigned to the intrinsic images.

e. Perform relaxational computations over the whole of the region covered by the discontinuities.

This procedure is illustrated in the fig 2.2.1. It can be observed that Barrow and Tennenbaum were very aware of the need for *model-driven* computations but have chosen to alleviate the problem by alternatively requiring the model-descriptions at higher image processing level as is shown in the area marked ***. In fact, although it is not obviously evident, there is in fact a model-driven computation even at ** by making the observation that the model is that model of '*Lambertian* surfaces' of 3-D objects under incident light illumination which is the source of the constancy, tangentiality characteristics.

In as much as this has some similarities with the Marr approach to be described later, in ending, it may be timely to highlight what the author feels to be two major differences from the Marr approach:

a. In the case of Barrow and Tennenbaum, range information is primarily obtained from the image-formation characteristics on 2-D surfaces whereas Marr prescribes the use of stereo image and texture cues.

b. In Barrow, investigation of low-level process should focus on what type of information is being sought, and how it might be obtained from the image. In some ways, contrasting with that of Marr, the design of edge detection need not focus on *intensity-discontinuities* (which is a necessary part of the derivation of primal sketches of Marr) but should rely on the physical meaning of the *type* of edge as a function of a combination of parameters (x,y,a...)

$$E = f(x,y,a,b,c)$$

where x,y are the spatial coordinates and a,b,c are the intrinsic features of range, reflectance,orientations etc.

2.2.2 Computational Theory of Marr.

Although the works of Marr on Vision was primarily directed at neurophysicists, his ideas provided a very strong influence on research in vision in general. In a 3-D object recognition case , the approach proposed by Marr can be seen in fig 2.2.2.

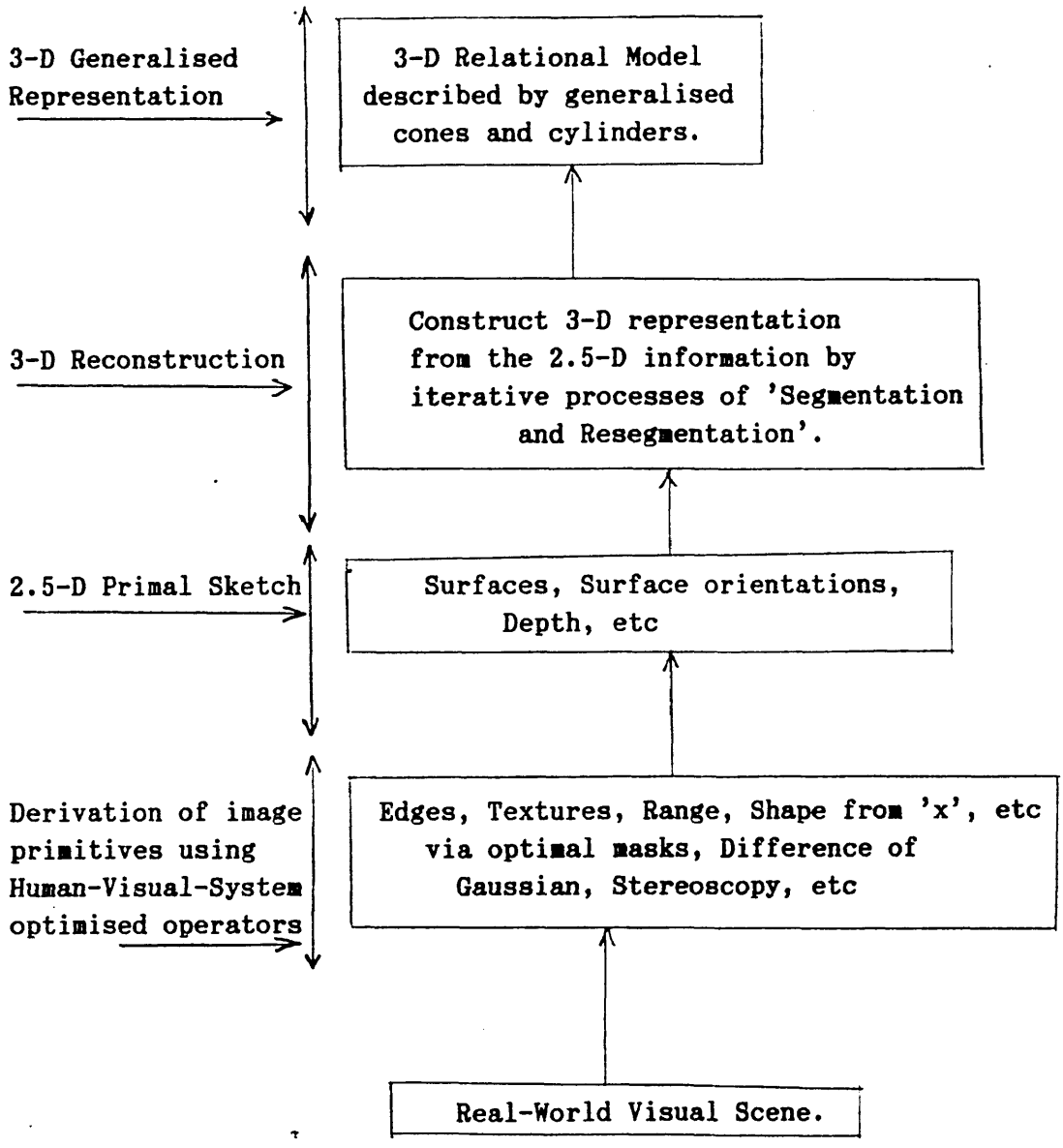


Fig.2.2.2 Marr's Computational Approach to Vision

The real-world objects are 3-D but the aim is to obtain the *2.5 D sketch* through the use of various techniques described as *resegmentation of 'x from y'* where various inference schemes are used to obtain for example *shape from shading, shape from textures, etc.*. The 2.5 D surface map has to be feature-extracted or resegmented to finally yield a representation containing consistent relational structure descriptions of the objects. Finally, the resulting relational structure could be converted to the form of a visible-surface labelled-graph or similar, from where object recognition could proceed.

A brief summary of the pertinent steps representative of the approach by Marr can be described below:

a. From the visual scene (3-D objects and surfaces), the first step is to use as much of the available information (stereo-opsis, texture, shape from x ,etc) to obtain the 2.5 D sketch. The 2.5 D sketch now contains, amongst others, information on surface orientations, depth and discontinuities of visible surfaces.

b. Obtain the 3-D model representation from the 2.5 D by an iteration process called *segmentation and resegmentation* on the early vision processes.

c. Formulate the 3-D model in a knowledge representation using some canonical symbols (e.g. generalised cones) to form a relational structure for interpretation.

Marr put some considerable emphasis on the human vision system in deriving information even at low-levels (e.g. difference of Gaussian Mask for optimal edges etc) and the use of object representations that are more recoverable (e.g. generalised cones instead of skeletonal representations) and other properties that are more likely to provide reliable information to the higher cognition processes (e.g. image correspondence and disparity effects in stereoscopy, texture discrimination effects on region segmentation etc).

If it is possible to select 3 main issues that Marrs theory contributes to the work of **monocular object recognition**, the following will be a probable list.

a. The importance of early processing to extract visible surfaces from the shape and dispositions relative to the viewer.

b. The importance of the model-based information to derive the 3-D information of the surfaces.

c. The importance of having a stable shape representation using generalised cylinders and cones. This contrasts with other shape representations that are sensitive to shape deformations. (e.g. Blums grassfire 2-D shape representations).

2.2.3 Cellullar Array and Multiresolutional Approaches.

The cellullar array and pyramidal approaches have been grouped together in this chapter, as

pyramids can be regarded as a SIMD (Single Instruction Multiple Data) cellullar array, configured as a pyramid of processors [Tani83]. Pyramids have some advantages of being able to handle a more variable-distance pixel neighbourhood which are particularly needed for multiresolution image processing [Duf 85] Cellullar arrays on the other hand can also be regarded as special cases of kindred architectures that include MIMD (Multiple Instruction Multiple Data) [Kruse82], systolic arrays [Kung82], pipeline architectures [Sternberg83] and other Multiprocessor architectures that have been proposed to speed up the image transformations.

The cellullar array approach of the CLIP4/5 [Duf85] uses 8 processors with 32 bit storage in an architecture shown in fig 2.2.3

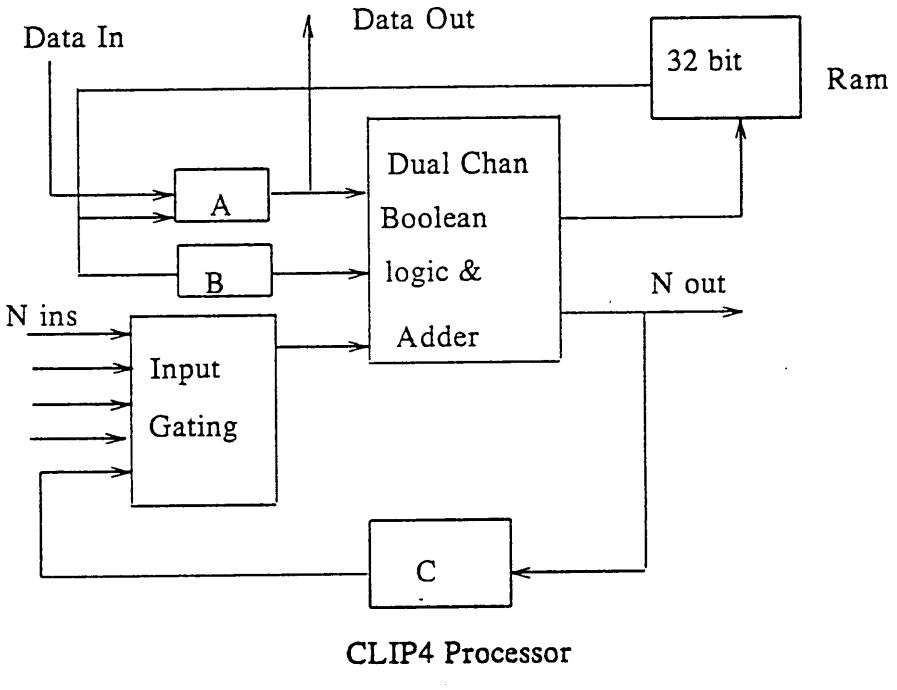


Fig 2.23 Basic Clip4 Architecture

Logical operations on the input neighbourhood pixels are carried under SIMD control on an input array of 96 x 96 pixels. Pyramids have been built from SIMD processors with each processor being able perform pattern matching and other bit-serial arithmetic [Tani83].

2.2.4 The Logical-Node Neural-like approaches of Aleksander.

Though the main function of Pattern Recognition is to provide a statement about the input image [Aleks83], there is a wide range of robot vision task that can be formulated as pattern recognition tasks [Pott83]. as illustrated in fig 2.4. The WISARD [Aleks84] uses Random Access Memory elements to perform logical *N-tuple Sampling* and *discriminator functions* defined by n-tuples of $\Sigma w_j i_j$ where w_j is the weight associated with the image pixel i_j. This overcomes some

difficulties of plain mask or template matching thereby allowing the system to be trained by examples. Industrial systems of WISARD have been built with near real- time performances for certain vision tasks [Aleks84]. Further, the judicious use of multiple discriminators and various decision-making regimes enable the WISARD to handle recognition of objects subjected to varying degrees of geometric variations like rotation, translation and scale. Recent extensions in the form of *Pyramidal Probabilistic Logic Nodes* show promise as an engineering- implementable *connectionist* architecture for pattern recognition. These have the property of providing a probabilistic output depending on the state of the RAM and they show promise as Pattern Recognition devices with generalisational properties [Aleks87].

2.2.5 Multiresolutional Approaches

The *recognition cone* [Uhr72] is chosen for a discussion as it is one of the earliest models of a class of conceptual image computation models based on the hierarchical computation schemes, from which later pyramidal multiresolution models were developed. The main motivation of the recognition cones is that of the integration of both *parallel and serial* processes in a scheme called *massively parallel and shallowly serial* computation by using arrays of processors of the order of 10^6 processing elements. This contrasts with the pure cellullar array approaches where up to 100^2 processing elements have been implemented which do not readily lend themselves to perform serial operations that are required in scene understanding. The recognition cone uses a pyramidal tapering-cone with massively parallel and shallowly serial operations, resolving some of the earlier difficulties of *overly parallel* systems for image processing.

Basically, a series of layers of *transforms* are applied in parallel at each layer, which successively *extracts and co-operatively coalesces information* by *synaptic-like* operations similar to neuronal structures [Uhr72]. These transforms are *learnt* or programmed. An example given by Uhr is that of programmed PEs (processing elements) that look for local horizontal edges at the lower levels and to look for motion at the higher levels. A illustration of the recognition cone is shown in fig. 2.2.5.

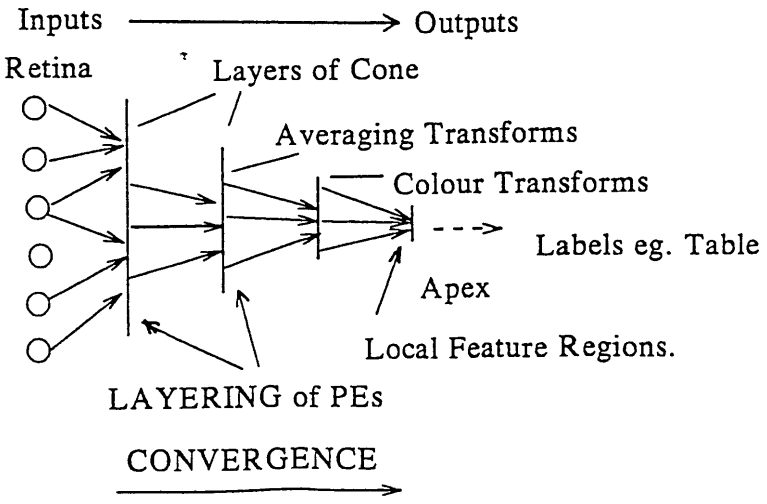
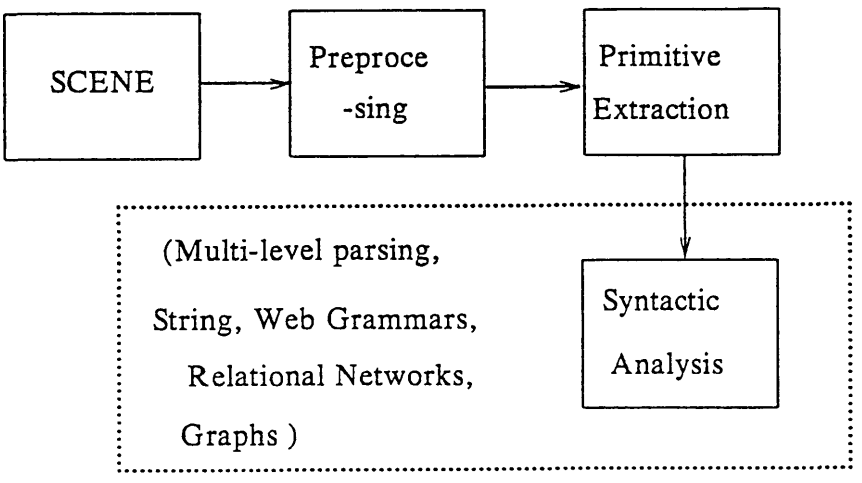


Fig 2.2.5 Illustration of the 'Layered recognition Cones'

Work on recognition cones have been mainly in conceptual and simulation models [Tani80].

2.2.6 Syntactic and Structural Approach of Fu.

No study in Image processing or robot vision is quite complete without referring to syntactic and structural methods as formulated by K.S. Fu. Syntactic approaches have been widely used in shape analysis, picture description [Shaw68] to other vision recognition problems (e.g. finger print recognition).



SYNTACTIC APPROACHES TO IMAGE PROCESSING.

Fig 2.6. A framework for Syntactic Approach to Vision.

The syntactic and structural methods arose from the more established field of formal language theory by extending the notion of alphabet, strings, grammar to image primitives. For example 'string grammars' have been used to describe complex shapes through the hierarchical chaining of embedded strings. Pattern recognition then becomes a problem of parsing the strings. More advanced methods have used *context sensitive grammars* (e.g. for closed boundary shapes) and *picture description languages*. In enabling syntactic pattern recognition to make use of the features of formal language theory, the following adaptations have been introduced:

Unlike formal languages, concatenation is seldom the the main relationship between symbols (shapes) (e.g. above, below, left of, etc) in syntactic pattern recognition.

The pattern primitives themselves need to be extracted by low-level image operations like enhancement, filtering, etc and can be quite fuzzy when compared to the symbols of language theory. A desirable criterion for the choice of the primitives include compact descriptions with little information loss. However, once the primitives are extracted, the relationship between the primitives can be make use of the power of formal language construct to solve the recognition problem (e.g. adjacency, collinearity, etc).

Fu [Fu74] points out that the main contributions of the Syntactic Methods in image processing have been in

a) the provision of a picture description language,

b) the application of high dimension grammar in the form of trees as opposed to arbitrary graphs for picture processing

c) the application of *web grammar* to more complex pictures where the picture intensity values are defined as a quadruple {n,t,p,s} of a finite set of non-terminals n, terminal symbol t, production pairs, start symbols. [Pfaltis and Rosenfeld69].

Although syntactic approaches are required to provide a formal theory on how to handle the higher level symbols, syntactical operations on these symbols are by no means trivial considering the amount of symbols that need to be processed. However an adjunct to the problem of the application of syntactic methods to robot vision is really the ability to extract consistent primitives so that coherent syntactical analysis can proceed.

2.3 Model-based Approaches

The model-based approaches of robot vision object recognition involves the matching of the input image with a set of predefined model of parts. Many of the successful industrial object recognition systems are based on model-based methods [Binford77]. Although most model- based methods use the approach illustrated in fig 2.3, they mainly differ in the way the features are extracted and how they are used in the matching process.

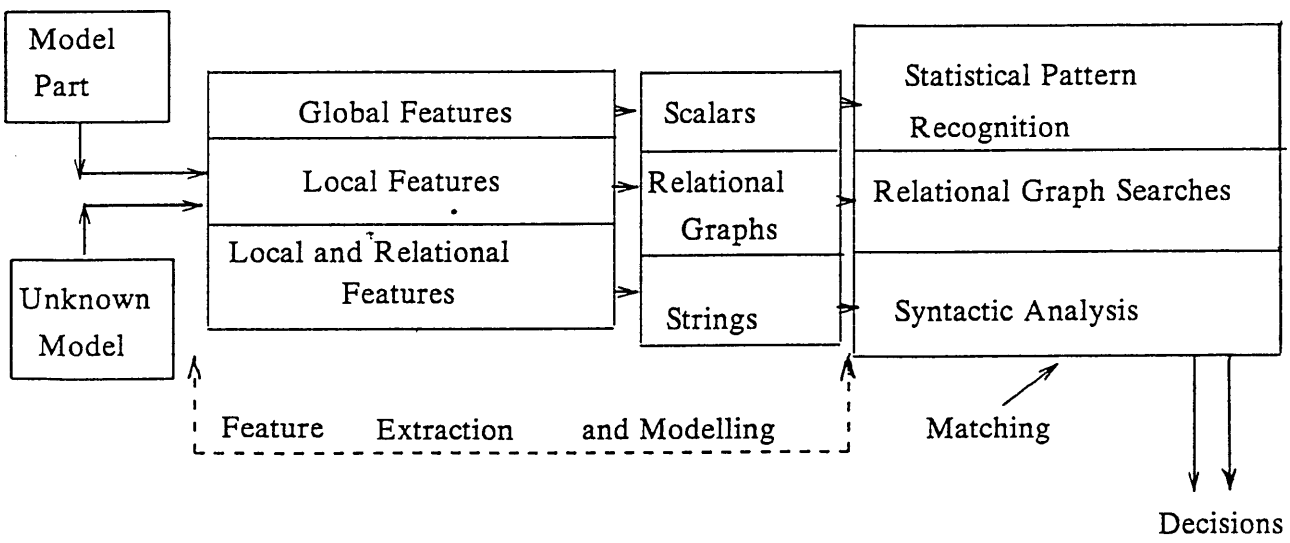


Fig2.3 Basic Operations in Model-Based Object Recognition

Broadly, there are 3 major methods in the model-based approach, namely

a) global feature method [Agin75] which makes use of the global features of the object like areas, perimeters and higher-order moments. The well-known SRI methods can be classified in this category.

b) local or structural feature methods [Perkins78] where more complex features of the object like curve segments and corners are used as features. This facilitates the use of syntactic methods to perform matching of the model.

c) the relational graph method which uses local and relational features [Yachida77] organised as graphs where the nodes describe the features and the arcs denote the relationships. This enables model- matching to be reduced as a sub-graph search where efficient algorithms are available to carry out such tasks.

2.3.1 The necessity of the Model-based approach in Industrial Robot Vision.

Due to the existence of the enormous amount of techniques proposed for Computer Vision, (image understanding, x from y, optical flow), there are many detractors as to the right solution to some very basic *Industrial Robot Vision* problems. Binford [Binford77] argues that 'no computer vision system can be expected to be so general that it can recognise objects without a model'. In fact a true dichotomy should not exist for model-based and non model-based approaches as we can always view the various techniques of computing x from y as precursors to *feature extraction* for some form of model- matching. For Industrial Robot Vision at least, the necessity of using approaches that are firmly *model-based* seem to be so absolute, at least at present, that successful schemes are nearly all model- based [Pugh83]. In chapter 6 and 7 the model-based approach will therefore be adopted in the application of multiresolutional foveal images for occluded object recognition.

2.4 Current trends in Robot Vision

Since robot vision applications have been introduced to the **industries** in the last decade, much of the concern have been in feature detection and matching for **binary** images. One trend in Robot Vision is to use more of *True grey-level* systems. A true grey level machine vision is one, not only capable of acquiring a grey-level image, but has the intended purpose of using as much of the grey-level information in as **high** a level of representation as possible. This means that **intensity thresholding** if ever necessary, is employed only at later stages of processing.

Whereas in earlier years, efforts were spent to obtain faster and better hardware (faster frame-grabbers, faster computers, subpixel accuracy determinations etc) there seems to be a greater awareness that overcoming the *ill-posedness* of machine vision should be a more desirable endeavour.

Active Vision is one such endeavour. There are a few definitions of Active Vision. One definition [Aloimonos and Bandyopadhay87] is that of having *active* observers. That is to say that the observer (camera image acquisition front end) should be allowed to move so that assumptions of smooth surfaces, reflectance characteristics need not impose restrictions that yield unstable solutions. This also enables machine vision to use the facility of *structure from motion* and other motion/velocity derived information like optic-flow computations in a better-posed and more stable environment.

Another definition of active vision is that of using *active sensing* and *multisensor fusion* [AAAI87] which uses additional sensing in the forms of active illuminations (laser ranging, structured lighting) and additional *multiple sensing systems.* In both of these definitions there are no preclusions from using binocular vision or other correspondence-derived computations.

It may be necessary to point out that there are also other actively pursued ways of finding possible solutions that extend beyond machine vision to areas of visual perception, through biologically and neurophysiologically-motivated investigations (neural computing, visual cortical receptive fields, biological retina). One such example, relevant to this thesis is that of the *Generalised Gabor Scheme of Image Representation* [Porat and Zeevi88] where a image for machine vision can be represented as a set of 2-D Gabor elementary functions, using position-dependent sampling and having some of the characteristics of the *multiresolutional* structure of the vertebrate retina.

2.5 Summary

The chapter has been organised into two main parts namely the first part on the theory of image processing that may lead to the longer range goal of image understanding and the second part describes the more immediate goal of robot vision in industrial environments emphasising on the need for model-based approaches. The model-based approach has been given this emphasis because in later chapters, the application of the multiresolutional image for the robot vision task of occluded object recognition and semantic-object segmentation shall rely on model-based techniques.

The contributions of Barrow and Tennenbaum have been described in terms of the generation of intrinsic images for all points of the image where subsequent object recognition is treated as the process of obtaining and matching as much of 3-D or 2.5 D information of the scene. Unlike the methods of Marr which incorporates the theory of stereo perception and texture to obtain depth information, active methods (e.g. laser ranging) have not been ruled out by Barrow as means to obtain the intrinsic images.

Marrs contribution is in a prescription for a computational theory with the right representations at all levels of image processing so that these representations will be uniquely recoverable.

Uhr's works have been instrumental in providing the impetus for the consideration of massively parallel and shallowly serial (the *100 step limitation*) using conceptual processors organised as conic

transforms. These ideas have led to the pyramidal image representations for robot vision using multiresolutional models.

The Syntactic Approaches of Fu have provided formal methods of using the established theory of languages in the field of 2-D picture analysis.

Aleksander's contributions have been in the engineering implementation of neural-like computational elements for adaptive pattern recognition and robot vision using Random Access Memory elements which are seen as good candidates for implementing neuronal-like networks with emergent properties.

The Model-based approaches have been discussed in the context that most successful industrial robot vision especially for object recognition have been based on some methods of model-based computation.

The discussions on current trends indicate that much effort is now being made to use multisensor-fusion active vision and stereo imaging to remove any ill-posedness as a solution to immediate robot vision problems.

Chapter 3

Computer Representation of Images for Object Recognition- Hierarchical and Non-Hierarchical.

This chapter surveys the ways in which an image can be represented as a computer data structure for the main purpose of object recognition. In this context, and following Rosenfeld [Rosen83], those data structures that assign uniform weightages to every pixel in the image as *non-hierarchical or traditional* and those that assign different weightages to pixels in different areas of the image as *non-traditional or hierarchical*. The relative strengths and limitations of these representations in relation to applications for object recognition will be discussed.

It will be noted that where an image is already segmented, and transformed into bi-level images, there is a wealth of shape representations that are both compact and elegant. Invariant object recognition of such planar shapes involve a further transformation of the shape into a domain for which the parameter of variance (e.g orientation) is non-sensitive to the recognition process. Such examples are moment invariance [Hsu70] and Fourier Descriptors. However, there are much fewer elegant ways of image data representation for images that are not segmented where invariant object recognition then becomes a more difficult process.

3.1 Relationship between Image Representation and Computer Data Structures.

Any image representation chosen will bear a relationship with the consequent data structure that the computer vision system needs to handle. Although strictly speaking Computer Data structure, image representation and image coding can be different entities, most literature cannot help but link these together in the discussion on image representation. One of the features of a good image representation is the capability of representing the digital image in the most compact form (hence the relationship with image coding). Although compactness of data usually leads to better data storage and faster manipulation, for the purpose of object recognition, compact data structures may not always provide robust object recognition as such representations may not be information preserving.

There are alternative ways of classifying image representation. As an example, image representation can be classified by the method in which the data is manipulated (e.g. graphs, trees, linked lists, matrices, orthogonal transforms etc), but with the increasing use of specialised hardware for computer vision systems, like pipeline, systolic, pyramidal processor, there is much overlap in possible classifications.

3.2 Hierarchical and Non-Hierarchical Image Representations.

Following [Rosen83], we shall now classify image representations into the hierarchical and non-hierarchical types.

If we represent an MxM image I(x,y) in terms of its pixel intensity i(x,y) and weights associated with that pixel as w(x,y), i.e.

$$I_{(x,y)} = \sum_{x,y=1}^{x,y=M} \left(w_{(x,y)} i_{(x,y)} \right)$$

then we say a hierarchical image has non-uniform weights whereas a non-hierarchical image has uniform weights. The

In hierarchical image representations, the Image $I_{(x,y)}$ has different weights $W_{(x,y)}$. The different weights are related to the hierarchy, which can be expressed in terms of the resolution, size of pixel or other metrics. Thus in multiresolutional pyramidal images, a pixel at the higher pyramid layer (lower resolution) has a higher weightage than one at a lower pyramid layer (higher resolution). This lends itself well to computational algorithms using hierarchical search trees and graph-traversals. **A Summary of Hierarchical and Non-hierarchical image representation is shown in fig 3.2**

3.3 Non-Herarchical Representations.

Usually, early processing attempts to segment the interesting parts of the image from the background and then generate a compact representation of the segmented object using the closed boundaries of the object.

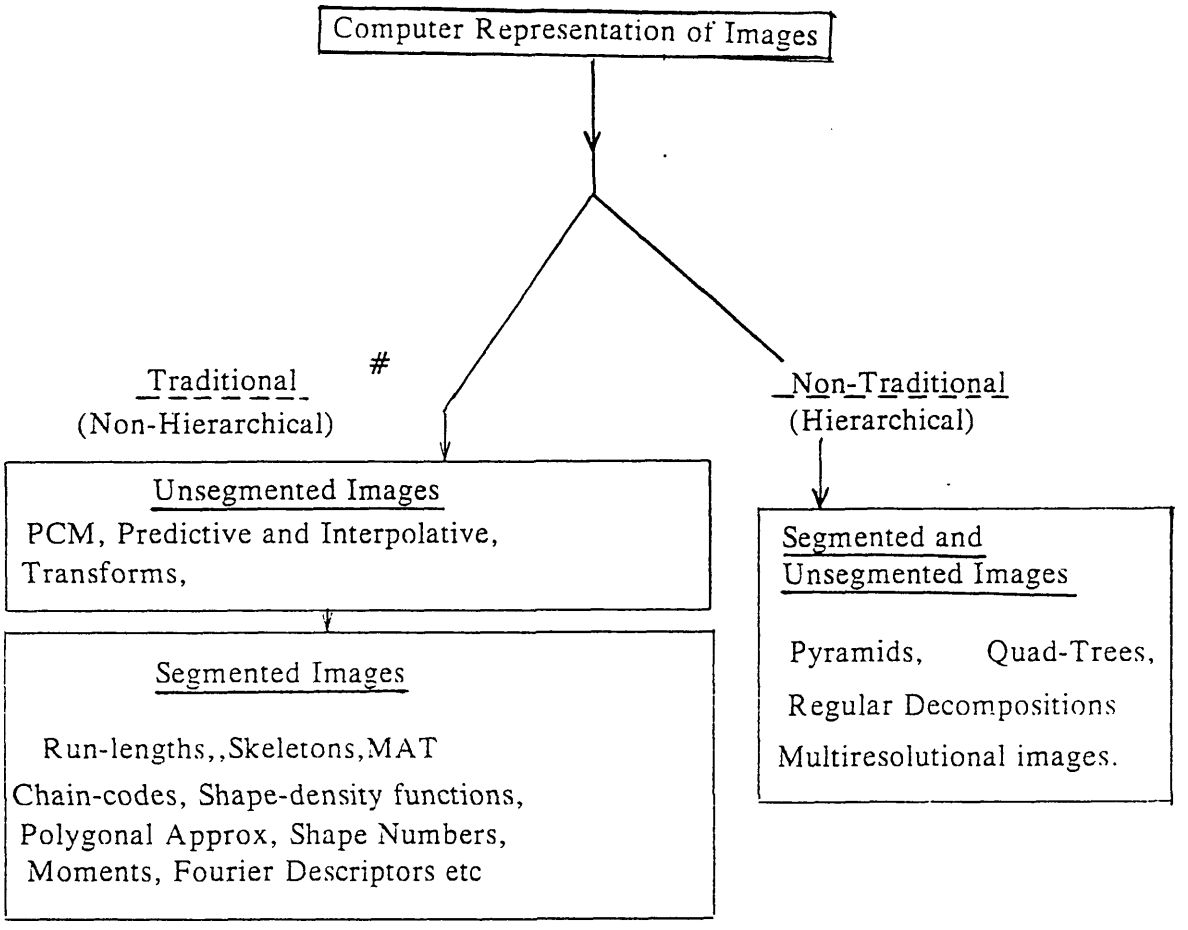
If the image is not easily segmented, usually due to distributions of grey level histograms that are not amenable to the various intensity thresholding methods (simple or dynamic thresholds), then edge detection methods [Marr-Hildreth, Roberts, Sobel, Kirsch etc] can be used to detect **intensity-discontinuities**, to be followed by global thresholding, connectivity analysis, thinning and skeletonisation methods to obtain a compact representation of the segmented object.

For images that are not easily segmented by the two methods described above, further AI motivated pixel tracking methods may be employed, though they are presently not as widely used in industry.

3.3.1 Non-Herarchical representations of segmented images.

Segmented images are usually bi-level either as a result of some thresholding processes on the original grey image or in some cases, special image acquisition front-ends like OpticRAM or *pre-engineered* structured lighting[1] (e.g silhouette lighting) are used to generate the binary image. Objects can be represented by its *boundary* description where the

[1] Pre-engineering is a common term used in industrial robot vision where the sceneis specially prepared for object recognition by methods of structured lighting, silhouetting, delivery of part from hopper to ensure orientations etc.



This classification follows that of [Rosen/Tani] to emphasis the newer purposeful hierarchical representations of pyramids etc.

Fig. 3.2 A Summary of Hierarchical and Non-hierarchical Image Representation.

image is first decomposed into regions of constant grey levels or regions in which the grey levels are considered sufficiently similar to be classified as belonging to the same regions. If there are 2^k grey levels in an image of 512 by 512 pixels, and assuming there are 100 distinct regions, the boundary representation will need 20,000 + k + 1800 bits to completely specify the image (with provisions for indexing, etc) which would otherwise require 512 x 512 x 2^k bits. Provided the boundaried regions are also consistent boundaries of the object, such a representation can be used for object recognition.

Where the image cannot be represented by disjunctive object boundaries, the *run-length* of the pixels can be used, where each run-length is the maximum length of some uniform grey level.

As an example, if a 512 x 512 image has 2^k grey levels then each run requires the specification by k+9 bits for the k grey levels and the length of the run. If there are 20 runs per row, we would need 20(k+9) bits compared to 512 bits.

Bilevel segmented images can be further made more compact for the purpose of object recognition by skeletonisation methods. Various skeletonisation methods have been proposed and they include the medial axis transforms, MAT [Blums67], and various thinning methods, the general aim of skeletonisation being to represent the shape by a smaller set of pixels.

Although skeletonal representation seems intuitively elegant, it is rather computationally intensive as it involves making a calculation of the distance from every interior point to every point in the boundary of the region, as in the case of the MAT and are sometimes not information preserving. Again, remembering that we are assuming a binary, homogenous-region object, (i.e. possibly a single object in the image), we could see that this representation is only used in very apt situations of object recognition where planar shapes are sufficient basis for recognition. *Polygonal* approximations [Pavlidis73] can be used to represent segmented shapes using polygons of varying segment lengths for its boundaries. For a closed curve, the approximation can be exact if the number of segments is the same as the number of points in the boundary. Polygonal approximation computation is really not trivial although the principle is simple because the problem can turn out to be a long iterative search with many non-unique solutions. Further more compact representation of an object shape have been suggested by Sklansky [Sklansky72] using the principle of the minimum perimeter fitting of the shape resulting only in very low fitting errors.

Other polygonal approximation uses *merging methods* to compute the minimisation of a LMS error fit between the approximations and the boundary by further *merge and split* methods which normally takes more time than the simpler methods.

Projection Methods can be used to reduce the 2-D description of the image into 1-D,

thus providing a simpler and faster data structure for object recognition. The general method is to project a boundary description of the shape to generate a linear signature which is sufficient to represent the object in the image ,though such signatures must be normalised for the starting point and size of the bounded area. Subsequently simple *chain-codes and shape numbers* can be used to encode the boundary by a set of line-segments of specific lengths and directions in a 4 or 8-connectivity grid but they have to be normalised with respect to the starting point and position of the object in the image. To overcome the problem of object size and position, *Shape-density functions* can be used. One method due to Nahin [Nahin74] uses a *slope density function* as a signature where the shape density function defines a histogram of the tangential angle values of the boundary.

Signatures can be transformed into equivalent *moments* corresponding to the variance and mean of the signatures, treated as a discrete random amplitude variable. This method is contrasted with the more familiar 2-D geometric moments of Hsu [Hsu70] which is commonly used in many object recognition applications. The method of geometric moments of object representation has some advantages of computation speed as moments can be computed quite straightforwardly. For the purpose of object recognition, it also enables a controllable degree of precision in recognition since only sufficient combinations of the moments need be used. This method also provides a simple way of obtaining geometrically invariant shape recognition and is the principal method in the more successful general-purpose object recognition modules of SRI [Agin75].

Fu [Fu82] described a projection method for shape representation using *Fourier Descriptors* where the boundary of the object, expressed as a sequence of 1-D complex numbers $(x+jy)$ is used to compute the Fourier Transform of the shape. The resulting Fourier Descriptors in the form of a sufficient combination of coefficients can be used for object recognition. As in the method of moments, suitable combinations of the descriptor coefficients can be used to uniquely describe the object under various linear geometric transformations.

3.3.2 Non-Herarchical Representations of Unsegmented Images.

Image representation for unsegmented images is more problematic. In coding theory this is related to the nature of the image that has not been *decorrelated.* Athough segmentation of an image into regions, objects etc is one of the early processes in conventional robot vision, so that subsequent representation of the segmented region of interest reduces the dimensionality of the data to be processed, there are many instances in which it is impossible or difficult to provide a semantically consistent segmentation [Marr82].

There also exists a great number of schemes to represent unsegmented stationary images that include schemes based on source encoding, variations of Pulse Code Modulation

(PCM), Predictive Coding (PC), Interpolative Coding (IP), maximal block coding (MBC) [Kunt80] and Transform Coding, a more detailed treatment being in [Huang79] and [IEEE72]. PCM in its simplest form encodes quantised levels of the image intensity for digital representation and transmission whilst the variation of Delta PCM (DPCM) encodes the quantised levels in a more efficient way using statistical dependencies of previously represented samples. In PC, linear or non-linear predictors use several previous samples in variations of adaptive methods where the causal prediction[3] of the pixel at location k, l is

$$I_k^p = \sum_{f=1}^{p} a_i I_{k-i}$$

where I_p is the predicted sample from past samples of I_{k-i}.

In IP, subsets of the pixels are used as the basis for the interpolative recovery of the remaining samples, with adaptations for minimum-aliasing error, depending on the rate of sample changes. MBC techniques use the 'maximum block' size of some *similarity measures* of the pixels to assign an optimum code-word, thus providing better image data compression and manipulation. The *Transform* representation of images has the principle motivation of transforming a set of data into another set of *less correlated* coefficients that can be viewed as matrix-vector multiplications, of which the *Karhunen-Loeve, Hotelling transforms* do provide optimal non-correlated coefficients. Other transform representation of non-segmented images using frequency-domain or spatial domain methods like the Haar, Hardmard, variations of Fast Fourier Transforms, Discrete Cosine Transforms are available and in some restrictive situations, can be adapted for object recognition.

3.4. Hierarchical representations of images.

As pointed out earlier, non-traditional or hierachical images are images where each pixel in the *hierarchy* of images have different weightages (or importance) according to its hierarchical level. Thus in multiresolutional pyramidal images, a pixel at the highest pyramid layer has a higher weightage than one at the lower level. This lends itself in computational structures to efficient search trees and graphs through hierarchical graph-traversal. It may be pertinent to mention that multiresolutional images have been studied in various forms with different emphasis, making it sometimes as though they are disjointed endeavours in image processing. They are studied under different names with *different emphasis* such as *recognition cones* [Uh72], Gaussian Pyramids [Tani80], Quad-tree images [Hunter70], Laplacian Pyramids [Burt86], Regular Decomposition Models [Klinger76].

[3] Causal prediction uses past pixels (i.e in previous lines and columns)

They however, have the common feature that *image* information is *summarised* as the resolution decreases.

Two examples of the *different* emphasis hierarchical images need to be pointed out:-

a) *hierarchical and pyramidal computation structures* for image processing.

b) *Regular Decomposition* image algorithms.

These hierarchical image representations have generated recent interest because of the following *possibilities*.

a) fast processing capability if *suitable* hardware is available,

b) the availability of image features at a wide range of resolution or scales enable processing not unlike that of humans

c) the ability to condense more global information into local information.

3.4.1 The Canonical Image Pyramid.

The canonical image pyramid is made up of N+1 planes of dimensions $2^n x 2^n$, stacked on top of one another, with the base plane of 2^N x 2^N containing the original resolution and the apex plane 2^0 x 2^0, being a single pixel. If the level is defined by L=P-n then each plane $f_L(x,y)$ is as shown in fig. 3.4.1 The planes are related to each other through a *consolidation* process using a *generating kernel* to map between planes. For any arbitrary level K, the plane $f_{k+1}(x,y)$ is derived from $f_k(x,y)$ using a selection of mapping functions Π where

$$f_{k+1}(p_o) = \Pi [f_k(p1), f_k(p2), f_k(p3), f_k(p4)]$$

The pyramidal planes so obtained, are subsequently used for image computations.

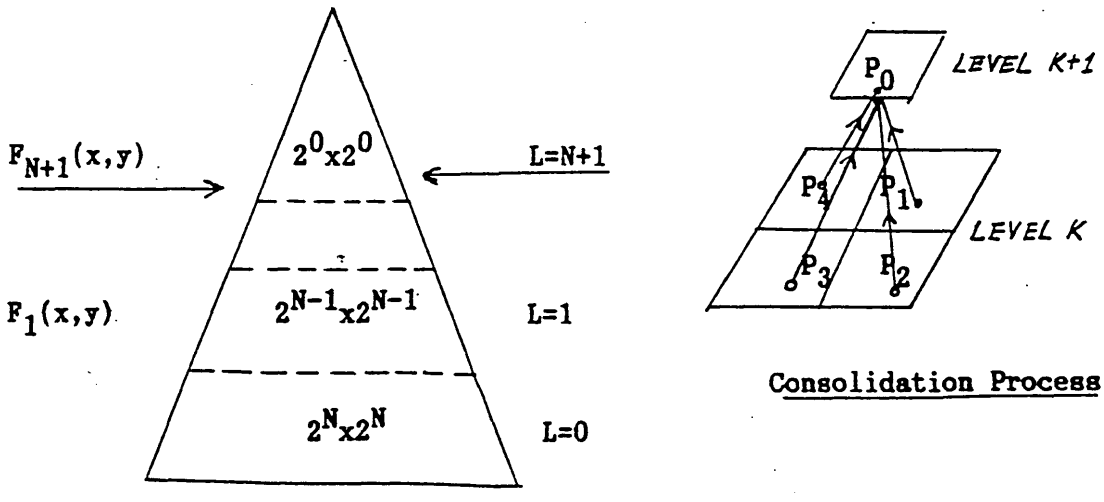


Fig. 3.4.1 The Canonical Pyramid and its consolidation Process. (a.b)

3.4.1.1 The Basic M-Pyramid

The most basic form of pyramid image representation is that using *M-pyramids* [Tani80]

An M-pyramid or a matrix-sequence pyramid P is a sequence {M(L), M(L-1)...M(0)} of arrays where

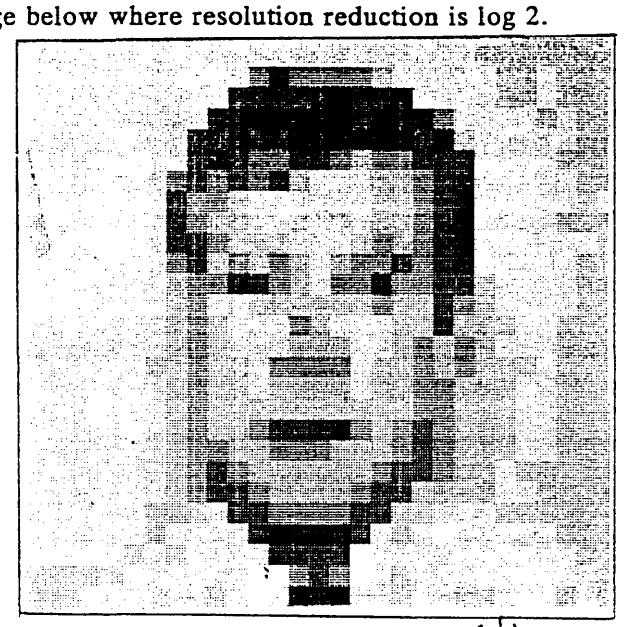
M(L) represents the original image,

M(i-1) represents a version of M(i) at a reduced resolution (usually half resolution), proceeding to M(0) which represents a single pixel.

Each reduced resolution image is obtained by subsampling the original image without generating any kernel function. This is shown for an typical image below where resolution reduction is log 2.

(c)

Fig. 3.4.1 c) A Gaussian Pyramid Image a plane 3



(d)

Fig. 3.4.1 d) A M-pyramid image at plane 3

3.4.2 The Basic T-Pyramid.

The *T-pyramid* [Tani80] describes the multiresolution image as a tree rather than as a matrix as in the M-pyramid. This allows an algorithm to traverse a tree according to the need for coarse or *refined* image information and thus provides faster processing times through a hierarchical search.

Formally, the T-pyramid is defined by :

1. a set of p-nodes $P=\{(k,i,j)\}$ for $0<=k<=L$ and $0<=i<=2^{L-1}$ and $0<=j<=2^{L-1}$
2. a binary relation F on P with $F(k,i,j)=(k-1,[i/2],[j/2])$
3. A kernel generating function binding V to P

Hence the p-nodes corresponds to the pixels of the M-pyramid and the complete set of p-nodes at the levels{k} represents the image of the T-pyramid at level {k}. F is defined as the parent function defined for all levels except for the root-node at {0,0,0}. Thus every p-node q except at the level{L} has 4 sons which are those p-nodes q' with $F(q')=q$.

For a generating kernel of a 'simple averaging' function, the T-pyramid can be concisely described as

$$V(k,i,j)=0.25\sum_{x=0,1}^{y=0,1} V(k=1,2i+x,2j+y)$$

for $k<L$

and $V(k,i,j)=A(i,j)$ for $k=L$

where A(i,j) represents the original image.

Similar equations can be obtained for 'sampling' generating kernels.

3.4.2 Computation in Pyramids.

Once an image is available in mulitresolutions, it would be possible to use the right resolution level to obtain features which are known to be most easily available at that desired resolution (eg. edge elements *edgels* as opposed to *edge points*). Further, in processing at the right resolution level we are also using less computation time as the total pixels in a reduced-resolution image is **reduced**

3.4.3 Pyramidal image applications:

The multiresolutional pyramid can be used to search for a pixel or image region of interest (ROI) by a hierarchical procedure [Tani80] by starting at the root of the pyramid and at each node, visit the sons which are non-zero until it terminates in level L. The spot coordinates of the ROI is then obtained from the search path in time proportional to the height of the pyramid.

Pyramidal *edge detection* may be recursively done as a tree traversal with searches needing to be done for those areas showing promise at the coarser image.

Using the quad-tree representation, the medial axis transform (MAT) is the set of points serving as centers of *maximal blocks* with their corresponding radii. In pyramidal *region analysis* the image is successively subdivided into quadrants until a block (possibly a single pixel in the case of unsegmented images) is reached that consists of 1s or 0s. This process is represented by a tree of degree

4 (i.e each non-leaf node has four sons) in which the root node represents the entire image.

Other pyramidal image operations for robot vision task of *connected-component labelling* is achieved by scanning the pixels not in a row by row basis as in traditional images, but by a sequential scan dictated by the tree structure in a *postorder traversal* in which updates and merging of labels are performed using metrics of *adjacency* that results in execution time proportional M log M where M is the number of pixels. Other features like *Euler number, perimeters, areas and moments* have similarly better execution times compared to those with *traditional* image representations.

3.4.4 Pyramidal Images as efficient image codes.

Proposals have been made to use the pyramidal hierarchy for *Image Storage and Transmission* [Tanii79] for both T-pyramids and M-pyramids.

For an image of n x n pixels, the M-pyramid requires a storage capacity equal to

$$n^2(1+1/4+1/16+\ ..) = 4n^{(2/3)/d} \text{ pixels.}$$

For T-pyramids, the data in the arcs may also need to be stored but compact storage can still be achieved by observing that the averaged nodes at various levels require a variable number of bits to encode the pixel intensities, e.g.

Level 0: k+2L bits per pixel

Level L-2:k+4 bits per pixel

Level L: k bits per pixel

It also provides an advantage of sending the coarser, low resolution image level first so that the recieving end can decide to continue or abandon the image without waiting for the fine-resolution image to be completely transmitted[Tani79].

Limitations of Hierarchical Pyramidal Image Representations.

Although the pyramidal structure for image representation provides some of the advantages described above, most of the work are done in simulations only as real pyramidal computers are still in development [Tani80].

The reduction of the resolution by log 2 can be at times too drastic which consequently has led to proposals of *overlapping pyramids* [Ferretti86], where pixels of intermediate resolutions can be generated. Many of the pyramidal algorithms do assume some 'a priori' knowledge of the image so that algorithms for search can start at that resolution. The time required to generate the pyramid using conventional computers is non-trivial, this having been used as a criticism of pyramidal approaches in image processing.

3.5 Other Hierarchical image representations.

Two proposals that have features of multiresolutional images need to be mentioned. These are

a) the logarithmic grid image representation of Weiman [Weiman79] in which a digitised image is represented by a more densely packed pixel distribution at the foveal area than at the periphery.

b) the *NASA* image in which pixels at the periphery are represented by a samples of the neighbourhood [Pratt78].

3.6 Summary.

The review of computer representations of image data in this chapter has attempted to make a distinction between the non-hierarchical and the hierarchical images, following the definitions of Rosenfeld and Tanimoto. This distinction has been chosen in order to relate them to the multiresolutional foveal image representation in this thesis which shares many features with the multiresolutional models. The importance of segmented image representations and the various methods available for planar shape representations in robot object recognition were described. For images that are not readily segmented, various methods, more in the domain of image coding structures were described, with reference to the need for object recognition. The various hierarchical data structures for images using quad-tree or pyramid descriptions were given more treatment in their techniques for image searches, image feature extraction, etc. Some of the limitations of hierarchical image representations were described.

Chapter 4

The Multiresolutional Fovea-Centred Image Representation.

The Multiresolutional Fovea-Centred Image Representation (MF image for short) attempts to represent an image of a scene as it would be seen by the human eye. Only the area in the retina corresponding to the fovea pit, subtending an angle of 1.5 degrees, has the highest resolution. The resolution outside the fovea decreases gradually till the limit of peripheral vision. Such a reduction in resolution being due to combinations of factors that include the distribution of photoreceptors in the retina, the optics of the eye and the visual information-transmission characteristics of the the visual pathways.

This chapter attempts to gather the more recent trends in image processing to make a case for the need of such a MF image for robot vision. It will briefly describe the pertinent aspects of the Human Visual System (HVS) in the light of the psychophysics of vision, the physiology of the eye, the neuronal processes from the retina to the visual cortex and the psychology of object perception. It will draw attention to the the experimental confirmations of the existence of spatial-bandpass channels in human vision [Robson68], the arguments of Fischler [Fisch78] for *partial segmentation* and Ballard [Ball87] for a need to rethink some present day approaches to computer vision.

The proposed structure of such a MF image, possessing certain features, making it more harmonious with the arguments above will be described. Formal definitions and comparisons with pyramidal multiresolutional systems will be made. The choice of the optimal median generating kernel is established.

4.1 The Human Visual System.

Visual Perception in humans must be sufficiently complex and efficient to make it one of the most advanced systems in creation. However, looking at its performance from other aspects, it has deficiencies and peculiarities making it seemingly not the best of Nature's design. Our contrast sensitivity, which is the ability to detect changes in light intensity dI, as a function of intensity I, denoted as the Weber Fraction dI/I, is linear about 0.2 and becomes non-linear outside this range as shown in fig 4.1a and b. This is compounded by the further logarithmic dependence of the Weber fraction on the surrounding Intensity Is, as shown leading to an overall choice in some image processing systems on performing operations on the logarithm of the intensity rather than on the absolute intensity. The uneven sensitivity of the eye to different spatial frequencies[1], being relatively insensitive to low and high spatial frequencies and more sensitive to mid spatial frequencies, gives rise to Mach Band effect. This effect leads us to think that since edges in images have high spatial frequencies, they could be processed with less precision. The human visual system also exhibits the 'Simultaneous-Contrast

[1] Spatial frequencies are frequencies of a signal expressed in the space domain (x,y) as opposed to say the time domain.

Fig. 4.1a Contrast Sensitivity functions

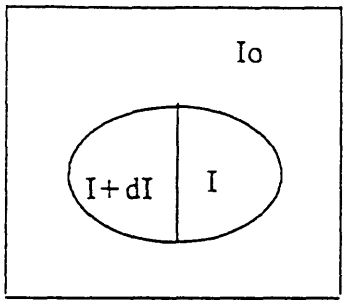


Fig. 4.1b Non-linearity of Weber Fraction

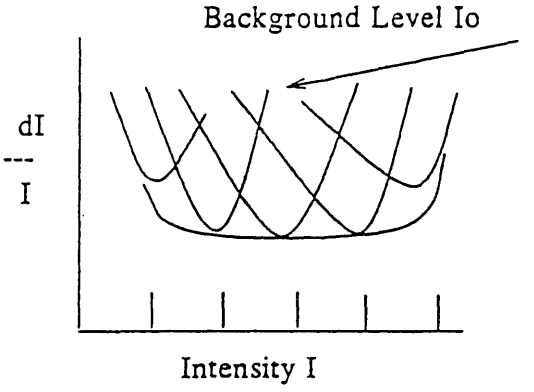
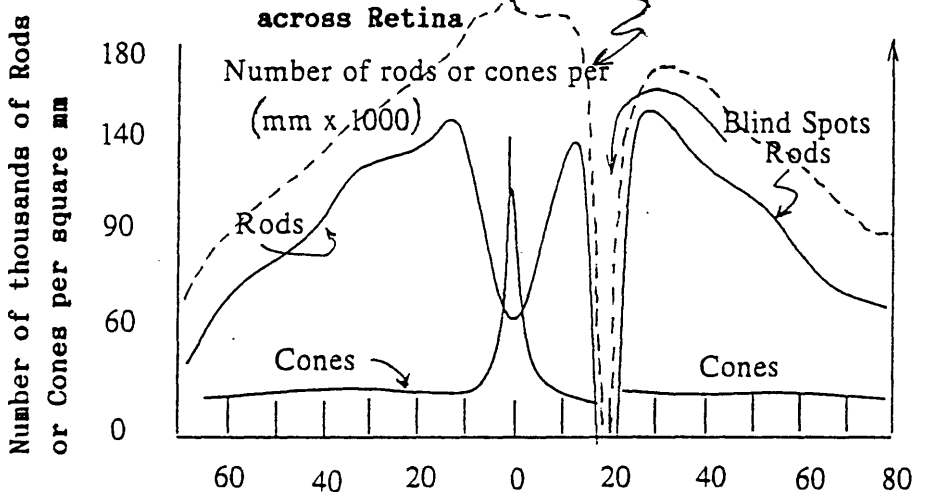


Fig. 4.2.1a Extrapolated distribution of Rods and Cones across Retina



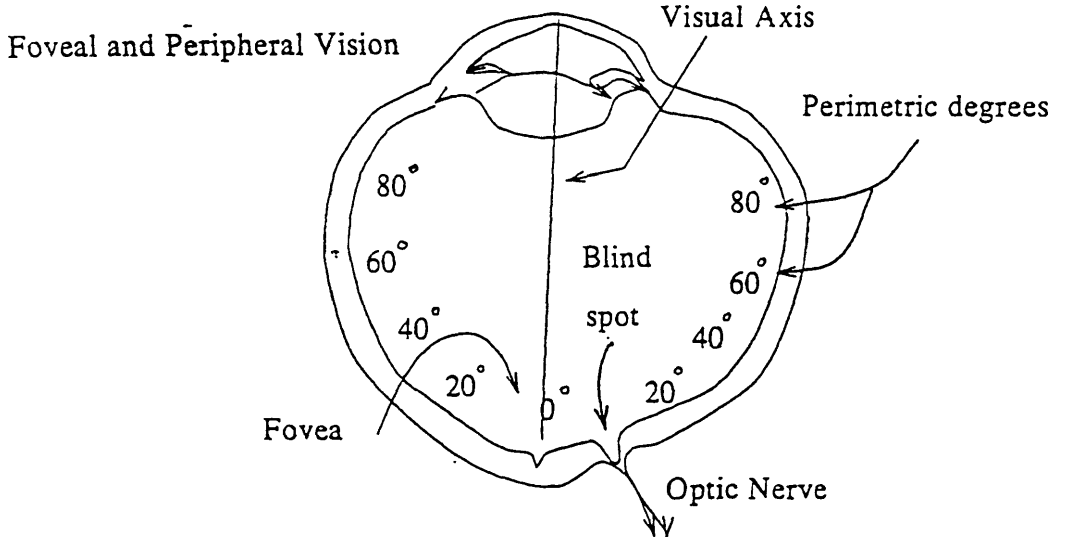


Fig. 4 Some Features of Biological Vision related to Multiresolutional Foveal Image Representation.

Perception Errors' [Cornsweet70] and other seemingly anomalous colour-related visual characteristics like *Chromatic Adaptation, Lands Colour Perception effects,* etc that present day knowledge is confounded as to how best to apply them for robot vision.

4.2 Relationship between Multiresolutional foveal images and the HVS.

The main motivations for MF images for robot vision have been drawn from the physiology of the eye, the neuronal processing in the visual pathways and the observed psychological processes in object perception which together show the existence of hierarchical, parallel and distributed processes, that together enable the HVS to handle visual perception at such high speeds.

4.2.1 Physiological and Neural Parallels.

Although the physiological and neural processes seem to be closely linked in the HVS, it is widely accepted that while much of the physiology of the visual system (structure of eye and nervous pathways) is a relatively well established science from pathological studies [Dowling87], the mechanism of computing or neuronal activity (e.g. activity of visual cells and their receptive fields) have only been better elucidated in recent years [Hubel62]. To simplify this discussion, the physiological visual system and the neuronal visual system shall be discussed separately.

a) Physiological Processes.

The physiology of the eye has been widely studied [Davson49]. Incident light on the eye's lens is sampled by an array of approximately 107 million photodetectors distributed as 100 million rods (intensity sensitive detectors) and 7 million cones (colour sensitive detectors) across the retina with their distribution in packing density shown in fig 4.2.1a. The rods are principally found at the peripheral area of the retina and the cones, mainly at the specialised area of about 0.55 mm across the retina near the visual axis, called the fovea. The density of these rods and cones are not uniform across the retina, showing a highest packing density of 150,000 photodetectors at the fovea and decreasing away towards the peripheral area of the retina up to a perimetric angle of about +/- 80 degrees. The distribution of *both rods and cones* (scotopic and photopic vision) is extrapolated as a dashed line in fig 4.2.1a showing that the composite photodetectors build up to a peak at the fovea and decreases away towards the perifovea.

Beyond the retina, as in fig 4.2.1b, the photodetectors are connected to the optic nerves via specialised cells called the bipolar, horizontal, amacrine and ganglion cells. The output from the ganglion cell is not just a function of a single photodetector but a region of photodetectors in the receptive field[2] of the retina.

[2] The Receptive Field (RF) of a visual cell is an area in the retina where visual stimuli affect that cell.

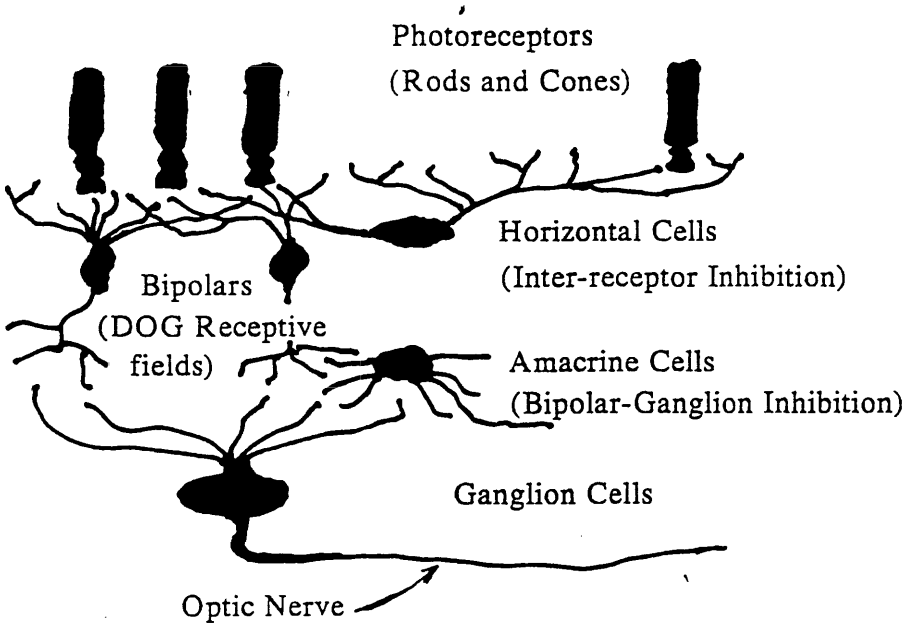
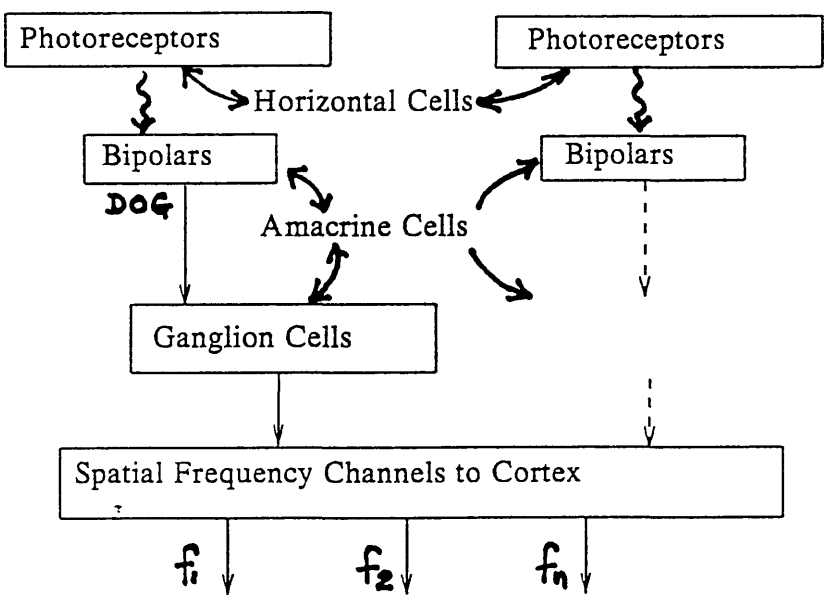


Fig. 4.2.1b Neuronal Processing in Vertebrate Retina



Simplified Block Diagram of Neuronal Processing

Fig. 4.2.1d Illustration of the generation of Spatial Frequency Channel in the Visual Processing

This complex interconnection reduces the dimensionality of the visual information by some early visual excitory and inhibitory processes (on-centre, off-surround) of the ganglion cells which, through more complex interactions with the other cells produce detection of shape and orientations [Hubel62]. This visual information is then passed to the Optic Chiasm where the right eye information is available to the left brain and the information from the left eye, to the right brain.

This leads to the lateral geniculate nucleus (LGN), where optical nerve fibres containing similar visual-field information are terminated near to one another. Although little information is processed here, the arrangement at the LGN is supposed to aid in the binocular perception of the image on reaching the visual cortex. At the visual cortex, the simple cells and complex cells, which are complex interconnections of ganglion cells, detect more complex orientations, shapes and other stimuli including motion. The signals from each eye are combined at this stage and cortical cells perform higher level processes on different classes of stimuli involving specific sizes and shapes. [Sakrison77].

b) Neuronal Processes.

A growing interest in *Connectionist* or *parallel distributed processing* is developing amongst the community of neural biologists, computer scientists, psychologists and engineers [Aleks89, Rumelhart87]. The processing at neuronal levels have provided many motivations to research in pattern recognition and machine vision. A summary of such neuronal level processing may be appropriate in order to appreciate any relevance to multiresolutional images of the human visual system.

The retina consists of 5 main types of cells illustrated in fig 4.2.1b.

Light stimuli are carried by the photoreceptors as membrane voltage variations proportional to the logarithm of the light intensity, to the bipolars and thence to the Ganglion Cells. Two interneuron cell-types, the horizontal cells mediate lateral interactions between the photoreceptors and the bipolars in the inner plexiform layer, whilst the Amacrine Cells mediate between the bipolars and the Ganglion Cells in the inner plexiform layer. The horizontal cells are thought to be a light sensitive resistive layer that aids in providing the dynamic range for the eye that far exceeds the performance of individual neurons. At the bipolar level, a difference of Gaussians (DOG) receptive field is created by photoreceptor-bipolar interactions providing the excitory centres. The horizontal cell-bipolar interactions provide the inhibitory peripheral response. The output of the Ganglion Cells is believed to be the earliest instance of directional-sensitive motion detection.

The outputs of the Ganglion Cells are carried by the axons which serve as an optic nerve, to the Visual Cortex. Beyond the Human Visual Cortex, much less is exactly known in terms

of neuronal-level processing.

4.2.2 Psychological Studies of Human Object Perception.

In the field of computer vision especially, psychological and psychovisual theories seem to be very overpowering, that pure engineering of vision systems without regard to such theories seem to be foolish. An attempt will be made here to point out certain principle differences between the way in which object recognition is performed by man and the way in which most current robot vision systems perform the same task. This should lead to some justifications for the the need of multiresolutional foveal images in robot vision.

Put simply, robot object recognition usually begins with partitioning of the image into *regions* or *primitive objects* (segmentation), characterised by a set of attributes, and using these attributes together with an a- priori set of inter-object relations (predicates and relations e.g. left of, etc) to provide a *label* of the object. Many of the sucessfull object recognition systems [Perkins78, Agin'75] in industry tend to use more fundamental techniques than is available in vision research, justifiably due to the requirement for robustness and speed.

They usually make use of a well-contrasted background and if possible, by simple thresholding, obtain the object of interest by performing hardware assisted connectivity and edge linking analysis to obtain the *connected-components*. Usually, the computed moments serve as features for object recognition [Agin'75].

This procedure differs from the human object recognition model in two main ways in the context of visually complex scenes (one extreme example being an outdoor scene with trees):-

a. Partial and complete Segmentation.

The term complete segmentation indicates an operation in which the scene is decomposed into disjoint regions with each object-region being given a distinct label as can be done in a preengineered conveyer belt of industrial parts (i.e. region of part and region of background). Partial segmentation means that the segmentation process need not finally produce closed regions. Partial Segmentation allows the incomplete segmentation of entities so that unclosed boundaries need not be forced to create an *unnatural* object. It is unlikely that an image can be neatly described by a *finite* set of discrete object entities. This implies that if complete segmentation of the leaves of the trees are required, the number of leaves and their positional interrelationships will not only be too many [Fisch78], but forced complete segmentation will also likely lead to semantically incoherent objects [Marr82].

b. The impracticability of a complete set of predicates.

The central theme in this argument by Fischler [Fisch78] is that due to the *non-isomorphic* way in which information is logically encoded in the conventional computer, it is impractical to have a complete set of predicates to describe the completely segmented objects in the scene. Non-isomorphism in information coding is described as the need to specifically encode *all* information and rules for unification e.g. if a predicate *left-of* is encoded, there is also a need to encode *'right-of'* as there is no way the complementary rule is automatically known by the computer. These arguments lead us to consider that partial segmentation would be a more natural consequence of the *global preattentive and local focused-attentive phases* in object recognition. Such a procedure is used in occluded object recognition in chapter 7

4.3 Spatial Frequency theory of HVS.

The *Spatial Frequency theory* of the HVS is a alternative theory of the HVS that seems to be gaining more support in recent years, compared to the *Feature Detector theory* which, before the 1970s was more more widely accepted. Psychologists have been puzzled by experiments that indicate that the HVS seem to exhibit some primitive Fourier Analysis [Sullivan82]. In fact it was only as late as two decades or more ago that an increasing use of patterned stimuli in the form of striped light, complex gratings etc have been seen in psychovisual experiments [Georgeson78]. The existence of *Separate Spatial Channels* in the human visual systems have been studied in [Robson68, Sachs71, Georgeson75].

Sources trace Robson as amongst the first to draw attention to the *existence of separate independently-tuned narrow spatial channels* Sachs confirmed the existence of these channels in the human visual systems by experimentally obtaining the psychometric functions using concurrent, simple and complex gratings for 2 frequencies centred on 14 cycles/deg, selectively tuned to a small range of spatial frequencies. Georgeson and Sullivan conducted experiments using high and low frequency gratings with 20% contrast that showed that the HVS was unable (within limits) to perceived the loss in contrast leading to their assumption that there exists narrow band-pass filters in the HVS.

In [Baker and Sullivan80], *multiple bandpass filtering* was used as universal *optimal filters* for computer tomography images with performances exceeding *matched filters* in cases of non a-priori knowledge images.

Though much of of the band-pass properties of the HVS have been experimentally confirmed, there is less study of this property from the point of the **graded resolution** retina structure. In Chapter 5, the contribution of the multiresolutional retina structure as an important component in

the generation of such spatial bandpass images will be modelled.

4.4 Structure of the MF image.

There are two types of MF images used in this thesis namely,

a) the Global MF image (GMF) corresponding to the images obtained during the preattentive phase of human object recognition

b) the local MF image (LMF) corresponding to the images obtained during the local focused-attentive phase.

The LMF is the basis for the *perifoveal-assisted boundary tracking* in chapter 6 and *intrinsic images* in chapter 7.

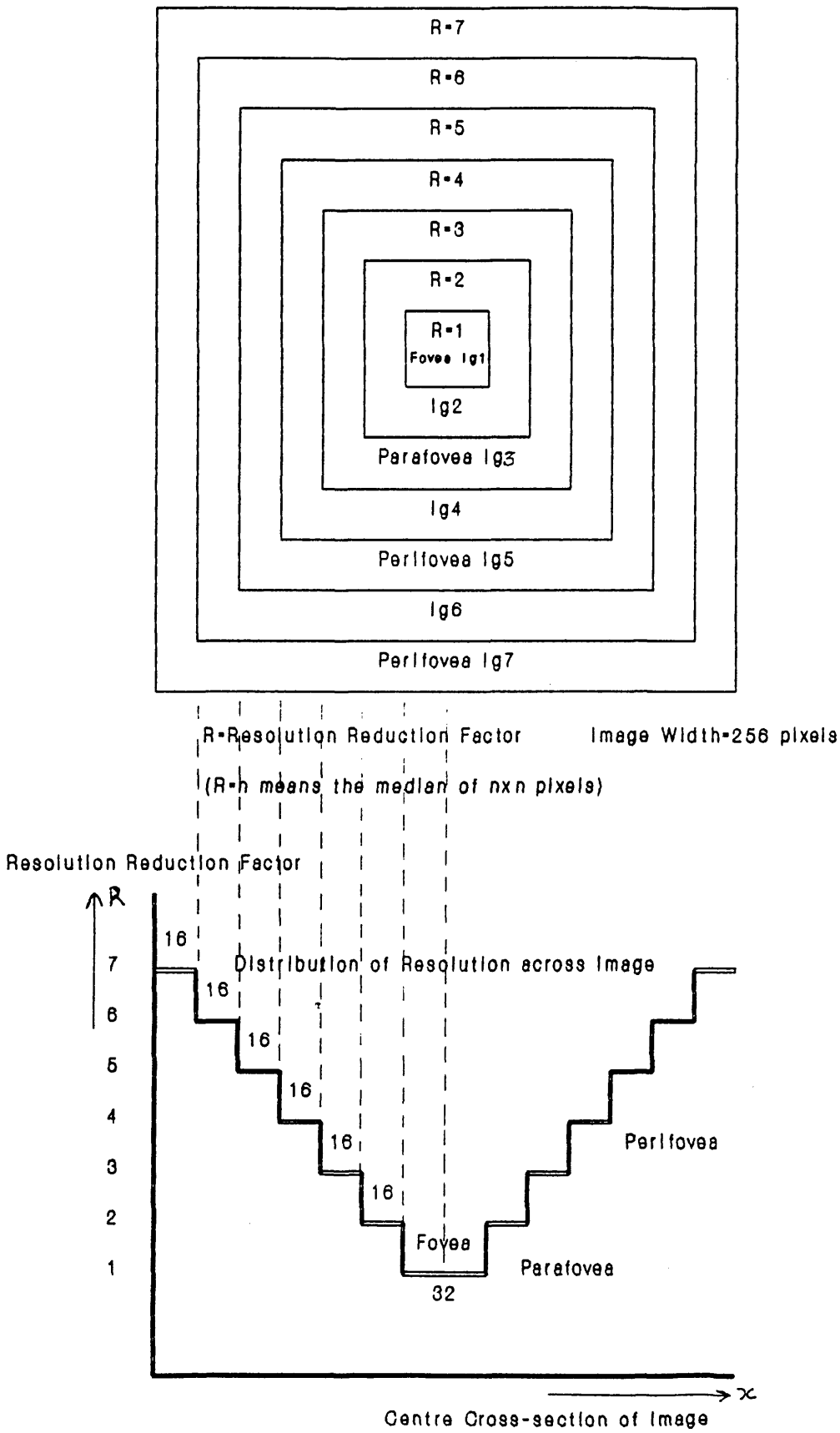
4.4.1 The GMF Image.

Referring to fig. 4.4.1a, the GMF image, I_G of a scene I_s is a set of n annular subimages $\{I_{g1} ..I_{gn}\}$ of different areas $\{a_1 ...a_n\}$ and of resolutions $\{r_1 ..r_n\}$ all sharing a common image axis, formally described by the following equations:

I_s is the input image made up of rectangular co-ordinates of pixels $\Sigma p_{(x,y)}$ for $1<x<M$, $1<y<N$ where M and N are the limits of the images.

$I_G = \{I_{g1} ..I_{g2} ..I_{gn}\}$ where each pixel in I_{gn} is generated from a function Π on a variable neighbourhood of pixels in I_s defined by the function f(n). Π is an operation on a neighbourhood of pixels of the input image e.g. average, mode, etc.

Fig. 4.4.1a Distribution of visual resolution across Multiresolutional Foveal Image



The subimage I_{g1} is called the foveal image and the set of images $\{I_{g2} ..I_{gn}\}$ is collectively called the periforveal image. Following the conventions of Tanimoto, the function Π is known as the *Generating Kernel* and the process is called a *consolidation* of a neighbourhood of pixels of the input image, $p_{(x,y)}$.

Each pixel in I_{gn} is given by $p_{Ign(x,y)} = \Pi \{f_{(n)} * p_s(n)\}$ where $f_{(n)}$ defines the neighbourhood of pixels for the nth perifoveal image and * is a convolution operation. The area of coverage of each subimage Ign is an annular area of common consolidation-neighbourhood $f_{(n)}$. This area is spatially isomorphic (pixel-to-pixel correspondence in x,y) to the input image since the reduced resolution images are not decimated as in the pyramidal image representations.

The resolution of each I_{gn} is defined as the number of similar valued pixels as a result of the consolidation. Hence the resolution of the foveal image is 1 and the resolution of I_{g7} is 7, the higher this number, the lower is the resolution. If the input image neighbourhood of pixels cover more than a line, such a consolidation is called a *non-causal* consolidation, as opposed to *causal* consolidation which takes its neighbourhood from pixels of one raster-line of the input image. It is obvious that the quad tree pyramid images use causal consolidations whereas MF images use causal consolidations. If the neighbourhood is a square k x k window then the resolution reduction will vary as the square of k.

The GMF images are illustrated in fig 4.4.1a and the typical consolidation processes are shown in fig 4.4.1c.

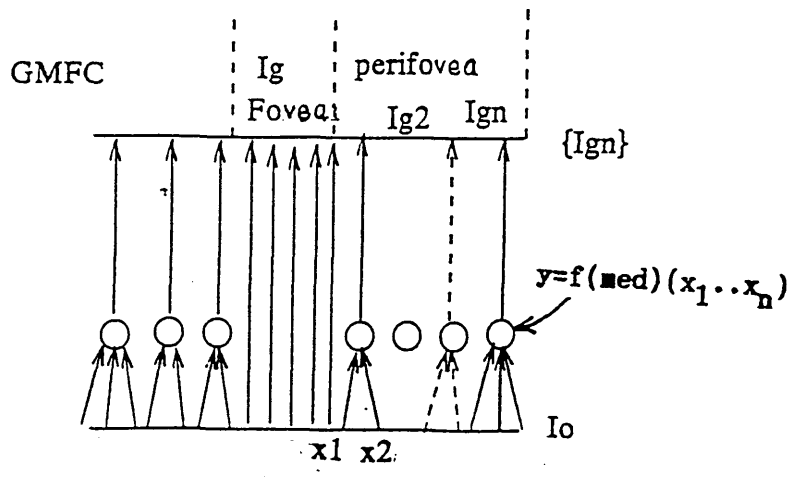


Fig. 4.4.1c Consolidation by Median Generating Kernel

The graded resolutions used here follow a simple linear rule of n=1 to n=N in steps of 1, with the centre image called the fovea image being an all-pass of the input image.

The co-ordinates of the common axis of all the GMF images define the foveation centre of the image. GMF images of different foveation centres are shown in fig 4.4.1d.

Throughout this investigation, the input image size has been chosen to be a 256x256 pixel image of 256 grey level intensities. Causal consolidation have been used with subimage physical dimensions being arbitrarily and experimentally chosen to meet the compromise of accuracy, image fidelity and speed of computation but generally guided by the distribution of the packing density of the photoreceptors across the retina as in fig 4.2.1b.

The table in fig 4.4.1d shows the image coverage area of the GMF images showing the computation advantages in consolidating the input images.

Fig 4.4.1d Table showing the coverage area and number of pixels for subimages.

I_g	R_{gn}	C_o	C_a	N_a
Ig1	1	1024	1024	1024
Ig2	2	4096	3072	1536
Ig3	3	9216	5120	1707
Ig4	4	16384	7168	1792
Ig5	5	25600	9216	1843
Ig6	6	36864	11264	1877
Ig7	7	50176	13312	1902
Ig8	8	65536	15360	1920
	Total		65536	13601

I_g = Perifoveal Images with fovea as I_{g1}

R_{gn} = Resolution Reduction in the I_{gn}

C_o = Overall Image Coverage area in from I_{g1} to I_{gn}

C_a = Coverage area of I_{gn}

N_a = Number of consolidated pixels in I_{gn}

4.4.2 Choice of Subimage dimensions.

The figures above are displayed in the graph of fig 4.4.1f. It shows the comparative time-

units to process a MF image and a uniform resolution image.

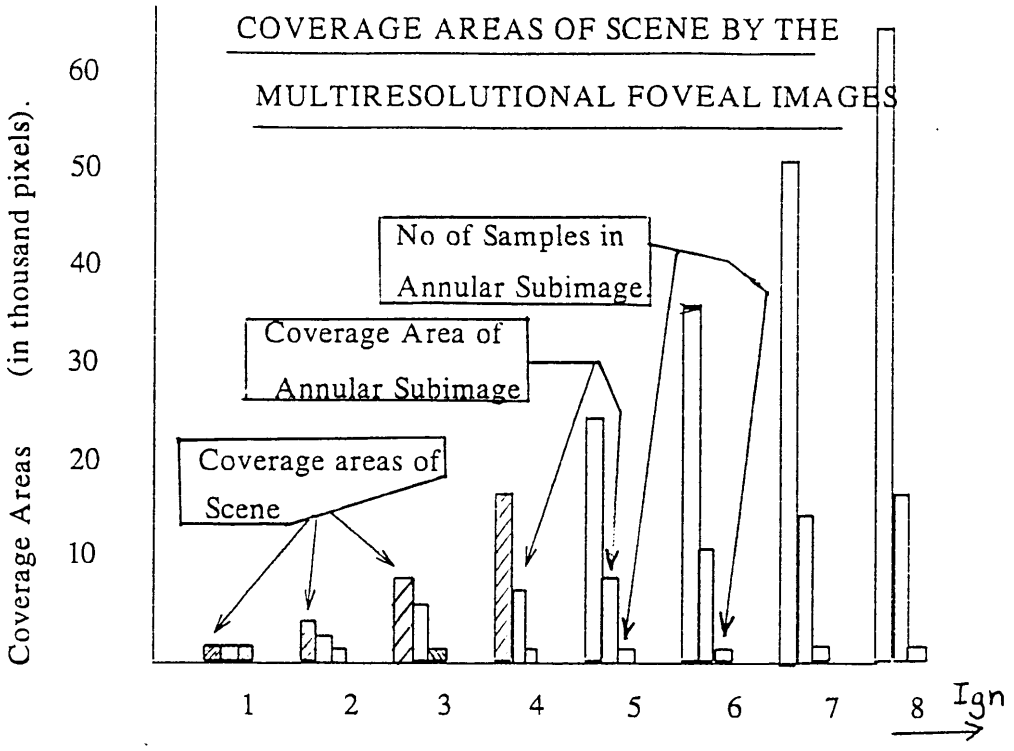


Fig.4.4.1d Subimage No. and Resolution Reduction.

It will be seen that a bigger foveal image will increase the computation time for the GMF image. Using a 32 by 32 foveal image, and a linear resolution reduction consolidation, the total number of perifoveal images is 8. The relative sizes of the perifoveal images are shown in the graph. It indicates that although the annular image coverage area increases rapidly from 1024 to 15360 in 8 resolution steps, the number of samples required to represent the annular images remain quite constant with a mean of 1700 pixels. This offers a faster processing of the image during the preattentive phase of vision.

4.4.3 Shape of GMF Images.

Although non-square concentric GMF images seem to better portray the Multiresolution Foveal Images in the HVS, the generation of such images have been more difficult and farther, the present-day raster line-scan pixel addressing seems not very suited for such processing. For ease in computation, square GMF concentric images have been adopted.

4.4.4 Resolution Reduction in GMF images.

The resolution reduction in the GMF image is a more important issue. This is because there is a intuitive need to have subimages that are more *contiguous* in resolutions so that unnatural artefacts in the image will not be created. Resolution reduction is also related to the choice of causal or non-

causal consolidation of input pixels. Even in causal consolidation, experiments with vertical consolidation, defined as consolidating input pixels in the vertical y co-ordinates as opposed to consolidating in the horizontal co-ordinates, illustrated in fig 4.4.4 show that the choice of vertical or horizontal consolidation seems to be not overly important.

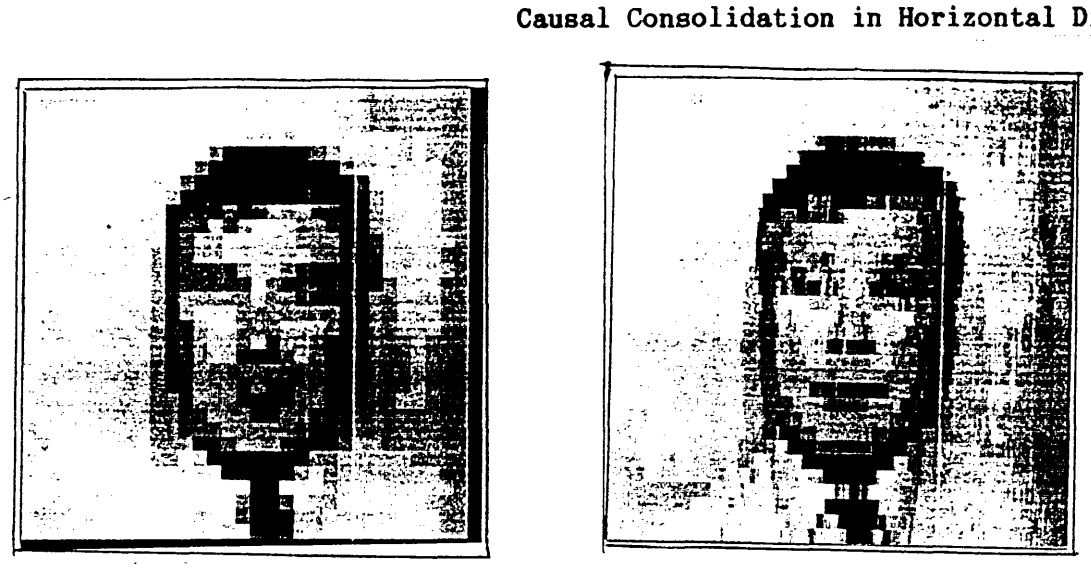


Fig. 4.4.4 Fidelity-loss in Vertical and horizontal consolidation.

4.5 Comparisons with Pyramidal Images.

In the canonical pyramidal images [Tani80], the input image of the n x m matrix of grey pixels need to satisfy the rule that $n = n' \times 2^L$. and $m = m' \times 2^L$, where n' and m' are integers and L is the number of pyramid levels with $3 < L < 10$ typically. Further, sizes of the lower resolution levels of the pyramid need to satisfy the matrices

$$P_i[n_i, m_i] \text{ for } i = 0..,L-1$$
$$n_i = n' \times 2^i$$
$$m_i = m' \times 2^i$$

L-1 level being the input image.

The Generating Kernel chosen in [Tani80] is the simple averaging Function obtained by averaging blocks of 4 pixels in level k to get a output pixel in level k-1 where

$$P_{k-1}(i,j) = 0.25[P_k(2i,2j) + P_k(2i+1,2j) + P_k(2i+1,2j+1) + P_k(2i,2j+1)].$$

It is obvious that the number of image samples to represent a 256 by 256 image is 4A/3 where A is the size of the original image. The pyramid of image gentl is shown in fig 4.5a and is compared to the GMF image of fig 4.5b both using the averaging Generating Kernel.

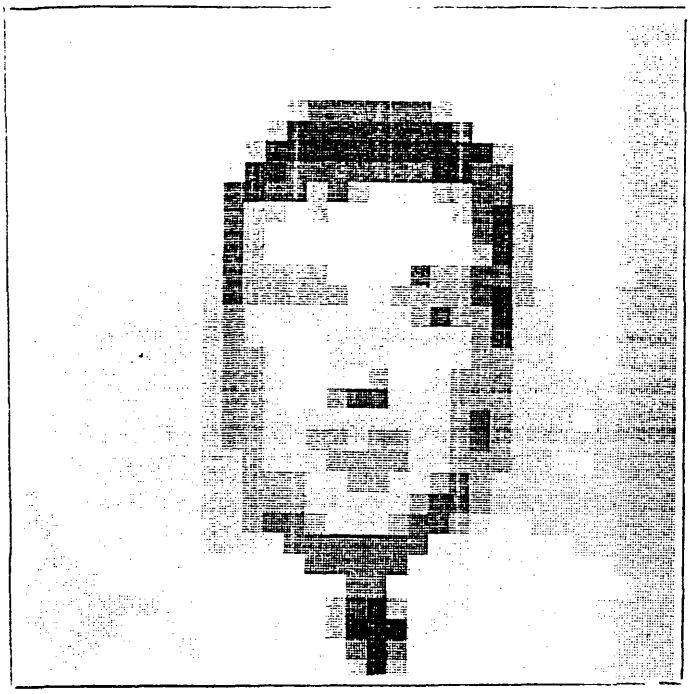
Fig. 4.5a Images at various levels of Canonical Pyramid
using Averaging Generating Kernel.
Layer 1 to layer 8

Layer 1

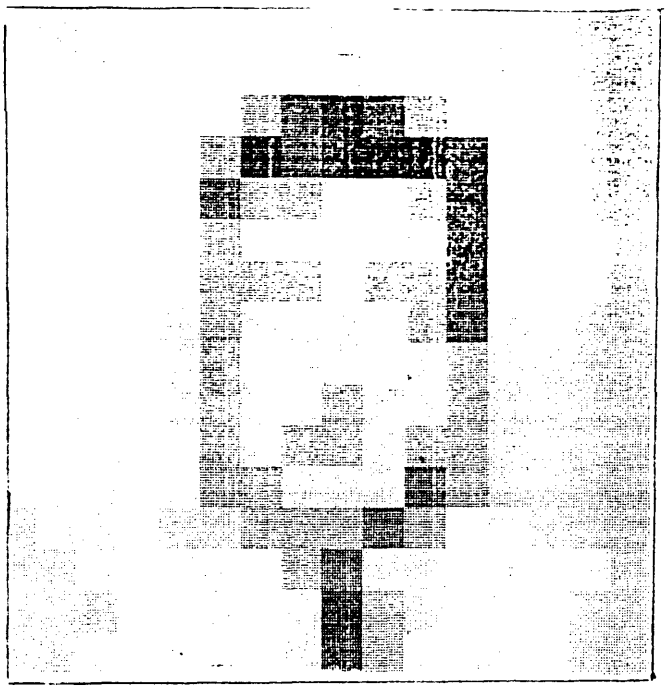
Layer 2

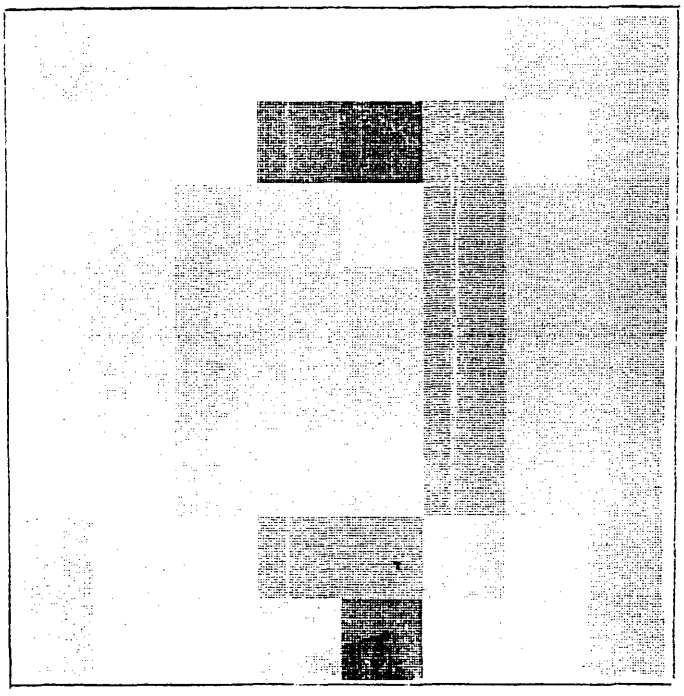
Layer 3

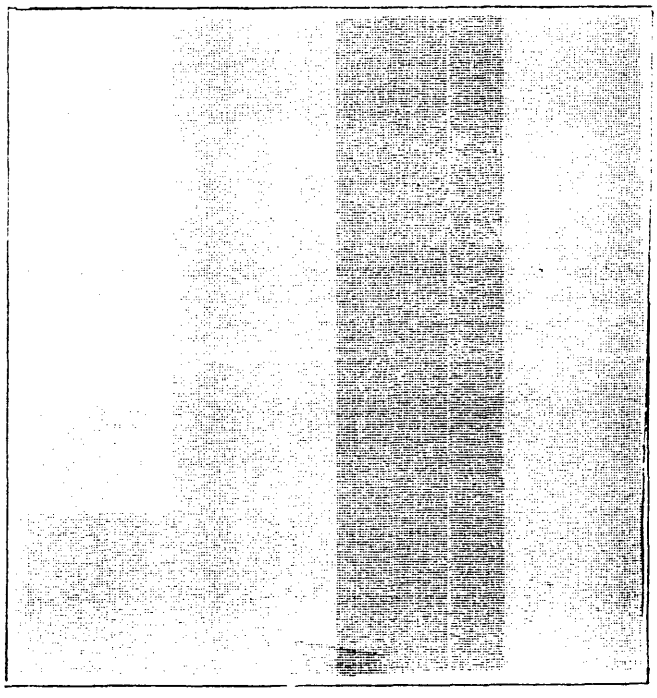
Layer 4



Layer 5







Layer 7

Layer 8

Original Image of Gent1

Multiresolutional Foveal Images with different foveation centres

Fi.g. 4.5b Multiresolutional Foveal Images using Averaging Kernels

Fig 4.5c shows the Fourier Components of the 4th layer of the pyramidal image Gentl (fig. 4.5a). The Fourier Components are grouped around the centre of the image with regions of significant Fourier components concentrated along the w_1 and w_2 axis. Image discontinuities at the top and bottom of the image contribute energy along the w_2 axis (vertical) and discontinuites at the 2 sides contribute energy along w_1 axis. It can be observed that the frequency components are not bunched near the origins of the w_1 and w_2 axis as in the GMF images of fig. 4.5d. This is because the foveal areas near the centre of the GMF image has sharp discontinuities which contribute to the high and low frequency components.

It is clear that to reconstruct the spatial image from the Fourier components of fig. 4.5c, one has to include more frequency components away from the origin, since the energy is more spread out, unlike the case of the GMF image in fig. 4.5d where the frequency components are more concentrated near the origin. In order to recover the image from the Fourier Components of the GMF image, it is only necessary to include those components extending to the limits of w_1 amd w_2 as desired by the fidelity measure. In both cases, we can see that low-pass filtering is the major operation on the orignal images, the difference being on the location and size of the windows involved.

It may be necessary to point out that the pyramidal layers (as in Tani80) are usually obtained through a *consolidation* and a *decimation* operation not exactly like the pyramidal layer shown in fig. 4.5a which is undecimated. This would mean that the Fourier components of the normally decimated pyramidal layer would be reduced in number, since the image dimensions would be similarly reduced.

In fig 4.5a $\log_2 mn$ levels of the pyramid are needed for the original image. In using the pyramid, it is necessary to establish a level called the minimum resolution level that will be the starting level for the top-down hierarchical processing in pyramids. This level is usually obtained a-priori or as a result of a successful search for structure as in [Klinger73] in an application for the the isolation of visual objects. Each level of the pyramid represents a progressively filtered and decimated version of the original image. This is shown in fig 4.5c and d where the Fourier Components of the pyramidal and GMF images show a gradual loss in high frequency components as consolidation progresses.

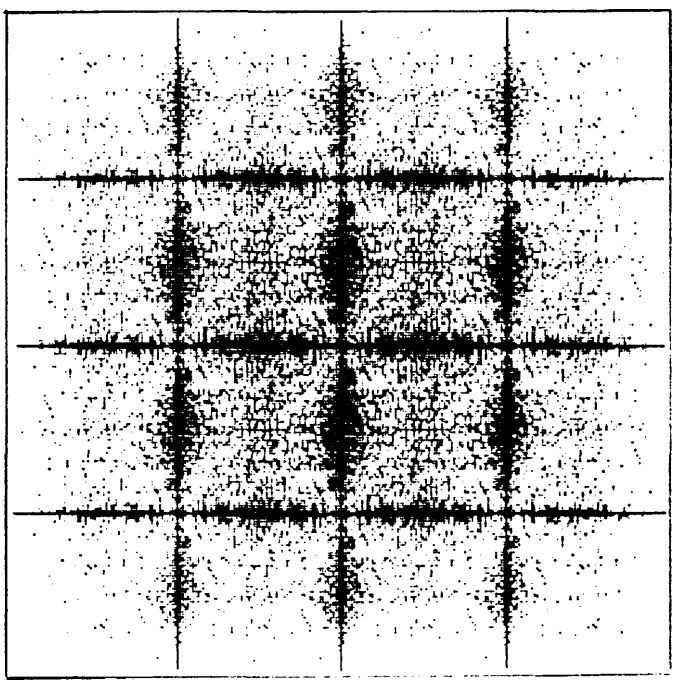


Fig. 4.5c Frequency Components in Pyradmial Images.

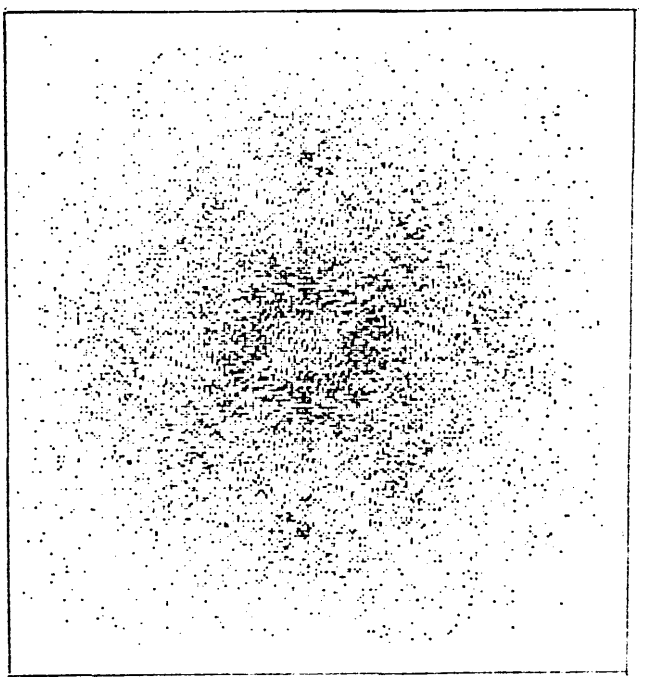


Fig. 4.5d Frequency Components in GMF Images.

Examining the pyramidal image representations of gentl and the GMF image of gentl, one can make the following observations:

a) Resolution reduction:

In the case of the Pyramidal images, processing attempts to start at the least computational-intensive lower resolution level. However, this level is not easily derived from information at a single level. It will be necessary to traverse (usually down) the levels to obtain the starting level. In contrast, in a GMF image, traversing levels is equivalent to moving between the foveal and perifoveal images. As the resolution reduction is more linear, it is expected that the initial level or perifoveal subimage easier to establish in MF images than in the pyramids. This is because the resolution reduction in exponents of 2 results in factors of 4,16,32,64 etc and thus cannot provide an *intermediate finer* resolution to resolve any ambiguities.

b) Image sequence during Computation.

In pyramidal images, each level is processed at a time, though traversing the layers is normally done after any single level is exhaustively searched. At any one instance, there is no equivalence of a datum subimage (fovea) to confirm any computations. The nearest datum image in the pyramid is the 1st level of the pyramid, which is formed from a 2x2 neighbourhood consolidation of the input image. This contrasts with the the MF image where a datum foveal image is composite within the MF image serving as a datum all-pass subimage of the input image. Such a datum subimage can serve as an arbiter of any perception ambiguity.

c) Dimensionality of computational data.

The computational data is the total amount of image data (in this case the amount of pixels) that needs to be processed. In the simplest case of a non-hierarchical image, the computational data is N

However in the 4-window pyramid , the computational data is the sum of the layers l, given by

$$\sum_{l=0}^{l=\log N} 4^l \approx \frac{4N^2}{3}$$

Similarly in the GMF representation, the total number of pixels to represent image I_{GMF} is

$$\sum_{n=1}^{n=r} \frac{a_n}{n}$$

where a_n is the area of perifoveal image I_{gn} and
r is the resolution reduction factor.

4.6. The Optimal Generating Kernel for GMF images.

The optimal Generating Kernel is that consolidation of the input pixels Π and its associated neighbourhood function $f_{(n)}$ such that the output pixel

$$P_{I_{gn}}(n,m) = \Pi * [f_{(n)}(P_{Is}(n,m)]$$

has the least measure of distortion. Whereas the shape of the neighbourhood function $f_{(n)}$ as been chosen to be causal and row-ordered (as opposed to column-ordered), the generating kernel can take on various forms that can *summarise* the data like averaging, median selection, mode selection or sampling.

Formally the averaging and median consolidation are defined as

$P_{I_{gn}}(n,m) = \text{Average}\{P_{I_s}(n,m)\} = 1/T\ \Sigma\{P_{I_s}(n,m)\}$

where $n,m \in S$, S being the set of coordinates of the neighbourhood,
and T is the neighbourhood window area.

The Median consolidation is

$$P_{I_{gn}}(m,n) = \text{Median}\{P_{I_s}(n,m)\} = q$$

where q is that value for which its probability of being less or more than its neighbourhood of pixels are equiprobable.

It is obvious that sampling (random or ordered) of neighbourhood pixels will provide a high distortion measure subject to the Sampling Theorem, though we may recall that this is one method used in image representation (Pratt82) of Satellite images where enormous amount of picture information need to be processed.

Before we need to decide on the choice of the averaging or the median generating kernel we proceed to discuss on distortion due also to causal or non-causal consolidation.

- **Causal or non-causal consolidation.**

Unlike time-series studies, where causal and non-causal can be related to a time datum t_o and usually in one dimension, the concept of causality in a spatial image is conveniently defined [Burt83] so that taking input pixels from a row-neighbour is considered causal. This ignores other factors like frames of images and perhaps non raster scan images. It is obvious that the consolidation on a neighbourhood that has less correlation is likely to generate more errors. This is evident from the linear additive properties of consolidation, an example being

shown in the case of median consolidation later, relating to the limit of perifoveal consolidation. Notwithstanding that there is a higher distortion in causal consolidation, this investigation has kept to the simple causal consolidation since a more elaborate scheme would be needed to maintain a more graduated change in resolution. using non-causal consolidation on 2-D images.

- **Experimental results of different generating kernels.**

A simple measure based on the mean square error (MSE) was used to to test the fidelity performance of the consolidation process defined as

$$MSE = \frac{1}{MxN} \sum_{i,j=1}^{N,M} [x_{(i,j)} - y_{(i,j)}]^2$$

where x and y are the consolidated and input pixels respectively.

From experimental results there seems to be little difference in using the median or averaging generating kernel, though it is obvious that sampling generates a high error (Sampling is not a consolidation in the strict sense of the word). This is because for real-world images[3] where adjacent pixels are correlated the average and the median returns the same result. However for images that have noise and other unnatural intensity discontinuities, the median filter has been shown to be superior [Huang79] as it preserves edges and smoothes out impulse noise. For want of a better generating kernel, the median consolidation is therefore chosen. At this point it may be necessary to point out one basic difference in the *median filtering* as used in image processing and the *median generating kernel* as used here. In the case of median filtering, each output pixel is replaced by the median of the pixels of the input window for unit- pixel sliding of the window. In the case of the MF images, a 1-D window of n pixels (n=1,2..8) is slided over the input image and *all* pixels in the output window is represented from this median. This process is equivalent to a *causal median filtering followed by a zero-order sample and hold.*

- **Upper bound on the perifoveal images.**

The upper bound on the perifoveal image is imposed by the high MSE error at the perifovea. This high MSE can be attributed to the fact that at the perifovea the input window pixels are less correlated leading to the breakdown of the linear additive property in median filtering. For real-world scenes where there is high correlation of neighbourhood pixels [Pratt82], the median of a product of a constant K and any sequence of pixels f(j) is

$$\text{med}\{k\ f(j)\} = K\ \text{med}\ \{f(j)\}\ \text{and}$$

$$\text{med}\{K + f(j)\} = K + \text{med}\ \{f(j)\}.$$

However, for 2 arbitrary sequence f(j) and g(j) this linearly additive property may not hold,

[3] Real-life scenes are typified by strong interpixel correlation.

i.e.

$$\text{med}\{f(j) + g(j)\} = \text{med}\{f(j)\} + \text{med}\{g(j)\}$$

As most real-world images have pixel decorrelation increasing with the spatial Euclidean Distance, the median consolidation of bigger windows will inevitably create a higher amount of artefacts in the output image. This imposes a upper limit to how far we can make good use of the perifoveal subimage without it becoming virtually useless to represent the scene. This can be observed in the perifoveal image $I_{(g7)}$.

4.7 Summary

In this chapter, the support for the Multiresolutional foveal image representation as a rational scheme for robot vision was built. These supports have been built on two main points, being:

a. Physiology of the HVS and in particular the obvious distribution of photoreceptors in the retina and the consequent visual-pathways processing indicate there is both 'hierarchical and multiresolutional' information processing.

b. A.I. and psychological theories of vision based on partial segmentation and preattentive/focused attentional processing influence us to consider that global approaches in **complete segmentation** and non-hierarchical 'searches for structure' limits the capability of effective robot vision.

Using the graded resolution of the retina as a model, the multiresolutional foveal images were developed. Experimental results show that the median generating kernel would not be a irrational choice for producing the MF images.

Chapter 5

Optimal Filtering for Robot Vision Using Multiresolutional Foveal Images.

In chapter 4 the multiresolutional foveal images were derived as an image representation using simple median generating kernel consolidation of input pixels having a basic resemblance to that of the eye. Only the fovea was unconsolidated, representing an all-pass channel. The role played by the fovea in processing an input image is only as **good as it can be supported by the perifovea.** While such hierarchical images (whether in the form of pyramidal hierarchy or in this form of multiresolutional fovea images) have been proposed for machine vision as good models to perform fast search for structure, region analysis etc., there is relatively little emphasis in their being used as filters, not to mention its possible use as a class of *optimal filters*.

Optimal filtering for image processing or machine vision is an extension of the theory of optimal filtering and associated *matched filtering* from digital communications and control. The basic question that has to be posed is this:-

Is optimal filtering applicable to robot vision?

If its applicable then it would not be too naive to assume that whatever filtering that is performed by the HVS, that filtering must have semblance to such optimality. This chapter shall provide further analysis of the MF images in their use as a *set of graded filters* having features that are optimal for image understanding. The question that has therefore to be answered is this :-

> If the multiresolutional images are the products of a set of simple low-pass filter channels, how could optimal filtering result from this set of low-pass channels?

5.1 Relationship between the Multiresolutional Images and Spatial Bandpass Filters in Robot Vision.

Visual perception theory is sometimes approached from two angles, namely the *feature detection theory* and **Spatial** Frequency theory. Up to the 70s, the neuronal organisation and response of the visual cells to lines and edges made the feature detection theory very acceptable. However, it breaks down in attempts to explain some visual illusions and also seems to be not consistent with the more recent findings in cell-responses to receptive fields. The importance of the spatial frequency theory can be traced to the paper by Campbell and Robson [Campbell68]. Campbell used psychophysical arguments and experimental evidence to suggest that the HVS might be analysing a visual image by its spatial frequency components[1] as shown in the fig 5.1.

---

[1] Frequency components in this context are the components of the image in the spatial xy space, not the electromagnetic/chromatic frequencies of the light spectrum.

This spatial frequency theory is now becoming more important as it is able to explain some visual illusions (e.g Poggendorf and the Muller- Lyer illusion) which the feature-detector theory finds it difficult to provide a good answer. In its relevance to the multiresolutional images, the spatial frequency theory can be summarised in the following points:-

a) the HVS contains several sensory channels, each selectively sensitive to a different, moderately narrow range of spatial frequencies.

b) The outputs from these channels are stochastically independent.

c) The channels are passed through separate thresholds. The importance of the spatial frequency theory having been mentioned, we will now proceed to see how such spatial bandpass channels as advocated in this theory can be considered as a class of optimal filters.

5.2 Optimal filtering in machine vision and image understanding.

Physiological measurements on single cells of the eye and the visual cortex seem to confirm that the retina image is analysed by a *set of quasilinear filters of overlapping frequencies, each filter being of the bandpass type of 1.5 octaves wide, which at any location in the retina, spans 4 octaves* [Georgeson75]. Sullivan [Sull84] argues that matched filtering is not practical in image understanding. This must presently be quite obvious, though if we look at machine vision about a decade ago, there was still intense pursuit in, for example, the design of better edge-detection masks which can be regarded as a form of matched filters for high frequency detection.

By definition, a matched filter is one whose characteristics is such that its transfer function is the complex conjugate of the signal to be detected. In the case of digital communications where the coding process has made the detection of the signal pulse from the background of noise as the prime interest, there is some a-priori knowledge about the type of noise, fading characteristics of the channel and the type of signal pulse constituting the signal. Such a-priori knowledge, especially in the approximate shape or time-occurence of the pulse is definitely useful in the design of matched filters in digital communications. However in image–understanding the signal (object? feature? line?) is not only of one type (for example an *edge* where one can therefore design a matched filter for high-spatial frequency detection) but also many of other types (e.g. regions, orientations, homogeneity etc.) so that there must be an infinite number of matched filters needed for one task.

Sullivan's argument is that the type of optimal filtering required for image understanding is a set of spatial bandpass filters. In [Baker and Sullivan80], attention was drawn to the use of bandpassed images as a desirable method for the analysis of computer tomography images. Although much of the spatial channel properties have been obtained from psychophysical observations, (study of the relationship between stimulus and perceptual appearance)

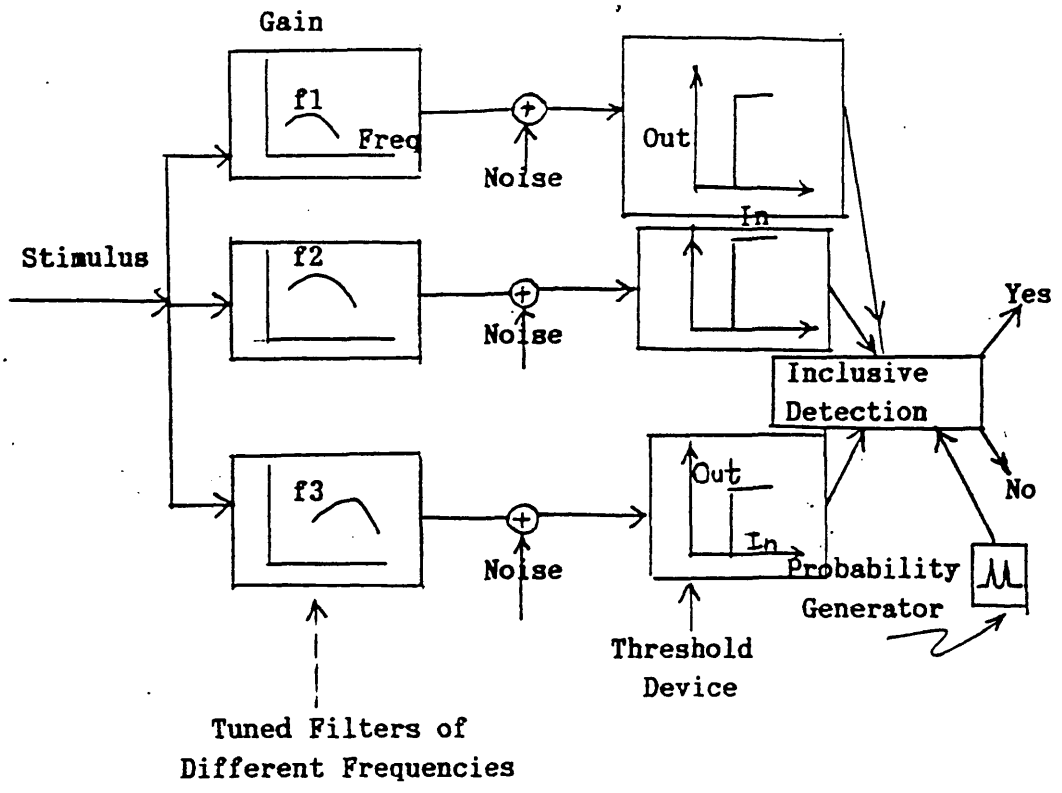


Fig. 5.1 Model of Spatial Frequency Channels
in HVS (taken from Campbell68)

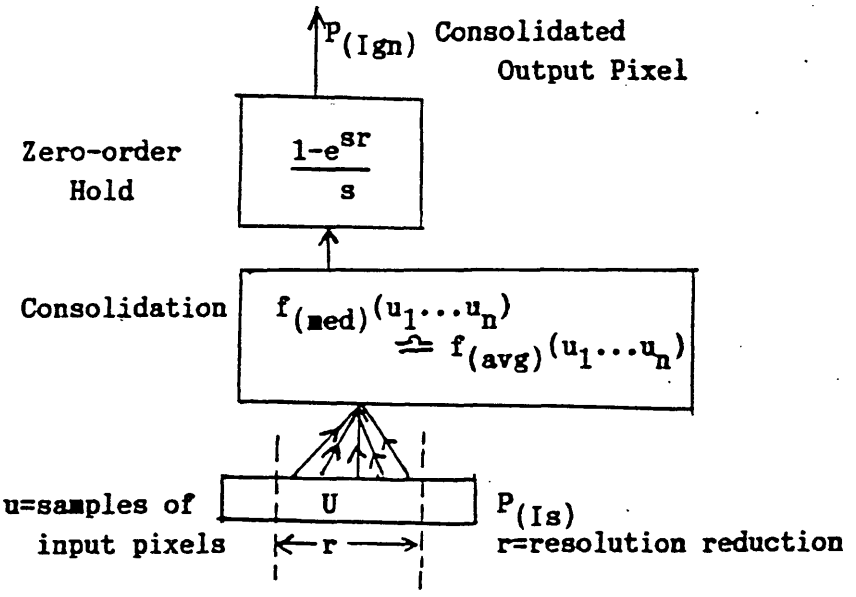


Fig 5.3a Equivalent Consolidation in MF Images.

there is not much literature that studies how such spatial channels are actually formed in the earlier visual pathways, i.e. at the retina fovea and perifovea. In particular there is little mention of the generation of the bandpass spatial channels due to the multiresolutional nature of the retina[2].

It will be shown here using simple linear time invariant (LTI) system representations that the set of lowpass filters under the conditions of *inhibition and receptive fields* perform like bandpass filters with the bandpass characteristics dependent, amongst other factors, on the size of the receptive fields.

5.3 Filtering in Multiresolutional Images.

We note that the the exact *consolidation* process in generating the multiresolutional Images was that of a median filtering followed by a zero order sample-and-hold as shown in fig 5.3a.

Noting that for a real-world image where there is high correlation between adjacent pixels, there is actually a small difference in using the averaging generating kernel or the median generating kernel, both being smoothing filters [Huang79]. In fact if the statistical distribution of the image intensities in an image is symmetric, the median equals the mean or average. We shall therefore assume the use of the averaging filter in place of the median for the purpose of the frequency domain analysis of the filters, as averaging filters are better represented as polynomials in digital filter analysis. Representing the images as a 1-D array of pixels as in fig 5.3b,

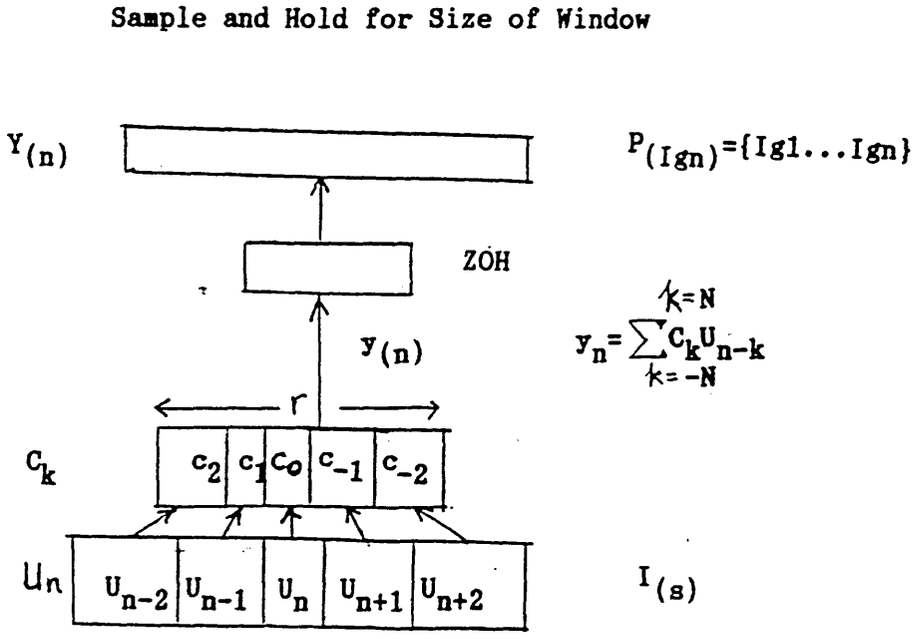


Fig 5.3b MF Image as a 1-D Non-recursive Window Filter.

[2] By definition [Ganong81], the retina comprises the photoreceptors, the neuron cells (bipolars, ganglion etc.) and optic nerve fibres.

where

$P_{(I_s)}$ is a row of the input pixels $\{u_n\}$

$P_{(Ign)}$ is a row of the output pixels $\{y_n\}$

C_k is a row of coefficients,

we can then represent a consolidation of the input image as a recursive equation of polynomials in n

$$y_n = \sum_{k=-N}^{N} C_k U_{n-k}$$

In the multiresolutional images, C_k is chosen to be equal-valued and unity (i.e. unweighted filtering) and the resolution reduction r generates the output pixel

$$y_n = 1/r\ (u_{n-r} + u_{n-1} + u_n + u_{n+1} + u_{n+r})$$

Such filters represent a class of discrete non-recursive filters variously known as FIR (finite impulse response) filters, transversal filters or moving average filters).

The output image P_{Ign} can be specifically represented, for a odd- size window, as

$$y_n = \frac{1}{2N-1} \sum_{m=-N}^{n} U_{n-1} \qquad ---(1)$$

The transfer function $H_{(w)}$ of such a filtering process can be obtained by using the eigen-function in w, e^{jwt} as

$$H_{(w)} = 1/r[e^{-Njw} + .. + e^{-jw} + 1 + e^{jw} + e^{Njw}] \quad ---(2)$$

Using Eulers Identity $\cos x = 1/2\ (e^{jx} + e^{-jx})$, we have

$$H_{(w)} = 1/r[1 + 2\cos w + 2\cos 2w + ...2\cos Nw] \quad ---(3)$$

Rearranging ---(2) into a geometric progression of e^{jx} terms,

$$H_{(w)} = 1/r\ [e^{-jNw} + e^{-j(N-1)w} + .. + e^{jNw}] \quad ---(4)$$

$$H_{(w)} = \sin (N+0.5)w/(2N+1)\sin(w/2) \quad ---(5)$$

The graph of eq. 5 is plotted in fig 5.3c.

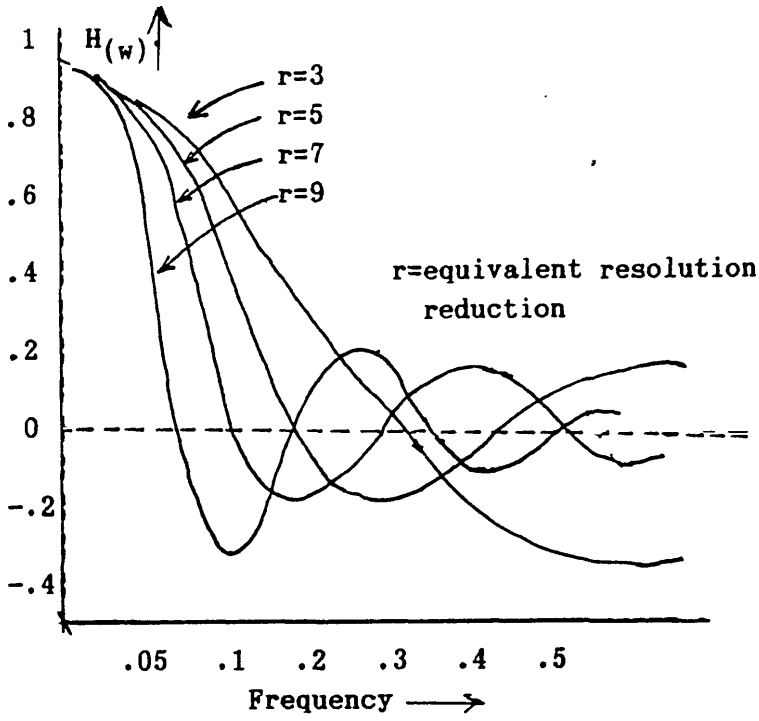
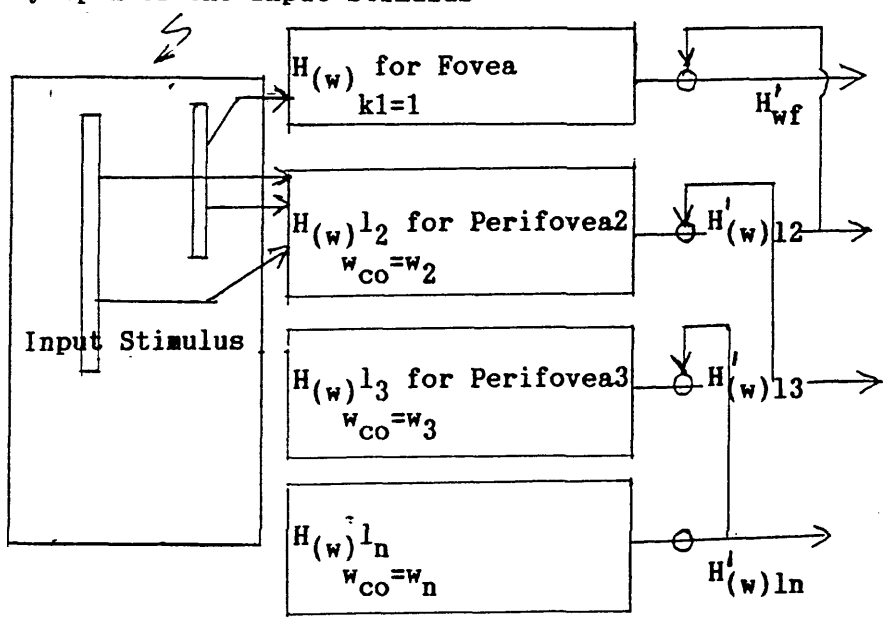


Fig. 5.3c Equivalent Low-pass Filtering in Perifoveal Channels.

Variable dimensions of receptive fields represented by span of the input stimulus



Note: The general form of $H_{(w)_{ln}} = 1/(1+jwt_n)$

where $1/t_n$ is the cutoff frequency associated with the perifovea and dependent on r

Fig. 5.4a

Equivalent foveal and perifoveal filters subjected to different receptive fields.

The effect of the sample-and-hold is equivalent to taking the output sample every r pixel distance introducing a bigger error, though the general filter characteristics remain unchanged (i.e $|H_{(w)}|$ non- monotonically decreasing with w).

It shows the following features:-

a. The general shape of the filtering is that of a low-pass filter being periodic with w since $|H_{(w)}|$ decreases non-monotonically with w. (The effect of the sample and hold process in the actual implementation is equivalent to taking samples every r pixel distance and holding for the duration. This only introduces aliasing and other sampling related errors but the equivalent $|H_{(w)}|$ is also non-monotonically decreasing with w.)

b. The bigger the number of terms in C_k the lower is the cut-off frequency, i.e the cut-off frequency of the low-pass filters decrease with r, the resolution reduction factor.

c. These equivalent low-pass filters generated the perifoveal images are periodic with w. (This must be expected since they are a class of unweighted, non-recursive filters.)

It should be possible to employ different filtering schemes to obtain better filter characteristics (e.g least-square quadratic smoothing) or plain Gaussian, but such is not the purpose of this chapter. Suffice now to show that the perifovea images are equivalent to low- pass filters through the filtering scheme in eq. 5. We will now proceed to show that such a set of low-pass filters under conditions of variations of the receptive fields and *inhibitory conduction* can act as a set of bandpass filters with *variable bandwiths and variable centre-frequencies* thus conforming to some of the characteristics of optimal filtering as in [Sull85].

5.4 Simplified analogies with receptive fields and neuronal inhibitory conduction.

Sachs [Sach71] attributes the existence of spatial channels using electrophysiological evidence that neural substrates in the ganglion cells **with overlapping receptive fields** and **antagonistic centre-surround activity** can generate bandpass channels. The location and bandpass region of such a filter would be a function of the **dimensions** of the receptive fields. Studies by Enroth-Cugell [Enroth66] reveal that receptive fields of **different dimensions** exist within the different visual regions of the retina and that the outputs of the cells have **subtractive** functions. Also in [Ratcliff63], lateral inhibition conduction, amongst the other neuronal conduction mechanisms, was considered the most important contributor to the frequency response of the eye.

We proceed now to make the following simplifications where the foveal and perifoveal filters are regarded as a linear time-invariant transfer functions as in fig. 5.4a.

Using the one-sided symmetry[1] of the eye and allowing all the filter gains to be unity, let the set of the retina channels be made up of a parallel cascade of n foveal and perifoveal channels $\{H_{(w)1}, H_{(w)2}, \ldots H_{(w)n}\}$ where the subscript 1 represents the fovea and (2..n) represents the perifoveal channels. Then, the foveal channel being an all-pass channel is

$$H_{(w)1} = 1 = k$$

where k is a constant, independent of frequency w

and the perifoveal channels, modelled as a simple first-order low-pass channel is

$$H_{(w)i} = \frac{1}{1 + jwT_i}$$

with i indexing the ith perifoveal channel of cut-off frequencies

$$1/(T_i).$$

Clearly the graded resolutional filters indicate a graduated cut-off frequenciesarranged in the order

$$T_2 < T_3 < T_3 \ldots$$

The variable-size receptive field in one dimension is equivalent to a region of common stimuli or common inputs, equivalent to an electrical short circuit of the filter inputs. Using the Cornsweet model of neuronal cell inhibition shown in fig 5.4b

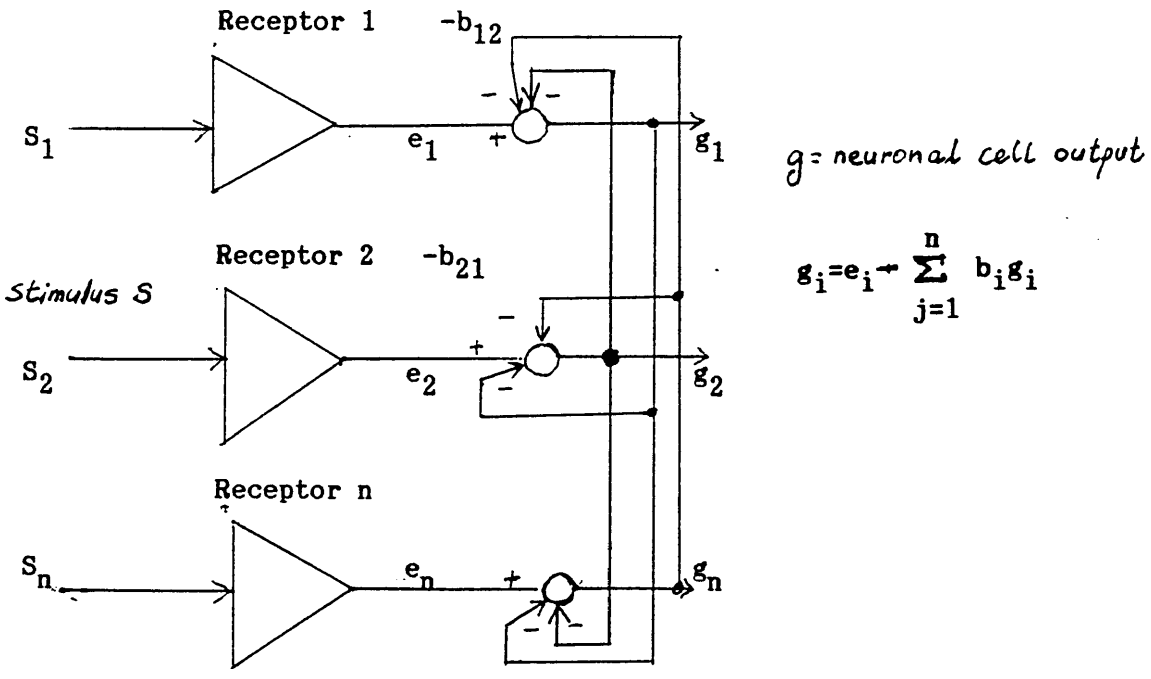


Fig. 5.4b Cornsweet Model of Backward Inhibition of Neuronal Cells

1. As the perifoveal regions are concentric about the fovea, there is symmetry extending radially from the fovea.

The model in fig 5.4a shall now be analysed for different sizes and locations of receptive fields.

5.4.1. Receptive fields near fovea and the generation of high-pass filters.

Assumptions: Small receptive field covering fovea and perifovea2.

Inhibition: Foveal channel and perifovea2 outputs are subtractive.

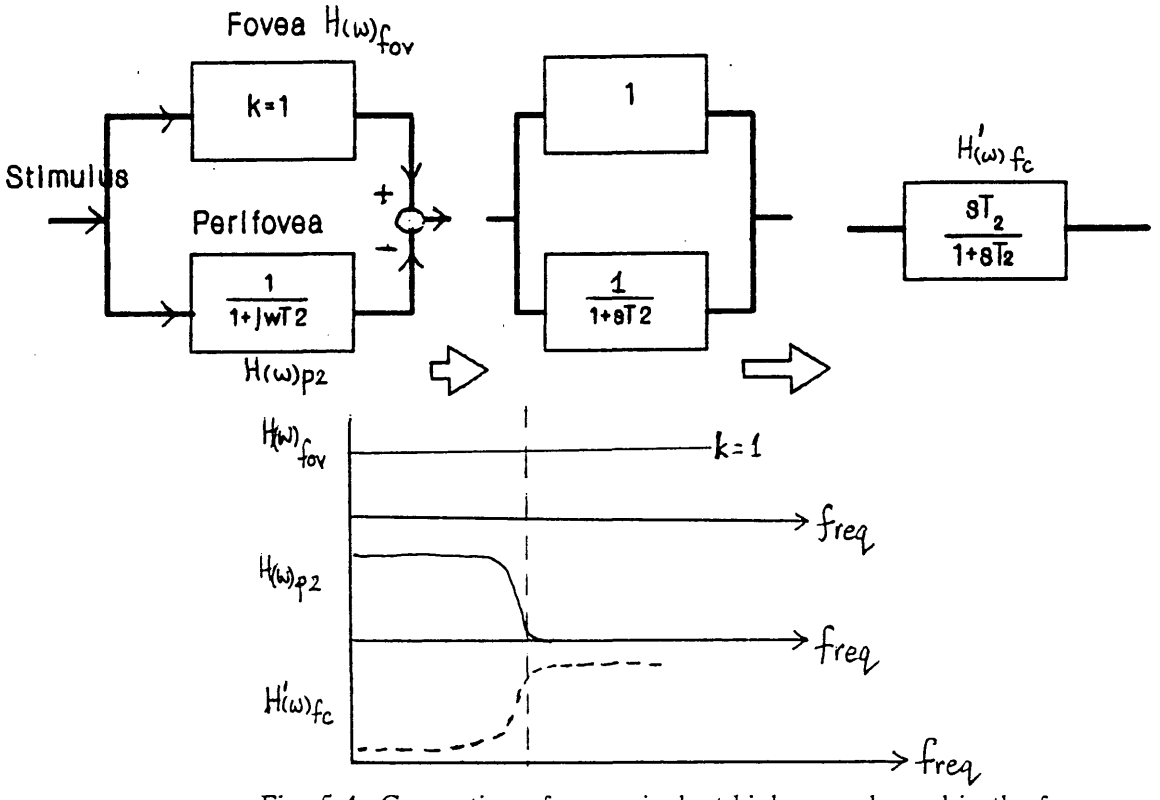


Fig. 5.4c Generation of an equivalent high-pass channel in the fovea.

The equivalent transfer function of the fovea channel (fc) interface is

$$H'_w(fc) = 1 - \frac{1}{1+sT_2}$$

$$H'_w(fc) = \frac{sT_2}{1+sT_2} \qquad \text{---- (5.4.1)}$$

which is equivalent to high pass filter with upper cut-off frequency

$$\frac{1}{T_2}$$

Therefore we see that an *ideal high pass channel*[1] is formed at the fovea channel through a common receptive field at the fovea region, with perifovea2 channel inhibiting.

5.4.2. Bandpass Channels at the Perifovea.

Receptive field: Small receptive field covering perifovea2 and perifovea3

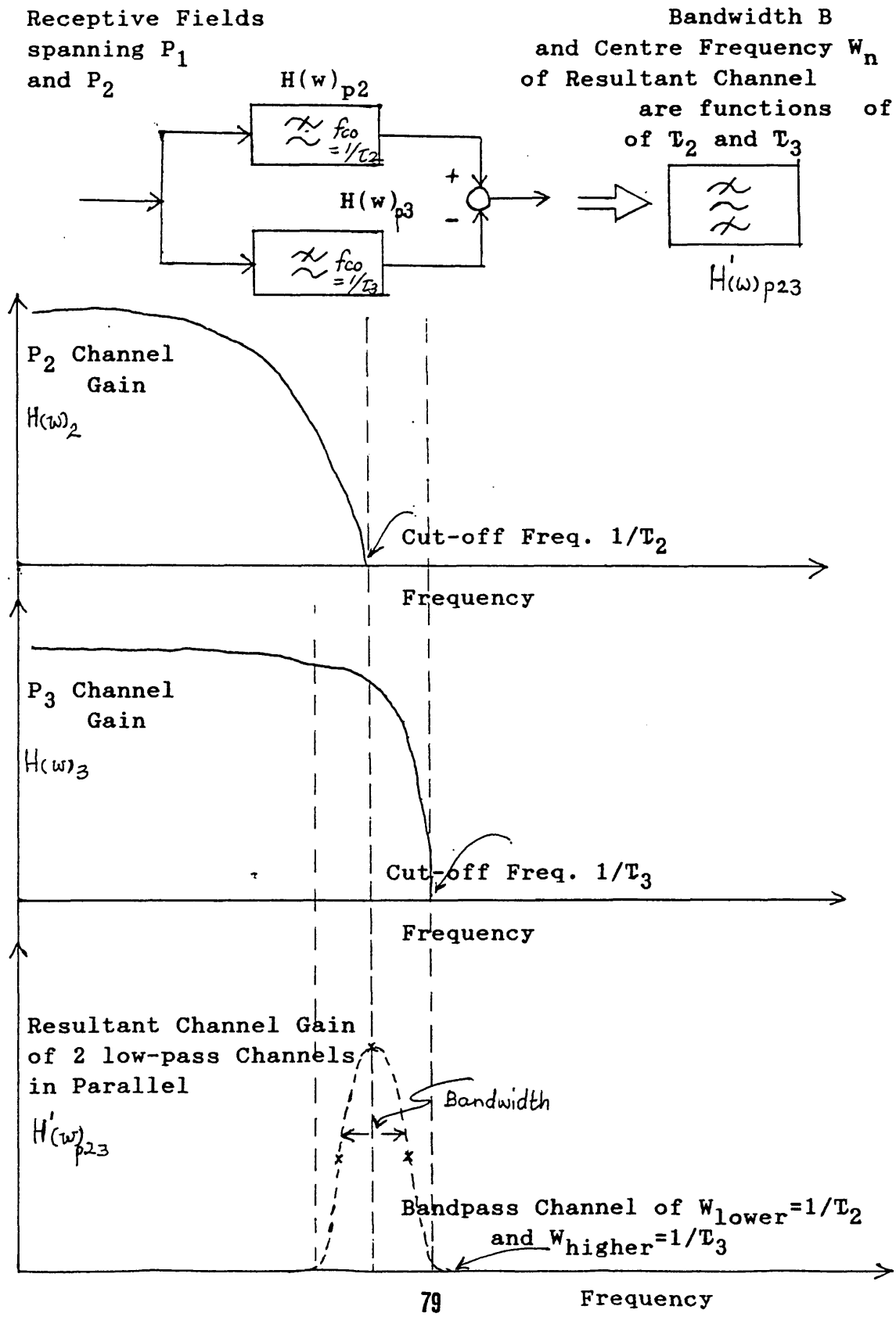
---

1. The existence of a high-pass channel was suggested in [Campbell-Robson68] through psycho-visual experiments.

Fig. 5.4d Generation of Bandpass Channels at Perifovea2

Note: Inhibition: perifovea2 and perifovea3 channels are subtractive.

P_1 and P_2 are the
Perifovea 1 and Perifovea 2 respectively



The equivalent transfer function of the perifovea23(p2 (fig. 5.4d) is

$$H'_{(w)_{p23}} = 1/(1+sT_2) - (1/(1+sT_3)$$
$$= s(T_3-T_2)/(s^2+s(T_2+T_3)+1) \text{-- (5.4.2)}$$

Rearranging eqn (5.4.2) into the canonical form of a bandpass filter,

$$H'_{(w)_{p23}} = 2\zeta w_n s / (s^2+2\zeta w_n+w_n^2)$$

where $2\zeta w_n$ is the bandwidth

and w_n is the centre frequency of the passband

Representing $H'_{(w)_{p23}}$ in terms of filter constants A, B, C, we get

$$H'_{(w)p23} = \frac{AsB}{s^2+sB+C} \quad \text{--- (5.4.3)}$$

we then obtain

A = gain of the bandpass channel = $(T_3-T_2)(T_2T_3)$

B = bandwidth of the bandpass channel = $(T_3+T_2)/(T_2+T_3)$

C = centre frequency of the passband = $\sqrt{(1/(T_2T_3))}$

all A, B, C being dependent on the perifoveal cut-off frequencies T_2 and T_3

We can therefore see that from this simplified analysis that the emergent characteristics of the cascade of low-pass filters or the multiresolutional filtering at the perifoveal regions is the generation of a set of graded bandpass filters.

Overall, the common stimulus in the receptive fields and the inhibitory action results in the following:-

a) a high-pass filter at the fovea region (*not just an all-pass filter*)

b) a set of bandpass filters of variable bandwidth, gain, and cutoff frequencies depending on the associated perifoveal regions. (See Fig. 5.4e)

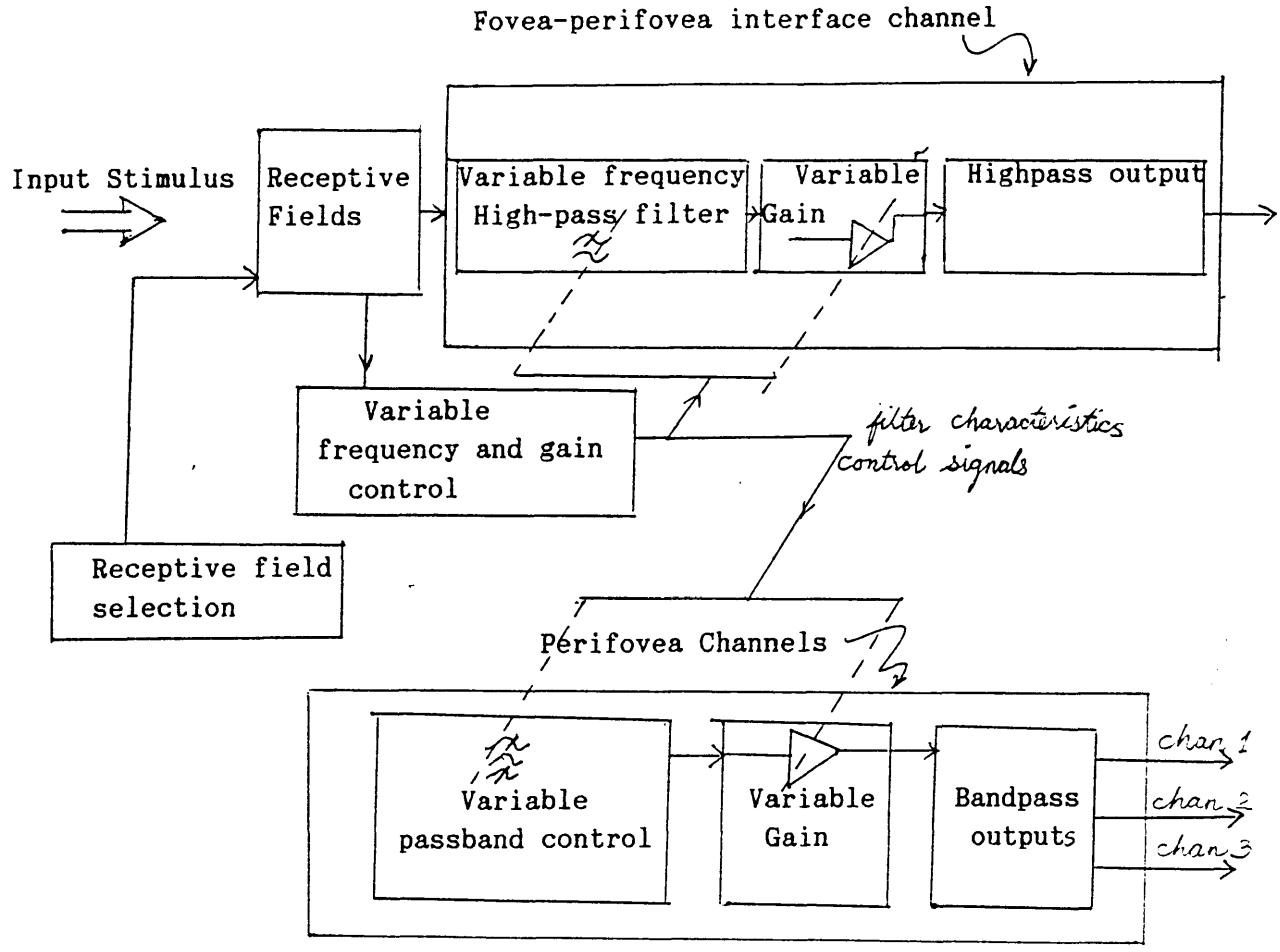


Fig. 5.4e Spatial highpass and bandpass channels from the Multiresolutional Foveal image processing.

Addendum.

An Explanation on the generation of bandpassed images from Multiresolutional Foveal Vision .

Using the illustration of fig. 5.4a the group of input stimuli to the eye (which must now be regarded as an arrangement of an all-pass filter (fovea) and progressively low-passed filters (perifovea)) is actually the scene to be foveated. What is important is to consider that the size and position of the receptive fields are a function of the movements of the eye, meaning that as the eye changes its foveation either by ballistic eye movements or by saccadic movements, the receptive fields change accordingly. These stimuli will be filtered accordingly and will be subtractively combined at their outputs. This subtractive function is based on the fact that the dominant mode of neuronal processing in the visual cells seem to be lateral inhibition (Cornsweet 70).

Consider the first case when the eye looks at a scene using channels near the fovea. In this case the receptive fields will affect both the fovea and the immediate perifovea-2 as shown in the stimulus area marked 'H'. It will be seen that their respective outputs, when subtracted will cause the existence of a high-pass filter (which is quite unlike the normal high-pass filtering using image sharpening). Thus what is significant here is that although the eye is known only to be made up of all-pass and low-pass channels, by the change in the receptive fields , due to eye movements we do also perform high-pass filtering.

In the second case, if we should move our eyes so that we now have the scene stimuli acting on the perifovea-2 and perifovea-3 area, as represented by the receptive field area marked 'B', we say that the outputs of these channels will be equivalent to bandpass channels. This can be explained by the fact that the subtractive outputs from graded low-pass channels are equivalent to bandpassed outputs, as can be seen from the graph in fig. 5.4d.

Although the above illustration seem obvious in linear time-invariant system theory, it is not very conventional to imagine that filter outputs should be made to subtract.

It is the change in the receptive field size and receptive field locations that create such high-pass and bandpass filters.

Hence referring to fig. 5.4e, we can explain that the selection of the receptive fields (as a result of changing foveations) will cause us to be doing an equivalent filtering of the scene using combinations of high-pass and bandpass channels. A significant point is that such filter characteristics are readily made variable (passband centre-frequency, passband bandwidth, etc) by the controlled movement of the eye.

Chapter 6

Perifovea-guided semantic line tracking.

In many applications of robot vision for object recognition, much of the initial difficulties, by necessity, have to be solved by proper pre-engineering of the scene using various methods of *structured lighting*, etc. so that the object contours can be obtained in an easier manner. Such a boundary becomes a part of the *feature extraction* process for further recognition of the object. A very real problem is to distinguish what constitutes the boundary of a object if the scene is visually complex. This is in fact compounded by the problem of defining what an *edge* is (e.g is a shadow of an object or is the intensity-discontinuity on its surface due to reflectance, etc., an edge?) since there is fuzziness in the concept of an "edge" [Levialdi83]. Therefore, before we can define a boundary, we need to define what an edge is, as one simple definition of an *object boundary* is that it is the outer edge of an object fulfilling certain semantic descriptions for that boundary (e.g. concavity, curvature, etc.). It must be plain that semantic boundaries of objects must yield a better feature for recognition than non-semantic ones (e.g one based on intensity threshold).

In this chapter we shall not concern ourselves with the complexity of edges but rather use standard ways of detecting intensity discontinuities defined as its gradient $\nabla f(x,y)$ specified as

$$\nabla f(x,y) = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}$$

giving its magnitude

$$\Delta = |\nabla f((x,y)| = \sqrt{(\frac{\partial f}{\partial x})^2 + (\frac{\partial f}{\partial y})^2}$$ and the

orientation or angle of the edge as

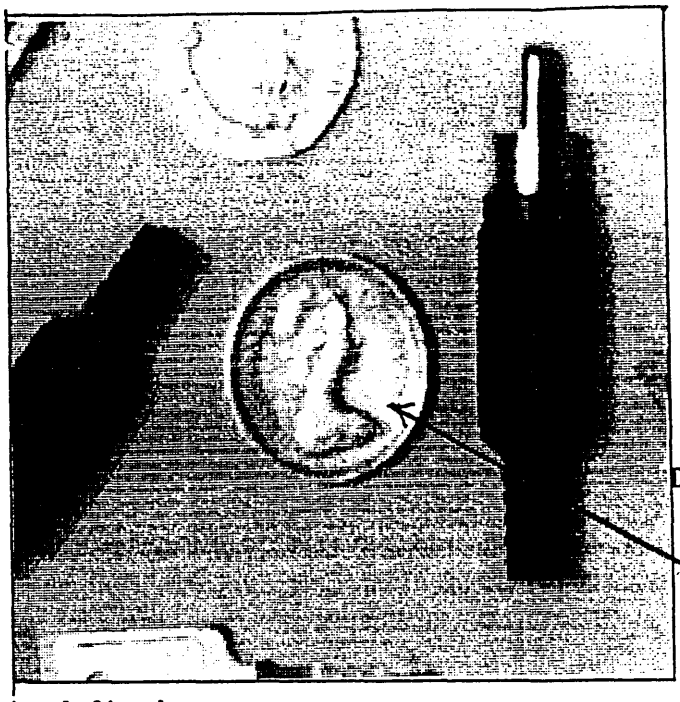
$$\theta = \tan^{-1} \frac{\frac{\partial f}{\partial x}}{\frac{\partial f}{\partial y}}$$

However, we shall attempt to use as much of the edge information as possible so that they could be matched against the semantic inputs to yield a semantic line. A *semantic line* becomes a line formed by fulfilling a set of semantic descriptions. This idea is not new as multidimensional space edge tracing has been proposed by [Liu78, Herman and Liu79]. However the use of *perifoveal support* to reduce computation search space and to suppress spurious edge points in order to obtain a truer line is the main focus in this chapter.

6.1 The Problem of defining a semantic line.

The problem of defining the boundary of an object in terms of image primitives (edges, regions, contours etc.) is a very complex one [Shirai71]. A simple illustration of this

difficulty is shown below:



where the task is defined as :

"Generate the semantic boundary of the object Coin1 (coin at the centre)".

Fig. 6.1a The difficulty of obtaining a semantic boundary.

We can use many edge detection masks to obtain image discontinuities. These range from from the simpler Sobel, Prewitt, compass gradients to the more complex ones, DOG, Cannys etc. Although what they yield is the basic information on ∇, robot vision has yet to make *full* use[1] of such information. Normally, an early attempt is made to threshold the points so that further processing can proceed, thus defining an edge as one that survives after the threshold. To use such edges means the sacrifice of some useful information since basically what is then left is a discontinuity pixel having $|\nabla f(x,y)|$ exceeding the threshold. Efforts to join such edges will at best yield non-semantic lines [Marr82]. In most cases, *edge orientation* information is not used.

If we can proceed a step ahead and try to use some prescribed information in an edge (e.g intensity profile of edge, edge neighbourhood metrics etc.) these edges are likely to provide a better semantic description of the objects of the scene. Obviously, such an attempt is bound to be difficult and time-consuming since there is a need to perform matching at every pixel with its neighbourhood, (e.g in a 5x5 neighbourhood, a point has to be matched to 24 of its neighbours for k semantic descriptions and proceed to its next point and do the same for 24 neighbours while bookkeeping its search path). It is perhaps for this reason that most of the edge-linking in industrial robot vision use simpler region growing or *erosion* techniques to perform connectivity, based on the non-semantic information of the edge points

1. As intensity-discontinuity thresholding, which is a non-linear operation, is done quite early in conventional robot vision, this means the loss of much of the information that intensity-discontinuity carries.

as it guarantees convergence.

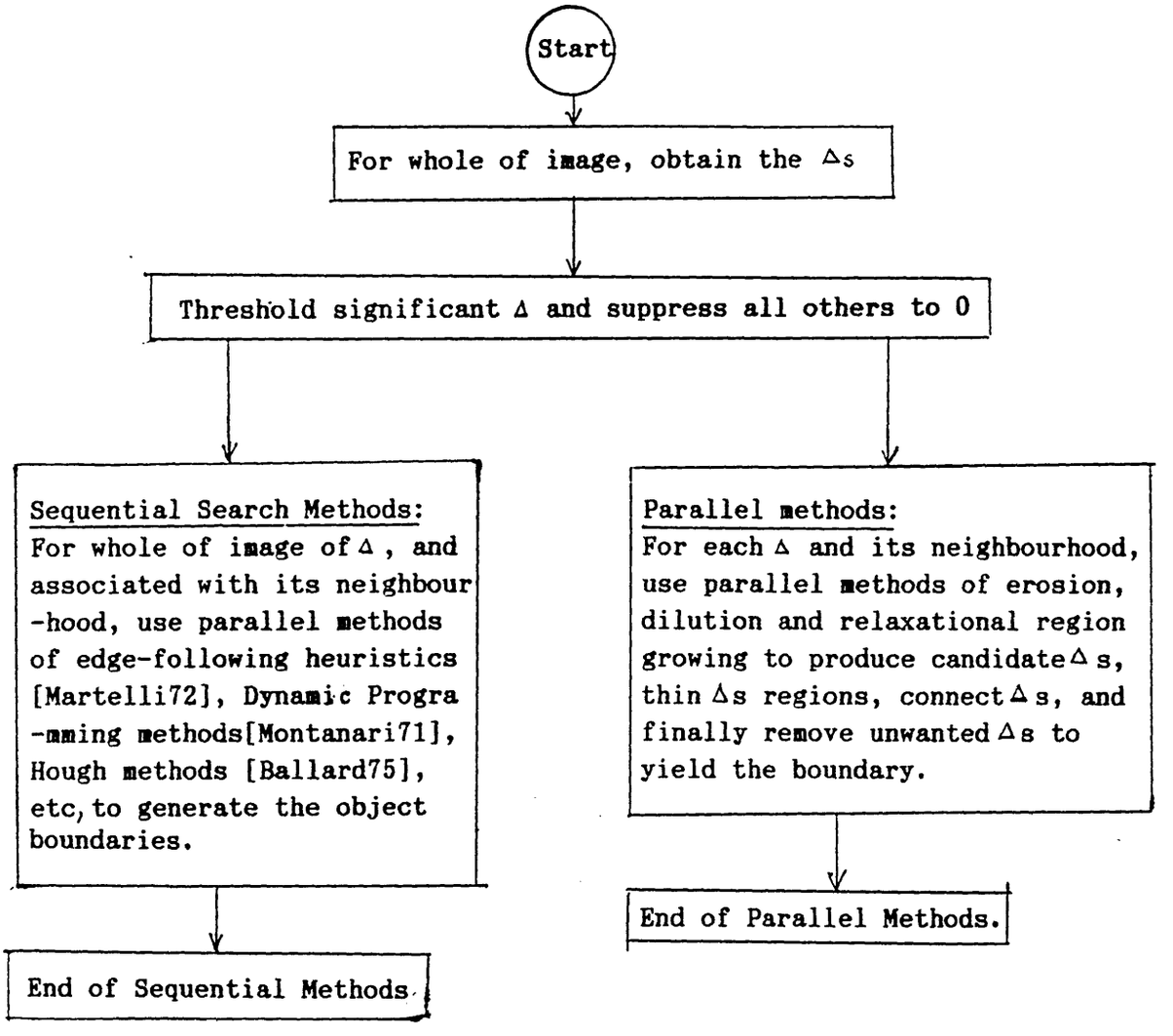
6.2 Related work in line following.

Methods involving special masks for the detection of vertical or horizontal lines like

$M_h \Rightarrow \begin{bmatrix} -0.5 & 1 & -0.5 \\ -0.5 & 1 & -0.5 \\ -0.5 & 1 & -0.5 \end{bmatrix}$ for horizontal lines are basic methods. Methods are available based on parameter-space transformations like the Hough Transforms which can detect any analytical line describable by a table [Duda72]. The method of Chow and Kaneko [Chow72] for the detection of the left ventricle uses a logarithmic filtering, interframe averaging, region-based dynamic thresholding of grey-level intensities and final extraction of the boundary using the gradient information of these thresholded pixels. Another method due to Lemkin [Lemkin79] uses a *Boundary Trace Transform* which is based on sequential thresholding of 'stable' regions, which show evidence of a boundary. Muerle [Muerle68] used a **region growing** method to perform segmentation of objects in a visually complex scene.

Boundary tracking belongs more to the *sequential methods* rather than the *parallel* methods of image processing [Rosen76]. In sequential methods, the results from each operation determines the next step whereas in parallel methods, the same set of operations are performed for the whole of the image as illustrated in fig. 6.2a.

Fig. 6.2a. A summary classification of parallel and sequential methods of boundary tracking.



In this context, the planning method of Kelly [Kelly] to obtain the boundary of a head could be classified as a sequential method using multiresolutions. Another method that is sequential is that of Martelli [Martelli72] which uses the concept of 'microedges' which are defined as edges if immediate neighbours **have different pixel intensities p(x,y)**

Microedge exists if $p(x,y) \neq p(x+1,y)$ or $p(x,y) \neq p(x,y+1)$.

These microedges are mapped into a graph where the search for the edges becomes a algorithmic-graph cost-indexed search operation which seems to be **serial** though there are **no** semantic inputs[1] in the search.

Montanari [Mont71] used a dynamic programming method to track curves in a noisy environment where a figure of merit is embedded in the curve. This figure of merit contains some specifications on the nature of the curves (e.g straight lines).

One method by Robinson [Robinson77] generates a line only when edge **orientations** fulfil a neighbourhood connectivity criterion. One rule that was used was that an Δ point is part of an edge if its directional neighbours comply with a set of orientation conditions. The results on test images showed very little spurious edge-points.

Grifiths [Grifith73] made use of the profile of the edge and other a–priori information to obtain straight lines and vertices. This seems to be using more information from the edge and the a-priori knowledge can be regarded as some similar semantic information. A sophisticated method based on a strong knowledge-base to find the lines of polyhedral objects is described in [Shirai73]. From the above methods we can note that sequential line following methods fall into two classes:

a) Those methods that tracks line and curves, using higher level knowledge by assuming that the Δ points are already **valid** parts of an edge. The algorithms assign the right edges to the right curves or lines.

b) Those methods [Grifiths, Robinson] that perform do not accept Δ points as valid parts of edges until some conditions are satisfied.

6.3 The ideal semantic-line tracking and the perifovea-guided semantic line tracking.

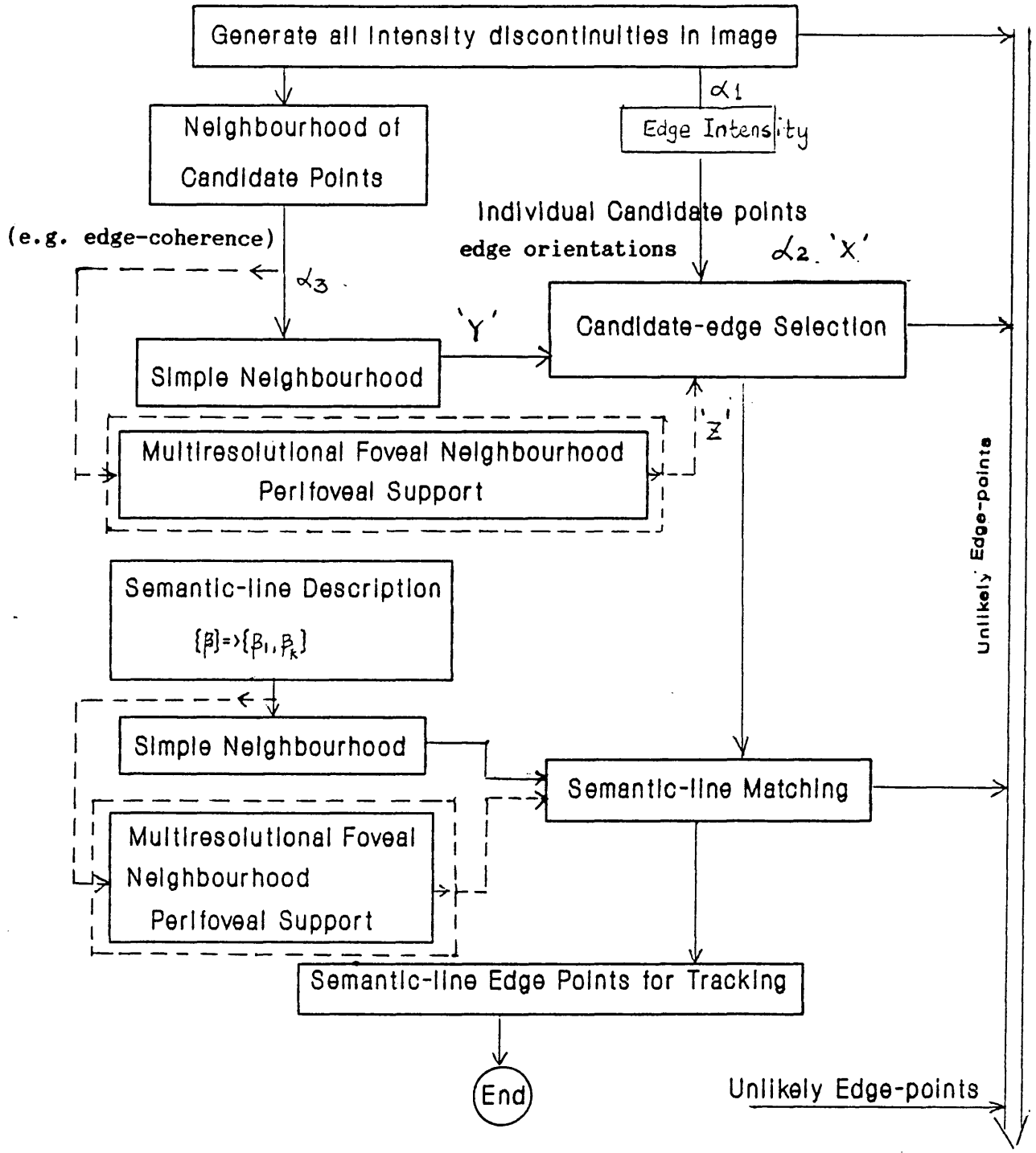
Before we discuss on the investigation of the perifovea-guided semantic line-tracking, it would be good to think about what an ideal semantic line-tracking would look like.

An ideal semantic boundary tracking or any generalised line-tracking method using a sequential search mode is illustrated in fig. 6.3a.

---

1. Semantic inputs must take into account more than the edge-strengths and need to include other possible metrics (e.g edge orientations, edge-neighbourhood statistical modalities, edge profile, etc), which also tend to make the problem very complex.

Fig. 6.3a Ideal Generalised Semantic-line Tracking Algorithm



Note: Perifoveal Support is a result of organising the neighbourhoods in a multiresolutional and hierarchical manner (i.e. near neighbours are more important than distant ones)

Ideally the program has two parts. The first part should be one collecting *evidence* of edges by creating all pixels meeting the set of conditions $\alpha=\{\alpha_1,\alpha_2..\alpha_L\}$ where the αs are information from the ∇ points,[1] e.g. orientations, edge intensity, edge profile, edge contrasts etc.}. A pixel E_i so obtained is not thresholded at this stage but shall be specified by its tuples $E_i=\{\alpha 1, ..\alpha_L\}$. The set of possible edge points $\{E_i\}$ are matched to the inputs of the semantic description β of the line $\{\beta\}=\{\beta_1, \beta_2..\beta_k\}$, where β specifies the semantic conditions for connectivity. The elements of β could be typically

β_1,β_2 = minimum and maximum length of any straight segment in the line,

β_3,β_4 = minimum and maximum curvature in a valid line.

etc. for all other βs.

The sequential search for the semantic line L becomes a mapping of

$$\{\alpha\},\{\beta\} \xrightarrow{a} \{L_s\}$$

This ideal appears easy to formulate but the real task is so daunting that it may sometimes be considered naive to think of this possibility. This can be seen from the fact that $\{\alpha s\}$ specifies low-level vision tasks for which the ambiguities are plentiful while $\{\beta\}$ specifies a higher-level vision task which is not as difficult if low-level tasks are done right. Success of the whole scheme not only depends on the sequential success from low-level to high-level but that relaxational search may be also be needed.

This situation is in fact emphasised in [Horn86] that "least work has been done on the organisation of edge fragments into larger entities, i.e lines and curves".

The inputs to the program has two parts, coarsely equivalent to the lower-vision part and the higher vision part. The lower level vision part generates all pixels which make them candidates for edge points using the set of conditions:

$$\{\alpha_1,\alpha_2,\alpha_n\} \Rightarrow \textbf{candidate edge-points } \{E_i\}$$

where α_s are the full set of elements of the tuples of the ∇ points (orientations θ_e), edge strengths S_e, edge contrasts C_e, profile of edge in X-direction etc.). The edge points E_i so obtained are not thresholded in any way. The full set of edge points $\{E_i\}$ shall be specified by their associated tuple $E_i=\{\alpha_1,\alpha_2,\alpha_L\}$.

The set of edge points for the whole of the image is matched against the semantic descriptions of the boundaries given as tuples in B where

$$\{B\}=\{\beta_1....\beta_k\}$$

each β specifying the connectivity of the edge points $\{E_i\}$. The elements of {B} could be typically,

β_1,β_2 = minimum and maximum straight-line runs of the boundary

---

1. This is to re-emphasise that Δ is a scalar whereas ∇ is a vector.

β_3,β_4 = minimum and maximum curvature of boundary line

β_5,β_6 = minimum and maximum edge-strength difference in adjacent edge-candidates of the edge line.

Finally a boundary line segment is generated if it satisfies the conditions in $\{\alpha s\}$ and $\{\beta s\}$.

6.4 Perifoveal-guided semantic line tracking.

In attempting to use perifovea-guided approach, simplifications will be made so that only three elements from $\{\alpha\}$ and one element from $\{\beta\}$ shall be used.

These are

α_1 = edge strengths

α_2 = edge orientations converted to modulo 8 cardinal directions east, north-east, north, etc. equivalent to 1 to 8.

α_3 = edge activity index (EAI), (α_3 ranges from 0 (low edge- activity) to 1 for (high edge activity), defined as a measure of how likely a ∇ generated when α_1 and α_2 have a finite positive values, are actually true edge-points and not a noise spike. This is based on a measure of the neighbourhood characteristics. Intuitively, if the neighbourhoods show no modality in the αs, then the ∇ is not likely to be in part of a valid edge. Although it is possible to obtain a EAI based on the weighted function of both α_1 and α_2, we will only use a simplified modality measure α_2.[1]

Using these criteria one can say for example,

Edge point valid if $(\alpha_1,\alpha_2,\alpha_3) \Rightarrow (0.7, 2, 0.8)$ i.e.

An edge point exists if the point satisfies the condition {0.7, 2, 0.8} meaning that it has a normalised edge-strength of 0.7, an edge orientation in the modulo-8 direction 2 and an edge-activity index of 0.8, thus showing a strong statistical edge-orientation modality in the neighbourhood.

---

1. In the experiments here, only the edge-orientations (α_2) have been used to measure EAI, defined as a simple measure of how similar the neighbourhood pixel orientations are, with respect to the candidate edge-point in question. (i.e. a simple vote count of the orientations). Hence if the candidate orientation is 2 and the neighbours show a dominant orientation of 6 (2 and 6 directions being orthogonal), then the candidate edge is not part of a line.

Similarly, the set of semantic inputs {βs} shall be the minimum, defined as:-

β_1 = maximum allowable differential in edge strength α_1 between two immediate neighbourhood edge-points.

β_2 = maximum allowable change in orientation between two immediate neighbourhood edge-points. β_2 ranges from 1 to 8.

We can see that β_1 and β_2 are semantic characteristics of the type of line we want to accept as a valid line. For example if we are interested in an object that has only a slight curvature, then β_2 need to be correspondingly specified for low values, perhaps not exceeding 2.

Using α_1 points at (x,y) and (x,y+1), $\overset{\Delta}{,}$ can be a part of a line if it fulfils the conditions

$$\alpha_1(x,y) - \alpha_1(x,y+1) <= \beta_1 \quad ---(1)$$

AND (logical and)

$$\alpha_2(x,y) - \alpha_2(x,y+1) <= \beta_2 \quad ---(2)$$

In this way semantic lines are traced (i.e created) for non-zero Δ points having valid combinations of $(\alpha_1,\alpha_2,\alpha_3)$ resulting in edge points defined only by (α_1,α_2) and linked only if conditions of β_1 and β_2 are simultaneously fulfilled. This process is illustrated in fig. 6.4a.

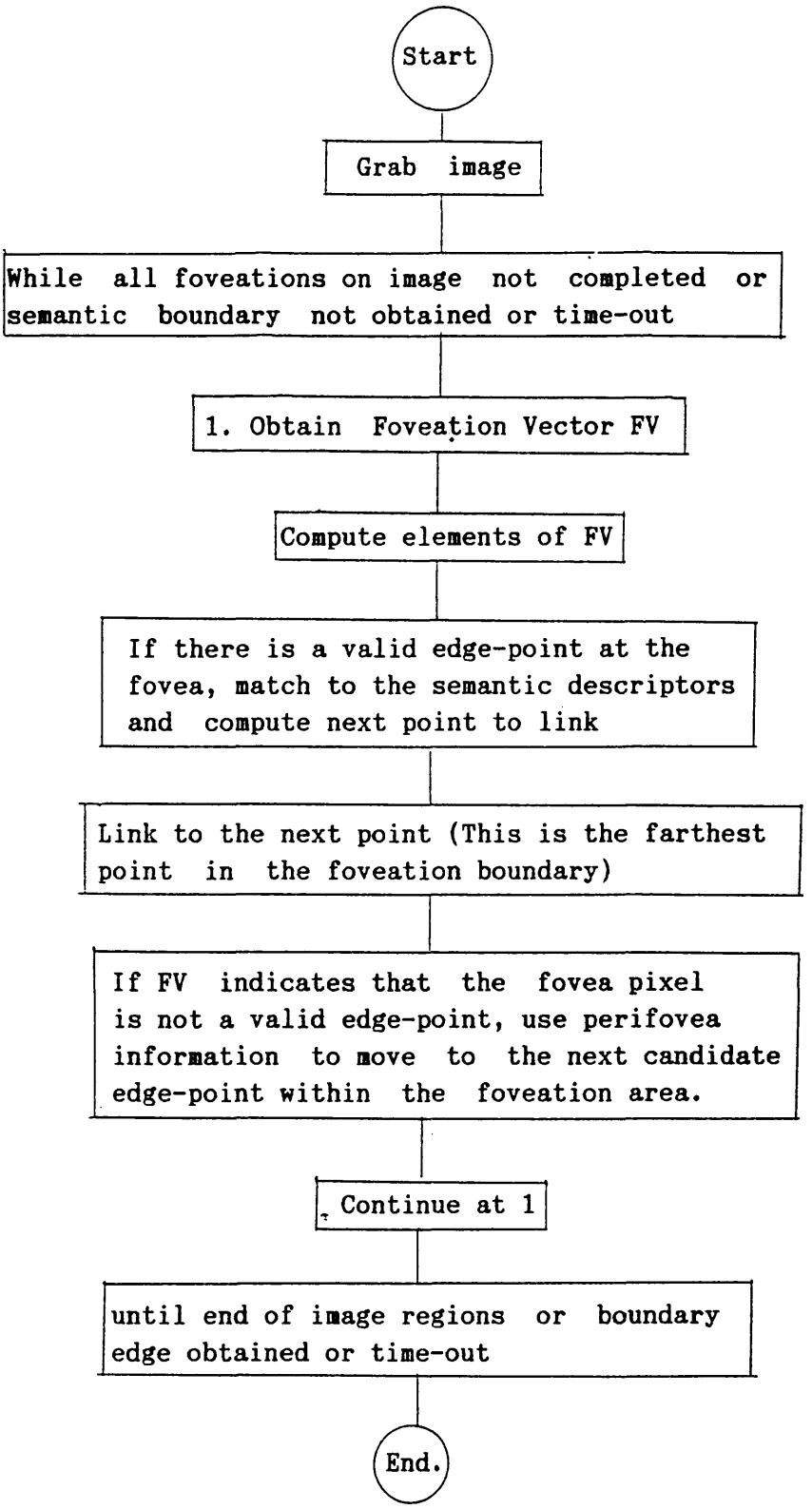
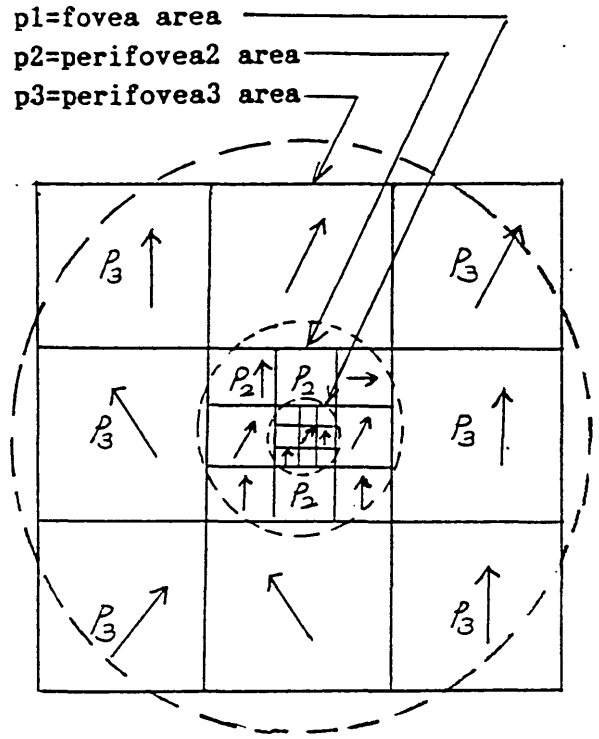


Fig. 6.4a Semantic Edge-linking using perifovea-support.

6.5 Perifovea support in the neighbourhood.

The neighbourhoods of pixels in the line-tracking algorithm here is not a uniform resolution neighbourhood of perhaps 1-unit Euclidean- distance pixels (e.g 3x3 window) but a larger neighbourhood of two perifoveal windows (27x27) illustrated in fig. 6.5a.



Modulo–8 notation for compass points
1=east 2=north-east,..
8=south-east.

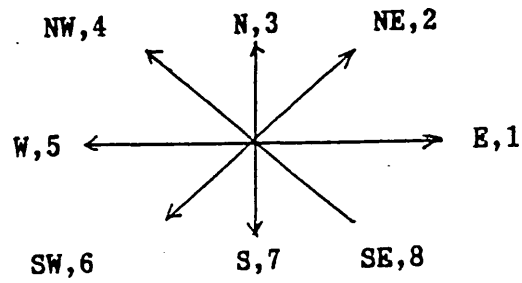
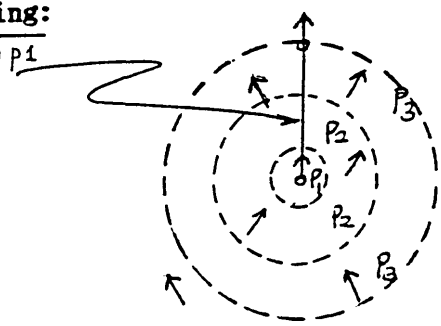


Fig. 6.5a Semantic Edge-linking using perifovea-support.

An Example of conditions for Perifovea supported line-tracking:

If candidate fovea pixel @ p1 has orientation of *north* and the perifovea in the *north* of it also fulfills conditions for support, then connect the fovea pixel to the *farthest* pixel in the perifoveal boundary p3.



- Computation of α_1 and α_2 in perifovea.

Referring to the fig. 6.5b and considering perifovea window 2.

i) Calculation for α_1

Let Y_i be the pixels in the fine-resolution image.

For each Y_i compute the vertical and horizontal components of the ∇ using the Sobel masks

$$\underset{v_i}{\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}} \text{ and } \underset{h_i}{\begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}}$$

yielding $\{v_1 .. v_9\}$ and $\{h_1 ... h_9\}$

The consolidated edge-strength of the single pixel Z is

$$\alpha_1^Z = \left\{ \frac{\sum_{i=1}^{i=9} h_i + \sum_{i=1}^{i=9} v_i}{2 \times i} \right\}$$

ii) Calculation for α_2

By similar consolidation, the orientation of the consolidated pixel in Z is

$$\alpha_2^Z = \tan^{-1} \frac{\sum_{i=1}^{i=9} h_i}{\sum_{i=1}^{i=9} v_i}$$

allowing for orientation normalisation.

iii) Similar computations are done for the perifovea 3 with the difference that i is now 27 instead of 9. **(i.e. 27 pixels to be collapsed into one equivalent perifovea3 pixel)**

In computation, the foveal and perifoveal αs are arranged as a vector constituting the support for the fovea, we call this the **foveation vector**, given by

$$\mathbf{FV} = \left(\alpha_f, \alpha_{p2}, \alpha_{p3}, o, p, q, r, s\right)$$

where

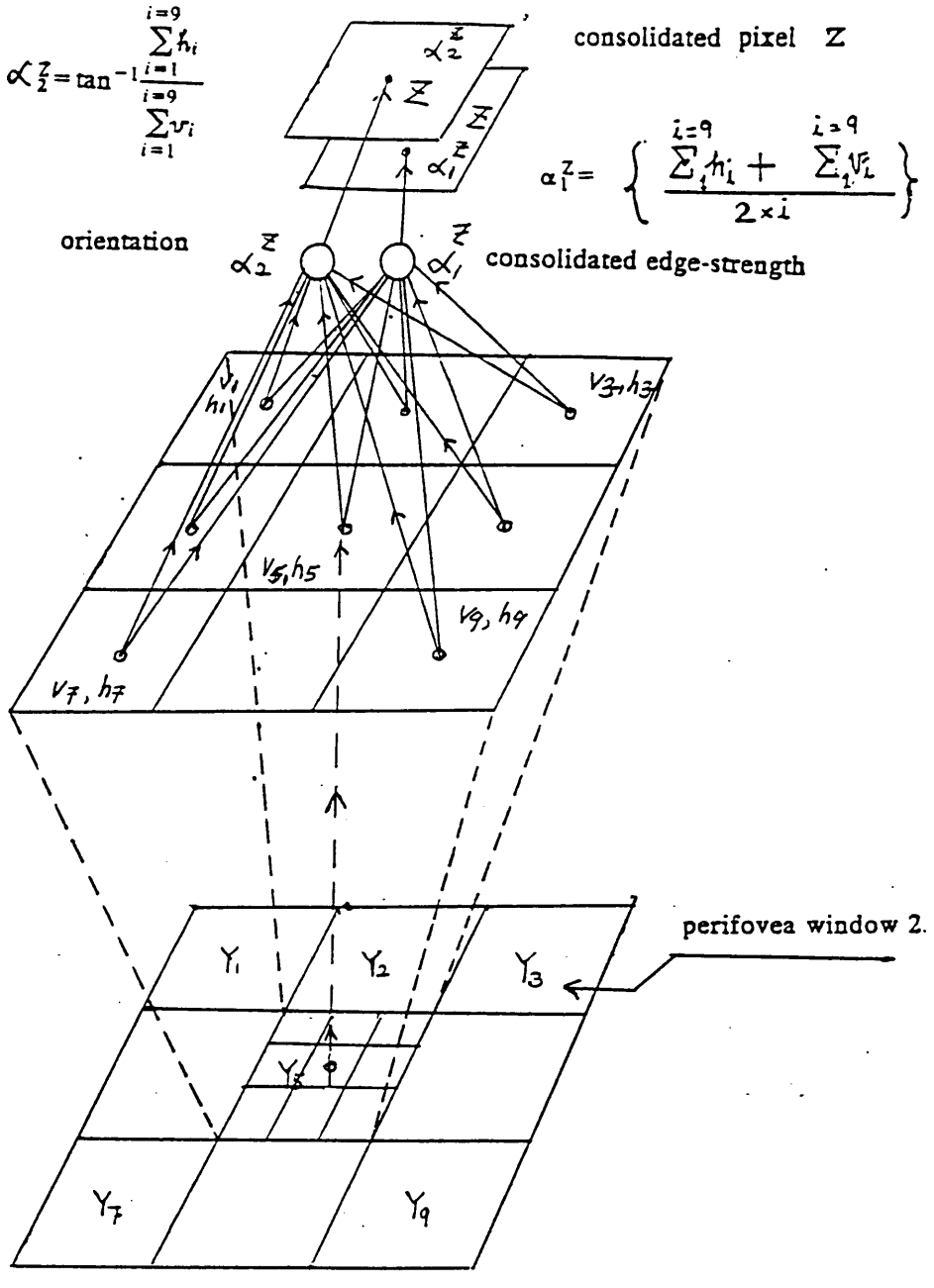
α_f specifies the 9 α_1 for the fovea, **p1**

α_{p2} specifies the 8 α_1 for the perifovea 2, P2

α_{p2} specifies the 8 α_1 for the perifovea 3, P3.

o,p,q,r,s specifies the decisions computed from the αs when matched with the set of semantic inputs in $\{\beta\}$. For example, o is set to 1 if the fovea centre pixel is a valid edge-point, p is the next point to which this pixel should be linked to, etc.

Fig. 6.5b Perifovea support by 3x3 consolidation in 'Z'



The fovea is a regular 3x3 pixel window where each pixel has values (α_1,α_2,α_3). Each coarser-resolution 'pixel' in perifovea window 1 is a consolidation of 9 pixels within itself as indicated by 'Z' in fig. 6.5b.

6.6 Experimental Results

The perifovea-supported algorithm was used to trace the semantic line (i.e. the outer boundary of the object "coin1" of the image fcoin of fig 6.1a. The lighting is specially selected to be *unstructured.*

Fig. 6.6a(i) shows the result of finding the edges where the information from α_1 is only used (as in most conventional edge-finding methods). The main variable that governs the merit of this result is only in the selection of the threshold value, which is done either interactively from visual inspection or from some optimal intensity-threshold methods (e.g histogram-equalisation) which is unlikely to yield semantic object boundaries, if the lighting is unstructured. .

Fig.6.6a(ii) shows the edge-points that satisfy the criteria for $\alpha_1 > 100$ and $\alpha_2 = 4$ (NW) i.e condition x of fig. 6.3a. Note that although boundary edge-points are obtained, the background edges are similarly generated as random background edges do also satisfy conditions for α_1 and α_2.

Fig.6.6a(iii) shows the edges that satisfy the additional conditions for α_3, (edge-intensity profile) where planar objects as is characterised by the flatter coin1 will show a edge profile α_3 that has a steeper rising and trailing gradient than the more diffused edges of the plugs.

In Fig. 6.6b, much of the background edge-points are eliminated by using the perifovea-supported edge-activity index criterion, since random edge orientations of the background do not show a statistical modal distribution unlike object boundaries.

Using perifovea-support to distinguish the background from the object boundaries results in a faster computation time since false triggers to link unlikely candidate-edges are suppressed whenever the perifovea region exhibits a low confidence in the existence of valid boundary-edges as illustrated in fig. 6.6c.

Finally another criterion for boundary edge-linking by including the semantic descriptor for β_4 (maximum allowable curvature of the boudary) was experimented. An earlier naive procedure using only foveal edge-orientations (i.e to say no perifoveal support) show spurious results. Such spuri-

-ious edges can be expected since single-pixel-distance edges are too close to give reliable information on changes in edge-curvatures. The result using conditions for β_4 to be less than 2, changing from the fovea to the perifovea, show more promising results, as in fig. 6.6d.

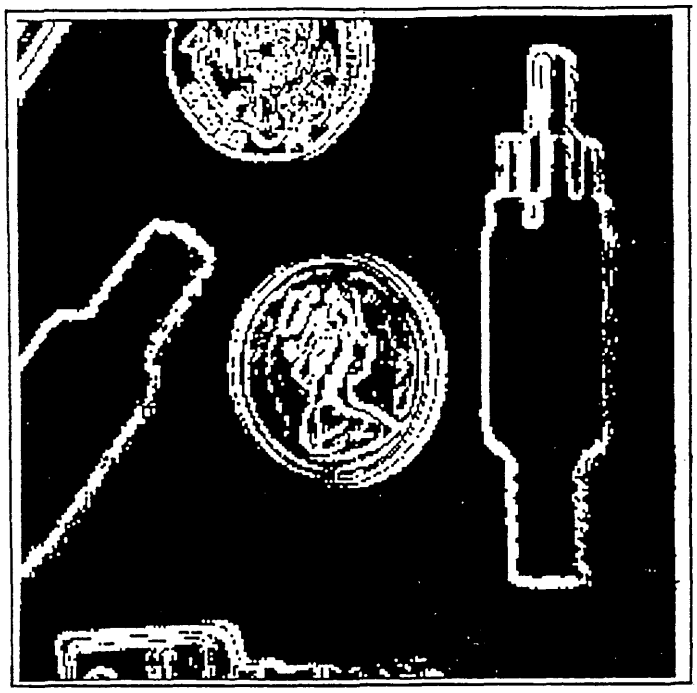


Fig. 6.6a(i). Image of fcoins with edge-strengths thresholded level of 100 grey intensity level (8-bit resolution grey levels)

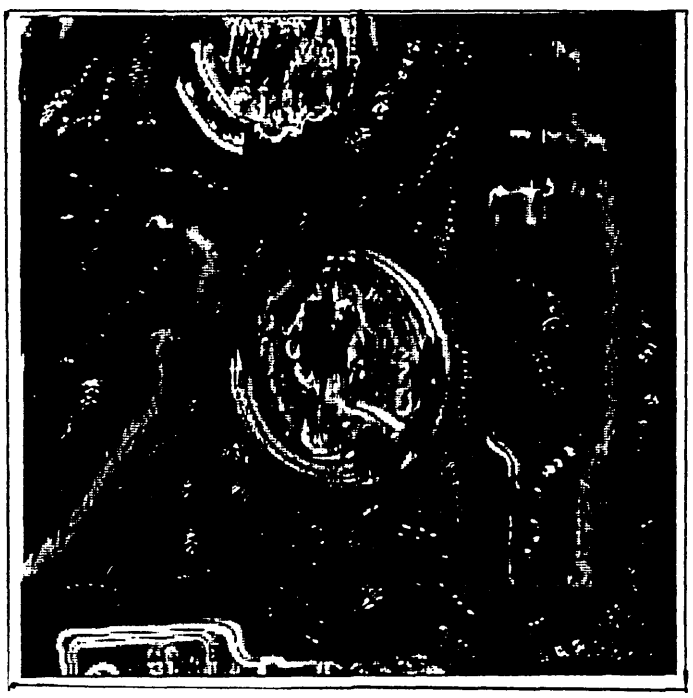


Fig. 6.6a(ii) Image for fcoins satisfying the condition for $\alpha_1 > 100, \alpha_2$ <> 4 (i.e NW)

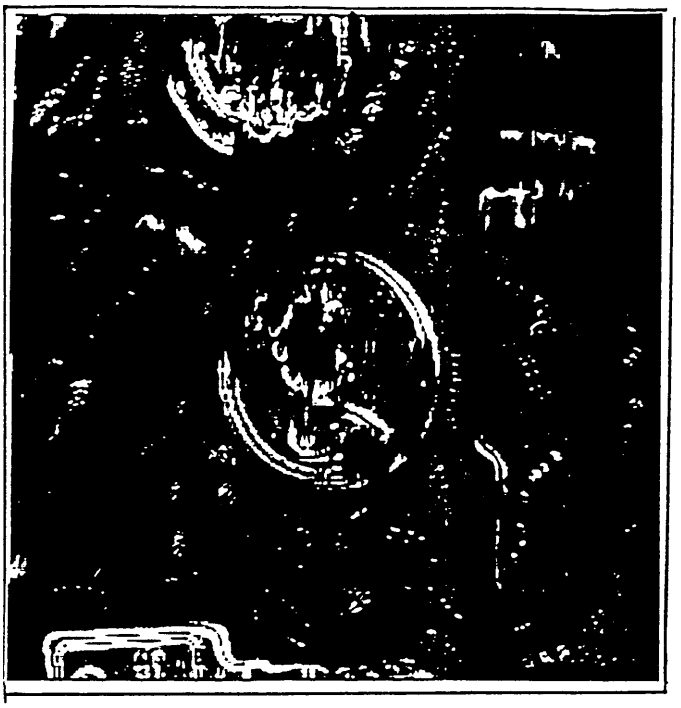


Fig. 6.6a(iii) Image of fcoins satisfying $\alpha_1, \alpha_2, \alpha_3$.

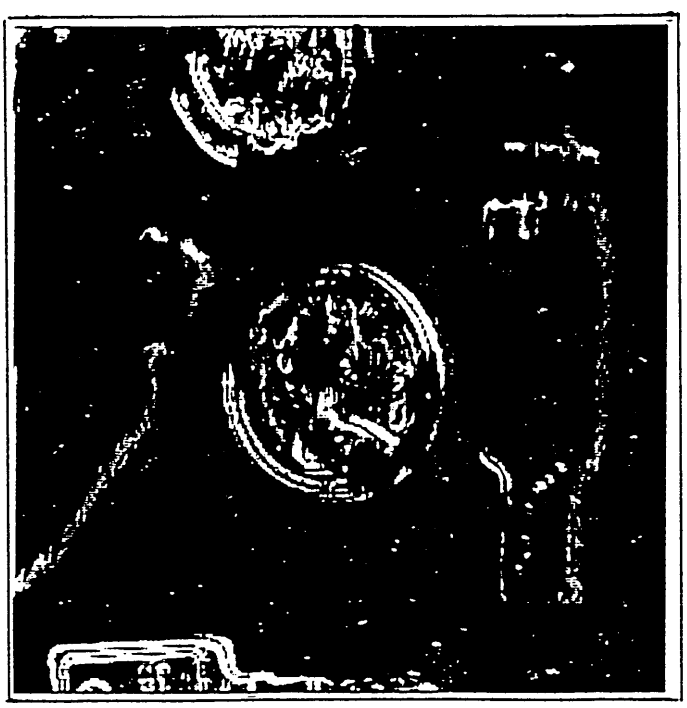


Fig. 6.6b Image of fcoins satisfying conditions of edge-activity index (EAI).

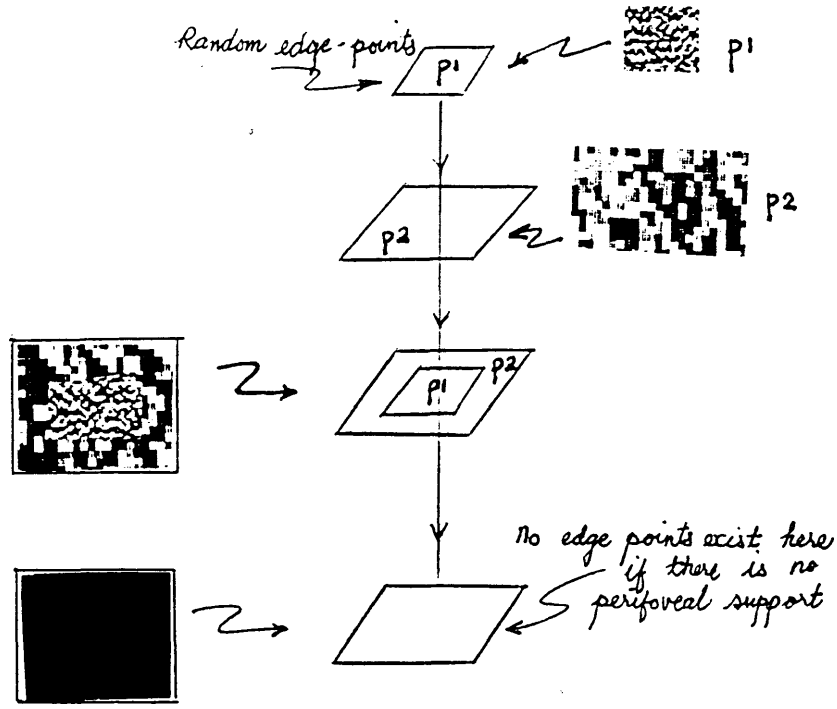


Fig. 6.6c Image of fcoins with elimination of background random edges using perifovea support to identify unlikely region of edge-points.

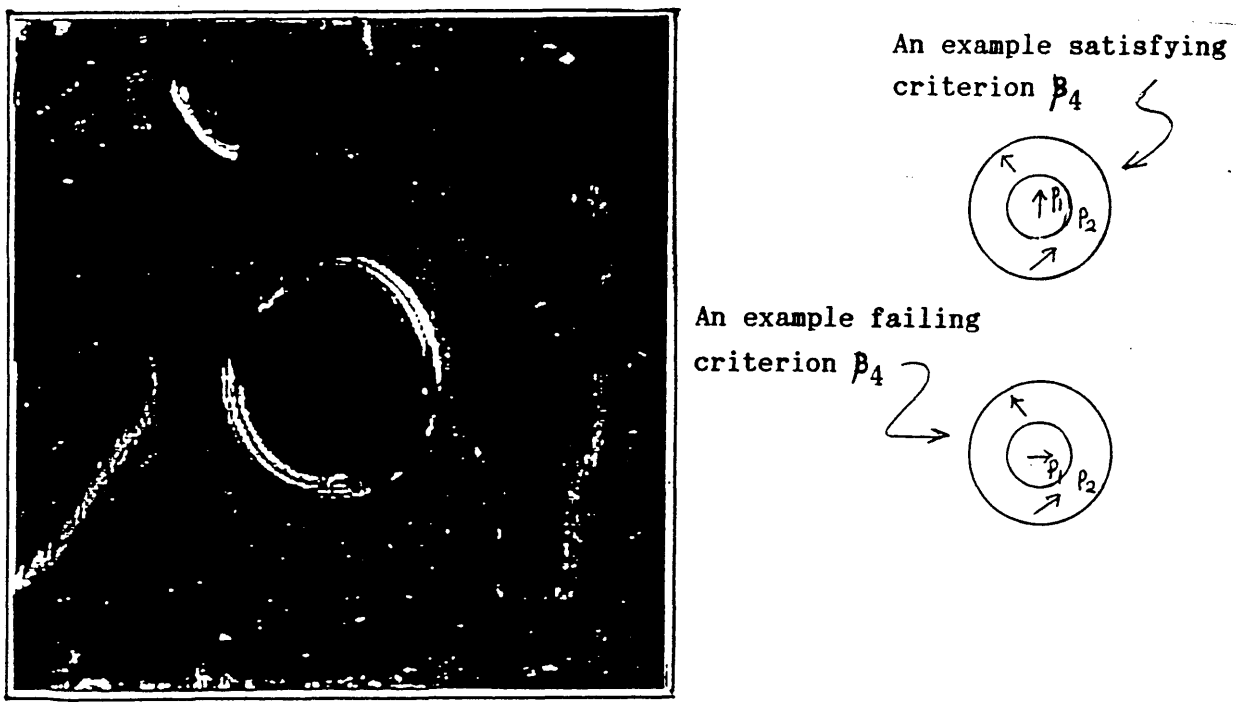


Fig. 6.6d Image of fcoins satisfying semantic-input criterion β_4.

6.7 Limitations in the Perifoveal supported algorithm.

The main limitation of this algorithm was that in actual computation, a long time was needed (approximately 5 minutes of computing time). This was not because of the matching of β but because of the computations to obtain the ∇ for each point and consolidating them in the perifovea. Once a valid edge point is obtained, linking was fast. In any program that required a search, there is a possibility of an infinite loop in the search since a local minima could be reached. This could be overcome by a counter to allow the number of iterative loops so that the program could cease on time-out. This was not provided in the algorithm. In using any perifovea-support it is necessary to establish an order of importance to the information derived from the perifovea 2 and perifovea 3. This is because the consolidated information in the perifovea 3 being further away has been subjected to more smoothing as compared to perifovea 2. The algorithm used here unfortunately assigned equal importance to the information from P2 and P3.

6.8 Summary

In this chapter, an attempt was made to use the perifovea as a support to guide in the tracing of semantic-lines. The importance of semantic edges/lines was discussed in the light of visually complex scenes. The techniques for sequential line tracing were reviewed, demarcating those that use inputs assuming that these are valid edge points and those that hypothesise on intensity-discontinuities until proved to be valid edge-points. The method of perifovea-supported line tracing was used in the second category. The perifoveal support provided a view of 27 x 27 pixels but consolidated at the perifoveal areas so that spurious information is smoothed out. Computation on the line-tracking decisions based on the resulting 25 pixels gave the search a speed gain since the decision variables were factored down by 30.

The results using perifovea support to track the object boundaries have been mixed. It must be pointed out that the primary aim of this endeavour is investigate if additional semantic inputs to a search program to obtain the object boundaries do yield better results. In the course of this, it was found that perifovea-support was necessary to constraint an otherwise exhaustive and falsely-triggered search. In designing the experiments to search for semantic object boundaries, the author has spent considerable time and effort to refine the algorithm. There seems to be many more descriptors that could be used to improve the results. However, at this point it is felt that any further effort may still be short of reliable results in practice since one fundamental fact always bears out. This is the much-discussed ill-posedness of the problem.

Addendum.

Although an initial attempt was made to perform semantic line-tracking using perifoveal support , this was not carried out successfully in the experiments. One reason was that unless the foveation fell near the outer boundary of the coin, all other object boundaries will also start a tracking process. Another reason was that it was very difficult to give a robust and richer semantic description of the coin using geometric properties of the outer boundaries only.

The result in fig. 6.6d where the semantic criterion for β_4 was complied, was not as a result of line-tracking but was a map of all intensity-discontinuities that satisfied the conditions for the multiresolutional $(\alpha_1, \alpha_2, \alpha_3)$ and β_4.

What the results prove is that using perifoveal support on a 27 x 27 pixel multiresolution window, strong intensity-discontinuities that would otherwise be considered as a possible object boundary, if computed at single unit pixel resolution level, would be rightly ignored since there will be no edge-coherence. This means that the background of the scene will not be mistaken to be possible boundaries of objects (the background being a mat-surface, strong intensity-discontinuity region , in an unstructured scene) .

Chapter 7

Multiresolutional Intrinsic Images for Occluded Object Recognition.

Intrinsic Images as a means of obtaining a more complete description of a visual scene was introduced in Chapter 3. The idea,originated from [Barrow78] was that by using models of illumination, reflectivity, surface slope, etc., and computing on the grey-level distribution, we should be able to generate these intrinsic images at *every* point of the 2-D scene. Such intrinsic images can be viewed as a means of deriving 3- dimensional information from a 2-D scene. It is obvious that to derive the intrinsic images for *all* points of the scene will be very time consuming and therefore would need to be adapted if real-time object recognition is the goal.

The methods using multiresolutional foveal images do not aim to process images globally throughout the computation process, as was pointed out in chapter 5. Upon obtaining an area of interest using the GMF, only the computation on a small multiresolutional window called the **Local Multiresolutional Foveal Image** (LMF) would be used to compute the finer details. We will investigate how we can use the LMF to generate some equivalent intrinsic images for occluded object recognition. The methods used here are quite unconventional but it aims to use as much of the grey-level information from a scene as possible,using LMF foveations.

This chapter starts with a concise discussion on the problems of occluded object recognition and introduces the techniques used to solve these problems. Posing the problem of occluded object recognition as a problem solvable by deriving basic 3-D information from the scene, the method of intrinsic images is proposed. However, adapting the intrinsic images to the multiresolutional nature of the LMF, the computations may need to be modified. The procedures and algorithms for an occluded object recognition experiment are described. These results are analysed and compared to two methods that have been proposed by others to solve the same task. The first method is that of using a multiresolutional pyramid model [Neveu86]. The second, and more industrial-like is that of the **local feature focus method** of [Bolles and Cain83].

7.1 Occluded and unoccluded object recognition.

The need to recognise an object that is partially visible or occluded occurs widely in Industry. This situations occur when objects or parts exit from some industrial processes (e.g metal stamping station) in such a manner that these parts are randomly oriented and overlap. This not only creates recognition ambiguities to the occluding object but also to the occluded object. It can be expected that unless means are provided, (e.g. hopper or gates) there is no special reason why parts should not be occluded at some stage of an industrial process.

7.2 The problem of occluded object recognition.

In retrospect, the problem of occluded object recognition in the earlier years of robot vision was actually a problem of trying to recognise occluded objects from binarised features (i.e mainly from shapes or outlines). Even today, in second generation robot vision systems [Rossol83] where processing starts from grey images, this is still by no means a trivial problem. The problem of occluded object recognition posed in this chapter is similar to that in [Neveu86] and will be described in 7.2.1 below:

7.2.1 Description of the experimental scene of occluded objects.

The scene is shown in fig 7.2.1a.

Objects in the scene: 5 planar objects (3 different keys and 2 similar coins).

Background: Any randomly chosen uniform-texture background (in this example, a briefcase).

Illumination: Normal unstructured study-lighting.

Amount of overlap allowed: From light to heavy overlap of the keys only. In this experiment, the coins do not overlap.

Touching of objects allowed?: yes

The problem to solve: To identify the occluding and occluded keys if any.

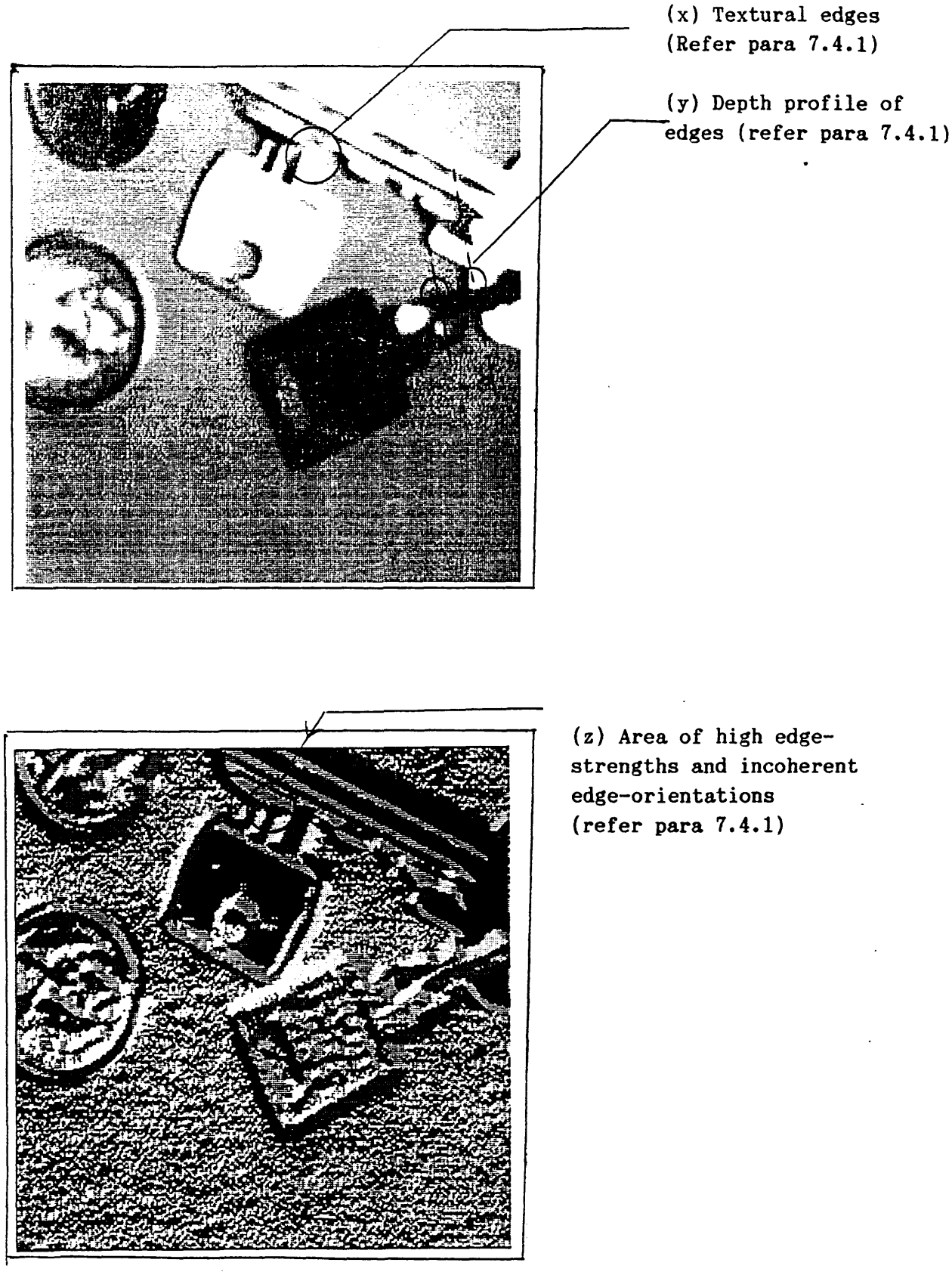
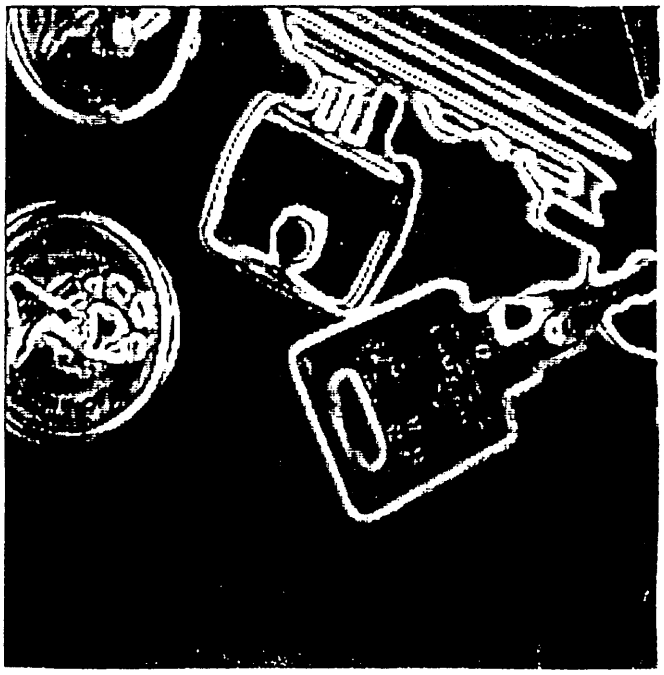


Fig. 7.2.1a The experimental scene for occluded object recognition.



Unthresholded ID

(Intensity-discontinuities)

Thresholded ID

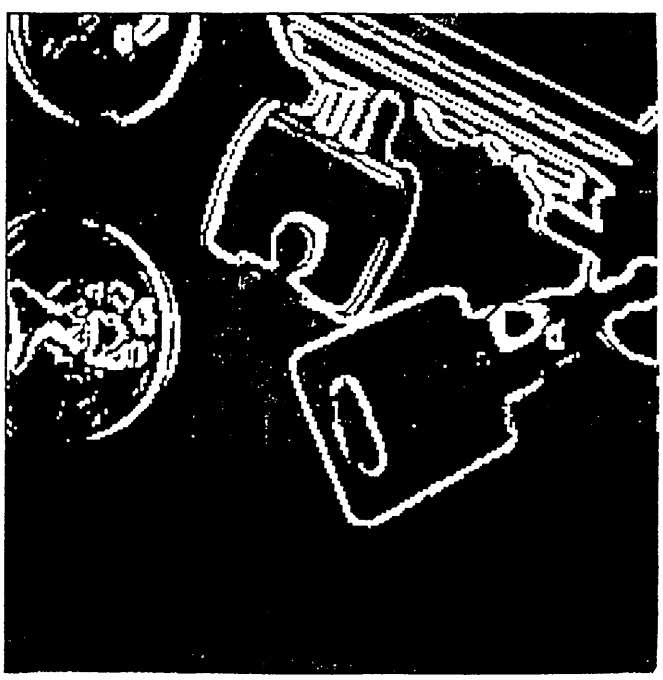


Fig. 7.2.1 a(ii) Intensity-discontinuity (ID) points of experimental Scene.

7.2.2 Minimum set of intrinsic images

To solve the problem of occluded object recognition as posed in paragraph 7.2.1 we will try a minimum set based on **texture** and **edge** characteristics. Therefore we will adapt the LMF to compute the following:-

i) Textural Edges.

ii) Depth from edge-profile

iii) Incoherent **edge-angles** in areas of strong **edge- strengths.**

Textural edges are edges between regions of differing textures. This is a useful method to identify an object from its background. Thus a production rule for occlusion may be stated as:

> "if an edge results from textures of two surfaces, such that one of the textures is not of the background, then occlusion of an object by another object is detected"

This is indicated by the area marked (X) in fig. 7.2.1a.

7.3 Survey of techniques for occluded object recognition.

Complex 3-D occluded object recognition is extremely difficult to solve. An example is the classical **bin of parts** problem. We will only study the case for 2-D objects (actually this means the 3rd dimension is small compared to the remaining two) that are overlapping or occluded so that uncertainties of the objects are in the plane of the camera.

Object recognition of unoccluded 2-d objects that have been pre-engineered is not too big a problem even with binary robot vision. Essentially, and the most common method employed in industry is to match the **geometric moments** (area, compactness etc.) as these can be obtained quite easily [Pugh83]. However, if objects are occluded, even this relatively universal method of moment matching fails as there is no simple way of distinguishing objects from additional moments that have been created due to either *touching* or *overlapping* or *partial views*.

Alternative faster methods that do not rely on moments, could use a model reference to reconstruct the object contours [Hattich83] or to use the **local features** of the objects (e.g Euler numbers, corners, and some extended features like the object boundary segment). The latter method is that of [Bolles83] that generates all the local features of the objects for matching.

Structured lighting methods involving the use of a specially *structured* illumination system such as a plane of light e.g. laser illumination [Agin73,CONSIGHT], slit of light [Shirai71], tapered light beam [Wei and Gini83] could be used to detect object overlap by inspecting the profile of the illumination. Although they seem easy to use, the problem is that of finding the illumination so that computation of the beam can start. Time of flight ranging using laser [Nitzan77], ultrasound [Waag and Gremiak76] or **photometric stereo** [Ikeuchi and Horn81] have been proposed to obtain a range map which of course could resolve occlusion ambiguities at some cost. Methods using passive stereoscopy can be used to produce the disparity map which encodes the range information in the occlusion, though these are relatively difficult to implement.

Methods to derive 3-D information from 2-D, usually referred to as the *recovery of x from y* (shape from shading, etc.) are theoretically possible. Other methods of occluded object recognition using global search for a linear or curve features using the Hough transformations in non-hierarchical ways or as a hierarchical search in a pyramidal image [Neveu86] have been also proposed.

From the above representative methods used for object recognition which have extensions of them being able to resolve occlusion, we can see that they belong to two groups. The first group tries to derive 3-D information from the scene and the second group tries to use the available binary features in the best way to resolve occlusion ambiguities. However, those methods that try to gather

3-D information are not widely used in industry due mainly to computational efforts.

If it is possible to use more than one source of information concerning the objects (surfaces, edges, reflectance, textures etc.) in as short a time as possible, it may be possible to use them for occluded object recognition. Such a method of deriving and *combining* intrinsic images information using the local multiresolutional foveal image is the pursuit of the later parts of the chapter.

7.4 Adapting intrinsic images to detect occlusion.

There are many intrinsic images that one can generate from a scene. The problem is to identify a possible *minimum and easily computable* set that will be just essential to perform occluded object recognition. From the theory of Intrinsic Images, we have many possibilities that include:

a) Shape from Shading

From optical physics we know that the amount of light reflected from a surface depends on the angles between the incident and emergent illumination at the point of illumination defined by a reflectivity function $r(\theta_i,\theta_e)$. Assuming some regularities, (smoothness, constant illumination etc.), the surface normal can be inferred by measuring the change in brightness at two points of the surface due to change in the angles θ in $r(\theta_i,\theta_e)$. Thus an occluded surface may have its surface orientation changed more abruptly due to its being occluded by another object which has a different surface orientation.

b) Shape from Edges or surface contours. As was pointed out in chapter 6, the so-called edge operators are essentially *intensity-discontinuity* detector. The outputs of these operators may not be **real edges** in the true sense of the word "**edge**" which normally means the outer boundary of an object. If we gather more information from the **vicinity** of these **edges**, we can infer about those edges that are

i) edges of shadows (usually not sharp changes)

ii) edges due to change in surface reflectivity (i.e different surface material)

iii) depth information from edges.

c) Shape from Surface

i) change in surface orientation with respect to the illumination (e.g convex surfaces have a higher centre brightness)

ii) change in range of surface (as a result of occluding or being occluded)

d) Shape from Texture

i) Change in surface orientations can be inferred from the computations of the *texture anisotropies* in the *texture microedges* of the surfaces.

ii) Contours can be inferred from change in textures, giving rise to **textured edges**.

The above computations are non-trivial even making some broad assumptions on the problem.

The problem associated with texture is to decide what the simplest descriptor for texture is, for the problem at hand. Consequently, a simple statistical texture descriptor for roughness R, based on the nth moment $M_n(z_i)$ of the pdf of the grey-level intensities $p(z_i)$

$$M_n(z_{(i)}) = \sum_{i=1}^{L} (z_i - m)^n p(z_i)$$

where

$$m = \sum_{i=1}^{L} z_i p(z_i)$$

being the mean, and R (normalised) for the 2nd moment $\sigma^2(z)$, being the variance, resulting in

$$R = 1 - \frac{1}{1+\rho^2(z)}$$ **is used.**

Thus R ranges from 0 for smooth surfaces (all z_i being equivalued) to 1 for rough surfaces. The value R is weighted with the mean to give the overall descriptor of texture T described as

$$T = \alpha m + \beta R \quad \text{----(7.4.1)}$$

where α and β are the weights for m and R.

- **Inferring occlusion from edge-profile.**

The edge-profile shall be used in the following way:

"If the edge profile P_s of the same object changes, indicating there is an increasing depth in the edge (as compared to the reference edge- profile on the flat background), the object is likely to be occluding another object".

This is reasoned out from the fact that an occluding object will be on top of another object and so the edge depth will change. A simple definition of edge-profile is illustrated below (fig. 7.4.1c).

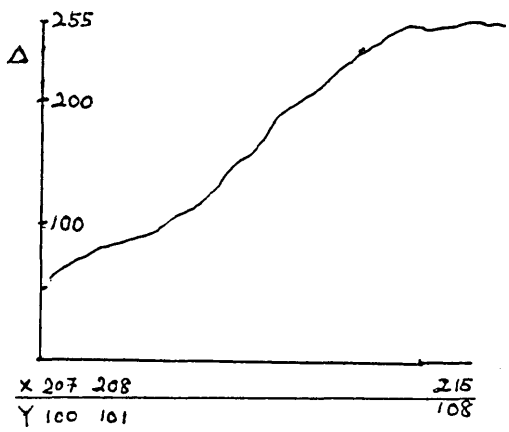
Fig. 7.4.1c Illustration of changing edge-strength profile across object boundary due to depth change.

Unthresholded edge-strength profile at region y1 of scene fig. 7.2.1a.

x pixel coordinate

y	207	208	209	210	211	212	213	214	215	216	217	218	219
100	54	18	20	34	16	18	20	10	28	36	20	12	12
101	40	20	26	32	10	10	10	12	6	8	34	16	20
102	16	10	34	30	8	6	6	2	10	50	24	38	8
103	20	14	10	24	4	14	14	10	12	46	56	20	16
104	24	8	32	70	50	46	46	46	36	52	74	52	48
105	18	40	116	122	114	132	128	122	146	124	124	122	144
106	78	144	196	162	212	212	216	194	222	238	168	198	224
107	166	248	255	252	255	255	255	255	248	255	254	252	242
108	242	255	255	255	255	250	222	234	252	246	228	216	212
109	255	255	244	220	190	118	92	94	158	150	134	102	112
110	228	164	102	98	60	16	42	12	52	40	64	10	34
111	76	58	64	116	106	78	90	40	60	12	66	26	50
112	236	184	158	170	158	110	98	60	44	32	78	20	50
113	255	255	222	184	168	144	120	70	8	16	38	24	60
114	255	255	255	214	188	156	138	102	50	10	30	72	52
115	255	255	238	224	172	132	128	74	60	42	30	24	54

Profile of edge-strength at region y2

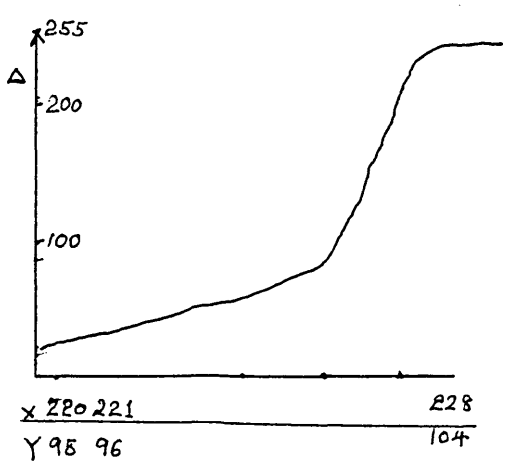


Unthresholded edge-strength profile of region y2 of of scene fig. 7.1.2a

x pixel coordinate

y	220	221	222	223	224	225	226	227	228	229	210	211	212
95	16	24	32	52	98	144	244	255	202	255	255	255	255
96	2	12	40	32	104	170	252	250	192	146	255	255	255
97	6	16	44	30	106	190	254	255	206	50	255	255	255
98	14	4	30	44	122	196	255	255	174	148	255	255	255
99	18	6	14	54	138	212	255	255	138	255	255	255	206
100	12	10	46	56	122	214	255	255	134	255	255	255	90
101	20	20	68	60	124	248	255	242	120	255	255	255	66
102	8	38	86	104	182	255	255	238	114	255	255	255	150
103	16	96	134	184	240	255	255	224	98	255	255	255	255
104	48	158	182	246	255	255	255	174	64	255	255	255	255
105	144	224	254	255	255	255	210	92	20	255	255	255	255
106	224	255	255	255	255	224	102	26	30	168	255	255	255
107	242	255	255	254	208	80	22	54	110	150	166	222	206
108	212	206	202	142	68	62	64	104	134	96	56	88	90
109	112	100	88	50	60	88	112	96	76	4	38	62	46
110	34	26	28	48	82	104	78	68	14	8	22	32	10

Profile of edge-strength at region y1



The edge-profile is derived from the unthresholded edge strengths s_e by using the histogram of its **edge strengths.**

The production rule to detect abnormal edge depth is:

> "If the edge-depth of an object and its background only, exceeds the threshold for its model edge-depth on a flat background surface, then the object is likely to be occluding". This is indicated by the location marked (Y) in fig.7.2.1a

We know that if two objects occlude, then the information from the **edge angles** at the point of occlusion will generally be such that they show the existence of strong **edge-strengths** at the point of occlusion. However, the associated **edge-angles** will not be **coherent** in the sense that the edge angles at the point of occlusion will change abruptly indicating that they are **not** the edge-angles of the same object.

The third production rule to detect occlusion may be stated as:

> "If there are strong edge-strengths in a region, i.e $\Sigma(s_1,s_2,s_n)/n >$ threshold s_t, such that their edge angles θs are not coherent i.e the variance in the edge-angles exceed a threshold
>
> $$\sigma^2(\theta_1 \ldots \theta_n) > \tau_\theta$$
>
> then an area of occlusion is likely to exist". Such an area is indicated as (Z) in fig. 7.2.1a.

7.5 Local multiresolutional foveal (LMF) images as Intrinsic Images.

In the method of intrinsic images due to Barrow, uniform resolution images were proposed. On problem that has to be solved is how best to adapt the multiresolutional nature of the LMF to generate the 3 minimum set of intrinsic images in 7.4.1. It turns out that for the case of textures where we only use a simple description of textures based on lower-order moments of its grey level probability distribution functions, namely variance and mean, the linear properties of this texture description does not change with resolution change.

In the case of the edge profile, we can expect that the reduced resolution at the periphery of the LMF to provide an advantage that spurious variations in the edge profile will be smoothed out.

The same applies to the edge-strengths and edge-angles, as spurious edge-strengths s_n and edge-angles θ_n at the periphery of the LMF will be smoothened out so that the information tends to be more reliable.

7.6 Algorithms.

The algorithms for the whole process of generating the intrinsic images and the search algorithm using perifovea support is shown in fig 7.6.

The grey-level of the image of the scene is first acquired. A random foveation is first attempted using the smaller window structured as a local multiresolutional foveal mask of 27 by 27 pixels as in section 6.5.

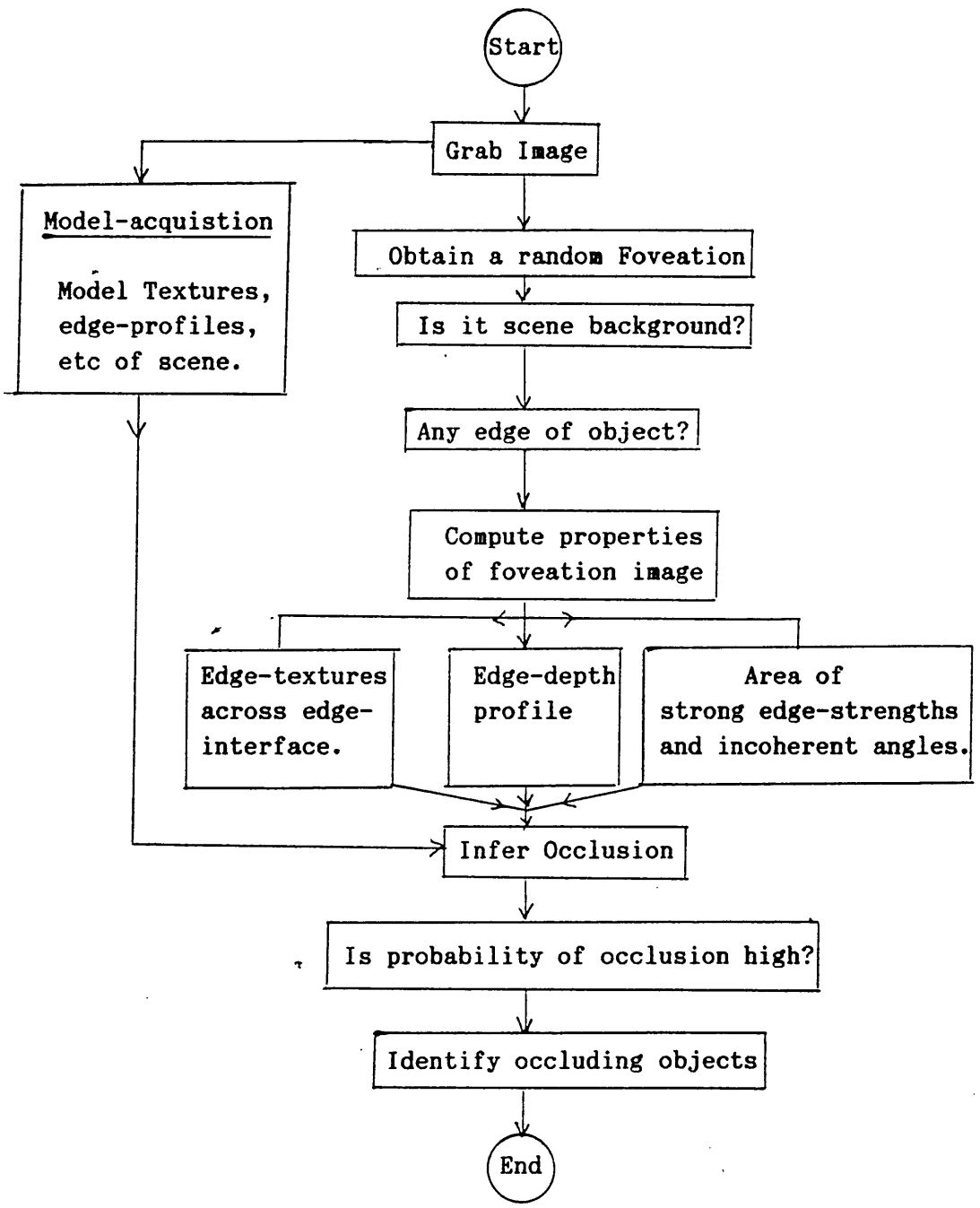
The aim of this foveation is to decide if the foveation image is solely a scene background or it also has some evidence of being part of an object. This is decided upon by generating the minimum set of the three intrinsic images. However from experiments it was found that sufficient confidence can be obtained from just the texture measure and the thresholded edge strengths, without regarding even the edge orientations.

Further foveations will be attempted if this foveation yields the subimage as belonging to the background. However, unlike the first foveation which is purely random and unguided, the subsequent foveation will make use of the information from the intrinsic images generated.

For example, if the foveation yields a higher distribution of edge-discontinuities of more uniform edge-discontinuity orientations within a part of the 27 x 27 pixels, then the centre of mask for the next foveation will be chosen to coincide with this part. This process continues until a foveation on a part of an object is obtained.

On obtaining a foveation which has a high probability of being part of an object (region of high edge strengths and coherent edge orientations), then this region becomes the new local focus point, from where further guided searches for conditions of object occlusion using the three production rules explained in section 7.4.

Fig. 7.6 Algorithm for Perifovea-supported search and generation of intrinsic images for object recognition of occluded images.



7.7 Experimental Results

Using the proposed algorthm of fig. 7.6 to infer the occluding/occluded objects, the results that have been computed from the test image are discussed below:-

The method using edge-textures seems to provide reasonably consistent results. This is because in this case, certain textures are readily identifiable, especially that of the background. Since it is necessary to compare two textural regions across a strong edge-inteface, the program had to first look for an area containing strong edge-lines, before it could compute the textures. The area in fig. 7.7 b was one of the areas detected and confirmed to contain object occlusion.

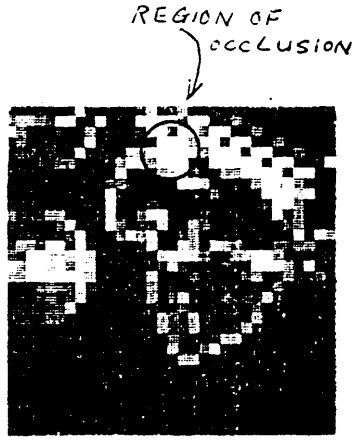
The criterion using edge-depth profile is only useful if an object edge and its *background* is obtained in a single foveation. However, it required a trace of the edge-line of the object, which turned out to be as labourious as the boudary-tracking algorithm used in chapter 6. The gradient of the edge-depth change along the edge-line turned out to be difficult to be used as a reliable model-input to the program. This is compounded by the fact that different poses of the same object with respect to the lighting produced different edge-depth profiles. This criterion turned out to be nearly unusable as a criterion for occluded object inference and was abandoned.

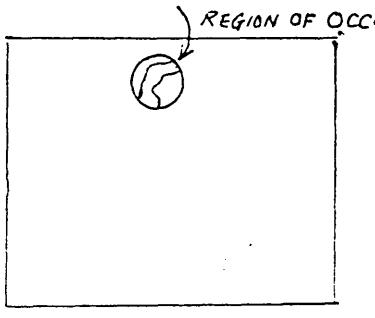
The use of the third criterion, i.e. strong edge-strengths but incoherent orientations was more useful as a means to detect inter-object occlusion. This can be seen in table 7.7b where the two keys mutually overlap. This region shows a high degree of edge-strengths (denoted by non-zero entries (0), since it was necessary to convert orientations to equivalent grey-levels for visual display purposes). Within a small area, denoted by the subimage 'za', the edge-orientations are incoherent since the orientations change rapidly across pixels in all directions. On inspection of the image, it turned out that this region had the bigger brass key occluding the smaller aluminium alloy key.

Fig. 7.7b A region of object occlusion using edge-texture of eqn. 7.4.1
(3 regions of distinct model-reference textures of .4, .6, .8 are detected in this region)

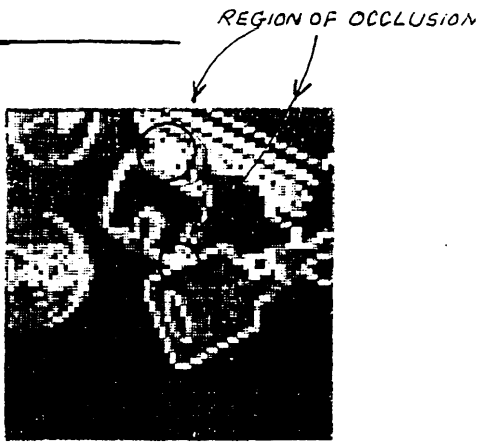
Pixel x coordinate

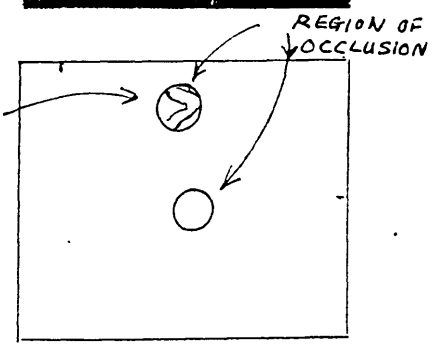
y	132	133	134	135	136	137	138	139	140	141	142	143	144
20	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8
21	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8
22	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8
23	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8
24	.8	.8	.8	.6	.6	.6	.6	.6	.6	.6	.8	.8	.8
25	.8	.8	.6	.6	.6	.6	.6	.6	.6	.4	.4	.4	.4
26	.8	.6	.6	.6	.6	.6	.6	.6	.4	.4	.4	.4	.4
27	.8	.6	.6	.6	.6	.6	.6	.6	.4	.4	.4	.4	.4
28	.8	.6	.6	.6	.6	.6	.6	.6	.4	.4	.4	.4	.4
29	.8	.6	.6	.6	.6	.6	.6	.4	.4	.4	.4	.4	.4
30	.8	.6	.6	.6	.6	.6	.4	.4	.4	.4	.4	.4	.4
31	.8	.6	.6	.6	.6	.6	.4	.4	.4	.4	.4	.4	.4
32	.8	.6	.6	.6	.6	.6	.4	.4	.4	.4	.4	.4	.4
33	.8	.6	.6	.6	.6	.6	.4	.4	.4	.4	.4	.4	.4
34	.8	.6	.6	.6	.6	.6	.6	.6	.4	.4	.4	.4	.4
35	.8	.6	.6	.6	.6	.6	.6	.6	.6	.4	.4	.4	.4
36	.4	.6	.6	.6	.6	.6	.6	.6	.6	.4	.4	.4	.4
37	.4	.6	.6	.6	.6	.6	.6	.6	.6	.4	.4	.4	.4
38	.4	.4	.6	.6	.6	.6	.6	.6	.6	.4	.4	.4	.4





y	132	133	134	135	136	147	148	149	150	151	152	153	154
20	78	84	112	88	120	206	255	255	255	255	122	202	255
21	100	80	52	22	74	74	150	230	255	255	255	220	122
22	72	158	162	102	26	68	90	74	186	255	255	255	255
23	36	50	118	178	122	58	56	76	56	96	154	228	255
24	90	30	12	68	104	104	150	102	50	50	10	66	152
25	255	134	98	104	58	104	124	162	168	106	60	34	56
26	255	255	252	122	80	62	46	62	140	176	86	34	98
27	255	255	255	255	244	138	52	32	72	82	104	80	96
28	255	255	255	255	255	255	156	82	60	78	132	128	102
29	132	168	255	255	255	255	255	226	70	62	120	112	146
30	255	255	255	255	182	255	255	255	184	50	80	196	255
31	255	255	255	255	214	184	162	188	170	154	196	198	250
32	255	255	255	255	255	206	208	120	48	192	255	255	255
33	255	255	255	255	228	146	80	12	112	255	255	255	202
34	96	255	255	240	142	60	112	82	188	255	255	236	70
35	130	255	255	255	226	160	176	190	255	255	255	144	12
36	254	255	255	255	146	128	136	255	255	255	240	62	0
37	255	255	255	255	100	126	78	255	255	255	144	10	0
38	255	255	255	210	84	110	148	255	255	255	86	0	0



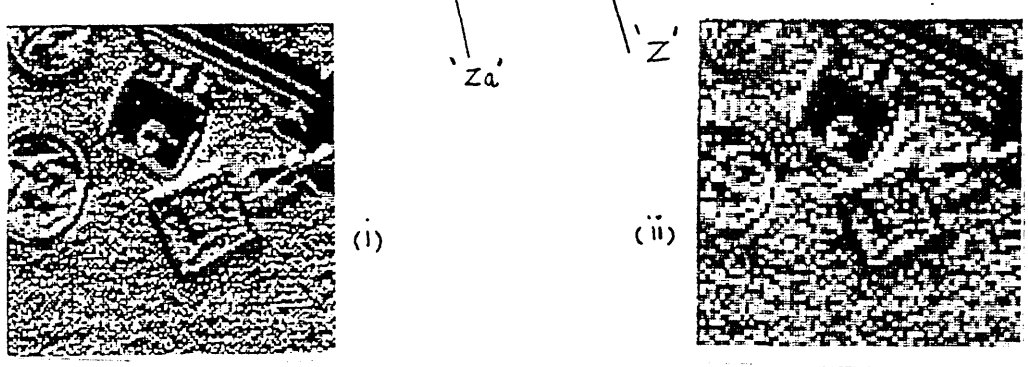


Edge-strengths in the region showing relatively strong edges.

Table 7.7b Region 'Z' of scene 7.2.1a showing a region of inter-object occlusion. (satisfying criterion 3)

Note: The entries in the table are edge-orientations of the 8 directions but scaled to increments of 30 to enable computer image printout as equivalent grey levels. Therefore 30 represents orientation 1 or 45 degrees, 60 represents orientation 2 or 90 degrees, etc. This region of high edge-stengths can be indicated by the absence of 0 entries, which is used to represent pixels having very low edge-strengths, whose edge-orientations can be ignored. Incoherent edge-orientations can be seen from the abundance of varying orientations in a typical subimage area 'za'.

y \ x	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143
20	150	150	150	150	150	180	90	150	240	30	60	90	30	150	150	240	30	60	60	60	60	60
21	150	150	150	150	150	180	150	240	150	240	30	30	60	60	30	240	240	90	60	30	30	30
22	150	150	150	150	150	150	150	150	150	150	240	180	90	60	60	60	60	180	240	240	30	30
23	240	150	150	150	150	150	150	150	150	150	150	150	180	150	180	60	60	60	60	0	240	240
24	30	150	180	180	150	150	150	150	150	150	150	150	150	150	180	150	0	30	30	60	30	30	240
25	30	30	150	150	150	150	150	150	150	150	150	150	150	150	150	150	150	240	240	0	30	60	60
26	30	30	60	90	90	180	150	150	150	150	150	150	150	150	150	150	150	150	150	150	240	30	60
27	30	60	90	90	90	90	60	90	240	150	150	150	150	150	150	150	150	150	150	150	150	150	240
28	60	90	90	90	60	60	60	60	30	240	0	150	150	150	150	150	150	150	150	150	150	150	240
29	90	90	90	90	60	60	60	60	30	30	60	90	90	90	180	150	150	150	150	150	150	150	150
30	90	90	90	90	60	30	30	30	30	60	90	90	90	60	60	90	90	90	180	150	180	150	150
31	90	90	90	90	30	240	240	30	30	90	90	90	90	60	60	60	90	90	90	90	180	150	150
32	90	90	90	240	240	240	240	240	90	90	90	90	90	60	60	60	60	60	90	90	90	180	150
33	90	90	90	240	240	240	240	240	90	90	90	90	90	60	30	30	30	60	90	90	90	90	150
34	90	90	30	240	240	240	240	240	90	90	90	90	90	30	30	0	240	240	150	240	150	90	150
35	90	90	240	240	240	240	240	180	90	90	90	90	90	240	240	240	240	240	240	240	180	90	150
36	90	60	240	240	240	240	240	90	90	90	90	90	240	240	240	240	240	240	240	240	180	90	150
37	60	240	240	240	240	240	90	90	90	90	90	90	240	240	240	240	240	150	240	150	90	90	150
38	30	240	240	240	240	240	90	90	90	90	90	90	240	240	240	240	240	240	240	180	90	90	150
39	240	240	240	240	240	90	90	90	90	90	90	240	240	240	240	240	240	240	150	90	90	90	150
40	240	240	240	240	240	90	90	90	90	90	90	240	240	240	240	240	240	240	90	90	90	90	150
41	240	240	240	240	180	90	90	90	90	90	240	240	240	240	240	240	240	150	90	90	90	90	150
42	240	240	240	240	90	90	90	90	90	90	240	240	240	240	240	240	240	180	90	90	90	90	150
43	0	30	30	90	90	90	90	90	90	240	240	240	240	240	240	240	240	0	90	90	90	90	150
44	30	30	60	90	90	90	90	90	90	240	240	240	240	240	240	240	240	90	90	90	90	90	150
45	60	60	60	90	90	90	90	90	0	240	240	240	240	240	240	240	240	90	90	90	90	90	150



Edge Angles computed at foveal resolution in (i) and at perifoveal resolution in (ii).

7.8 Limitations of the LMF intrinsic image approach.

The primary reason for investigating the intrinsic image approach was that of using as much of the grey-level information as possible to derive some simple 3-D information to resolve occlusion. It was reasoned that by combining a few essential intrinsic images, we could detect an occluding and an occluded object. The idea is feasible but the experimentations showed the following major difficulties.

i) The texture measure using the roughness criterion R was not reliable as R changed with angle of lighting. Unless a model is constrained to be in specific locations so that R is consistent, a single model of roughness is not sufficient. Perhaps additional measures of textures based on higher combinations of the moments of the pdf or structural texture descriptions will be necessary.

ii) In using the edge depth, the biggest problem was that of finding the exact position of the edge from where the depth can be inferred. This arises because in the experiment, the criterion for the detection of the edge is that it is an edge of a single object and its background, based on detection of 3 entities simultaneously (object, background texture, edge). Extrapolations would be needed to centre the LMF so that its centre is the centre of the edge. This turns out not to be a easy task as it needs further fine detail search for the maximum of the edge-strength as the centre of the LMF.

iii) The criterion of incoherent edge-angles at points of high edge strength activity is difficult to model, as the variance of the edge-angles is not a sufficient measure of incoherence. Again the problem is that of locating the centre of the LMF window so that we have the right population of the edge angles.

7.9 Comparisons with other methods for occluded object recognition.

It will be noticed from the above experimental results that though the method using multiresolutional computations for intrinsic image generation and perifovea-supported searches seem theoretically possible, there are certain limitations. These limitations will be discussed in the light of comparisons with two approaches that have been proposed to solve similar tasks. It needs to be reiterated that these two methods do make certain assumptions of *structured lighting* and a *segmented background* which is in contrast, not specifically demanded in the multiresolutional approach.

7.9.1 The Multiresolutional Pyramidal Method.

The method of occluded object recognition described in [Neveu86] is based on a model-based matching technique using the coarse-to-fine search properties of a Gaussian pyramid. The boundary segment of the object is stored as a node in a hierarchical graph. A generalised *Hough transform* is used to match the model nodes with regions in the corresponding levels of resolution. Its strength lies in its having adapted the robustness of Hough-space searches by a hierarchical reodering of the object features (edge, holes etc.).

This method assumes that the object is already *segmented* reliably so that the edges and shape are the features sufficient to detect occlusion. In general it does not use any grey level information from the image except that these have been operated upon by a standard edge-mask to obtain edges. It could be considered as a occluded-object recognition technique using *higher-level* information in a *fast hierarchical model of computation.* Though the Hough method is slow, it is robust. However, the Hough method does have limitations [Brown82] which shall not be discussed here.

7.9.2 The local feature focus method.

In [Bolles83], another method for the recognition of occluded workpieces is proposed. The edges of the occluded objects are avail-

able to the program by edge-mask operations. The resulting segmented scene containing entites called *object-features* (edges, corners, and segments of perimeters of objects) are ordered in a manner making use of the clusters of *mutually consistent features* to create certain hypothesis as to the existence of occlusion and hence infer on the occluded-object recognition.

There is an elegant emphasis in the use of *local features*, as opposed to more global ones, so that the feature-selection process could use a *smaller* set of focused-features to infer on the other features. A sophisticated program generates such features off-line, from the models of parts presented to it.

7.9.3 Some Comparisons

We can see that in general the above approaches do not intend to use the grey information in the image except to generate the edges. This of course must assume sufficient *preengineering* so that the edges thus created are semantic edges (i.e edges of the objects).

The use of hierarchical search using reduced-scale images is a useful strategy in [Neveu86]. Hierarchical search is indirectly incorporated in the intrinsic images by selecting a-priori what *features* have the highest occurence. In the multiresolutional fovea method here these features include *intrinsic foveation images* (e.g background) whereas in [Neuveu86] the features are edge features.

The local-feature-focus method needs to build up a complete database of the model during the model-acquisition phase. In contrast, multiresolutional foveal intrinsic images method do not demand as complete a model-base (e.g edge-segment runs).

In both the methods of Neveu and Bolles, a global search of the problem domain is performed. However, in the intrinsic image approach we should theoretically be able to perform a sequential non-global search, but computation difficulties of keeping some equivalent stack of the areas searched (possibility of infinite loops) have forced the author to use a global search, (top-left to bottom-right), this being not the intention of multiresolutional methods. It must be admitted that the method of Bolles is a more robust solution in industry and will be a likely method to be used for some time to come, in the light

of the discussions in chapter 2.

7.10. Summary.

The problem of occluded object recognition was proposed to be solved by using the method of intrinsic images generated from LMF images. The LMF images were to be combined in a novel algorithm that used the information from *textures, edge-depths and incoherent edge-angles* to detect any events of occluding and being occluded. The texture of the background *was* an important part of the search process.

The experimental results showed that although the concept is acceptable, the computation procedures required good formulations for

a) model descriptors and
b) good search methods.

The problem of good model descriptors was a bigger problem as simpler, or even more complex descriptors for textures, edges, etc. are bound to be constrained by some irregularities of the environment (e.g illumination intensity, illumination direction, change in texture due to angle of lighting, etc.).

Improved search methods in the computation could be helpful but it does not solve the basic problem of being given the right things (true edges, true object boundaries, etc) to search.

In some ways, the problem posed here for the solution by the multiresolutional foveal image method has been quite unconventional and consequently more difficult, as we have tried to place relatively little constraint on preengineeing, as was assumed in the two approaches in [Neveu86] and [Bolles83]. Whereas the boundary features (edge segments) of the binary objects were the inputs to the search programs of Neveu and Bolles, in this method we have tried to use a smaller set of features called the *minimum set of intrinsic images* for the problem at hand.

The multiresolutional fovea intrinsic image method could be improved by :-

a) a better *selection* of intrinsic images
b) a more robust model-reference (e.g texture measures, etc)
c) use of better model-matching (structural matching, Bayesian inference etc.)

It is felt that the intrinsic images do point towards a possible solution to this class of problems, for which the more conventional boundary-based methods would not likely be suitable. Any real success in this intrinsic image method would lie on the availability of computer and imaging hardware capable of generating multiresolutional fovea intrinsic images at a higher throughput rate. Further, we can only console ourselves that short of removing the *ill-posedness* of the problem, real-life occluded-object recognition will continue to be a difficult problem to solve.

Chapter 8

The Multiresolutional Foveal Image in a Hierarchical Progressive Image Transmission Scheme.

In robot vision, occasions arise where an image has to be processed at different sites using fast and efficient means (e.g defect inspection).

Gross coarse-information first image encoding or *Progressive transmission* (PT) schemes have been proposed with the aim of using minimum resources (time, storage, transmission channel) in conveying a image signal between two points in an *interactive* manner [Knowlton80]. It makes use of an encoding scheme that builds a hierarchy from the image and then sends these hierarchies in a prescribed manner to achieve this aim. An example that could be envisaged is that of two graphic workstations working on the same image (CAD image, CT images or just plain robot vision images) and needing the image to be sent to each other. The natural hierarchy in the multiresolutional foveal images was investigated for such a scheme. It was found that certain resulting properties give it additional advantages over the methods of Tanimoto [Tani79] and Burt [Burt83]

This chapter starts with a brief survey of relevant coding techniques for *images* and proceeds to develop the hierarchical transmission systems and algorithms for multiresolutional foveal images.

8.1 Brief Survey of Image[1] Coding Methods.

Possibilities of image coding arise from two main sources,

a) source coding, based on the statistical entropy of the image content and uses methods of *redundancy reduction*

b) Multiple component coding of *transformed* images. (transformation here means any operation on the image to obtain desired characteristics e.g region-segmented image, not necessarily *domain transformation* methods).

In robot vision, some transformations may be regarded as an encoding of the image (i.e if the object of interest is already segmented, then Blums medial axis transformation represents an efficient code of the object). The advantage of coding in this case is not so much of information compaction as being a suitable data structure for the purpose of data manipulation and matching in object recognition.

There are two fundamental steps in coding. The first is to decorrelate the input information. Decorrelated information provide many possibilities of coding. The second step is to devise coding methods to encode this decorrelated information. Therefore it may be possible to concisely say that coding methods generally involve novel techniques of achieving the above two steps in *information-preserving* (*lossless*) or *lossy* ways subjected to the constraints of the HVS.

Sampling and quantisation together form the most basic form of coding. Quantisation errors, aliasing

[1] It needs to be reminded that the images best suited for this hierarchical scheme are still- frame images as opposed to video signals with interframe information. However it seems these techniques could be extended to slow-scan systems.

errors and the Nyquist criterion set the upper bound on this method. If there is redundancy in the image, methods of *predictive coding* [Musmann79] and *transform coding* [Tescher79] can be used. Predictive coding uses neighbourhood pixels to achieve prediction either in causal or non-causal ways.

General transform methods involve the transformation of the image from one domain to another such that the transformed coefficients are less correlated with *energies* concentrated in a few coefficients, examples being Fourier, Walsh, Haar, Hadarmard, Karhunen-Loeve, Discrete Cosines etc. Amongst these transforms the Karhunen-Loeve can achieve the best energy compaction using only the first m of the n transform coefficients, although with long computation times.

Another category of image compaction makes use of the subjective perception in the HVS (e.g synthetic highs coding [Graham67]). Knowing the statistical distribution of the image, it is possible to use *optimal codeword assignment* methods which are variations of the Huffmans method [Huffman67].

[Kunt85]

Second generation methods involving *multiple-component coding* transforms the image into multiple components or signals, (edges, regions etc.) so that these signals can be tailored to their individual characteristics for efficient coding. Other methods based on the hierarchy of the subimages to be transmitted known as hierarchical coding schemes are becoming more important as interactive transmission becomes more common through ISDN networks (Integrated Services Digital Networks). The method of Multiresolutional fovea images could be classified in this category.

8.2 Multiresolutional Images and Progressive Transmission Codes.

Progressive Transmission possibilities are strictly speaking not solely restricted to multiresolutional images only. Any other method that can reorder its *transform coefficients* can similarly be progressively transmitted, though as we shall see later, they may need to be adapted to fit such schemes. In PT, series or images sent is such that successive images are progressively and non-destructively reconstructed using the property of **Global to local-order of perception** in the HVS [Spoehr82]. The method of Knowlton [Knowlton80] uses a regular decomposition of the original image into blocks so that the average of the block is represented by an optimal assignment on a look-up table. Hill [Hill83] extended the Knowlton approach for progressive colour images. This method allows a convenient way of interception of the transmission as the image is constructed in the spatial domain which coincides with the domain of visual perception.

Tanimoto [Tani79] uses the *Gaussian* pyramid images as the series of images for transmission. This uses decimated and reduced-resolution layers of the pyramid images which

are essentially low-pass versions of the original image in a manner having similarities with *interpolative reconstruction* and *maximal block* encoding.

In the *Laplacian* pyramid approach of Burt, [Burt] a preprocessing of the Gaussian pyramid images is done by a simple pixel-to-pixel subtraction of successive layers of the pyramid. This difference is then transmitted instead, as it allows for fewer digits for representation due to its magnitude reduction. Vector quantisation is further employed to yield higher data compression.

Progressive transmission using the *reordered coefficients* of transform coding methods like DCT, Hadarmard and Fourier [Takikawa84] have been proposed. However, these methods generally pose a difficulty of obtaining the right sequence of the hierarchy of coefficients since the transform domain does not by itself yield such information, [Ngan84] though in [Chen and Smith77], a priority of sequence transmission is obtained from the property of activity of the blocks of the original image.

8.3 Multiresolutional Foveal Images for Progressive Transmission.

The adaptation of the multiresolutional foveal images for use in progressive transmission schemes is based on the following procedure.

a. generation of progressively low-pass perifoveal images together with the all-pass foveal image which we call the GMF image I_G, being the set of $\{I_{g1}, I_{g2}..I_{gn}\}$ for $1<=n<=8$.

b. generation of a set of bandpass images $\{I_{b1}, I_{b2}..I_{bn}\}$ for $2<=n<=7$ by subtracting I_{gn} from the equivalent area a_n of the input scene image I_s,

$$a_n * I_{bn} = a_n * (I_s - I_{gn}) \quad ---(8.3)$$

where * indicates convolution operation.

c. the resulting bandpassed images I_{bn} now have properties of

i. being decorrelated

ii. having a smaller dynamic range of amplitudes that lend themselves well to vector quantisation methods.

iii. being representable in reduced samples of $1/r_n$ which results from the sample-and-hold process after consolidation, as explained in chapter 5, where r_n is the resolution reduction at nth perifoveal image.

d. Possible use of a regime of progressively refined image-reconstruction steps (inter-pixel interpolation.) This arises from the fact that there is always a true foveal image being available for interpolative reconstruction.

8.3.1 System Description for Progressive Transmission.

Refering to Fig. 8.3.1a, we can represent the kth instance of the reconstructed image as I_{Gk}, given by

$$I_{G_k} = \sum_{n=1}^{n=8} \frac{a_n}{r_n} * I_{gn}$$

$$= \{a_1/r_1 I_{g1} + a_2/r_2 I_{g2} + ..a_7/r_7 I_{g7}\}$$

where a_n, r_n are the subimage areas and resolution reduction factor in the image I_{gn}.

It is clear that the number of samples required to reconstruct the image I_{gn} is a_n/r_n.

In discussing Progressive Transmission, we have to introduce the idea of a set of *progressively transmitted* images $\{I_{G_{t0}}, I_{G_{t1}} ... I_{G_{tk}}\}$ where $I_{G_{tk}}$ is the composite image received after the kth discrete time of completion of transmission.

It is expected that on completion of any kth reconstructed image I_{Gtk} the recieving station can interactively decide with the sending station is further refinement of the image is required. If further refinements are required, the sending station will send the next $I_{g}n+1$ image. If further refinements are not required, then the Progressive Transmission terminates.

However, instead of simply sending the refinement images I_{gn} as low-pass images, an computation is done at the transmitting station to encode the low-pass images as band-pass images by a simple subtraction of nieghbourhood low-pass images. Such a procedure not only ensures that there is more decorrelation in the image data to be sent, so that this creates better opportunities for coding, it also calls for reduced number of bits for coding of the band-pass images.

The detail manner in which the sequence of I_{Gtk} are transmitted and reconstructed is illustrated in the alogorithm of fig. 8.3.1b. and described in the following paragraphs

Fig. 8.3.1a The Multiresolutional Foveal Image (MFR)

Progressive Transmission of Image I_{G_K} for k=1

(ie. first instance of reconstructed image I_{G1})

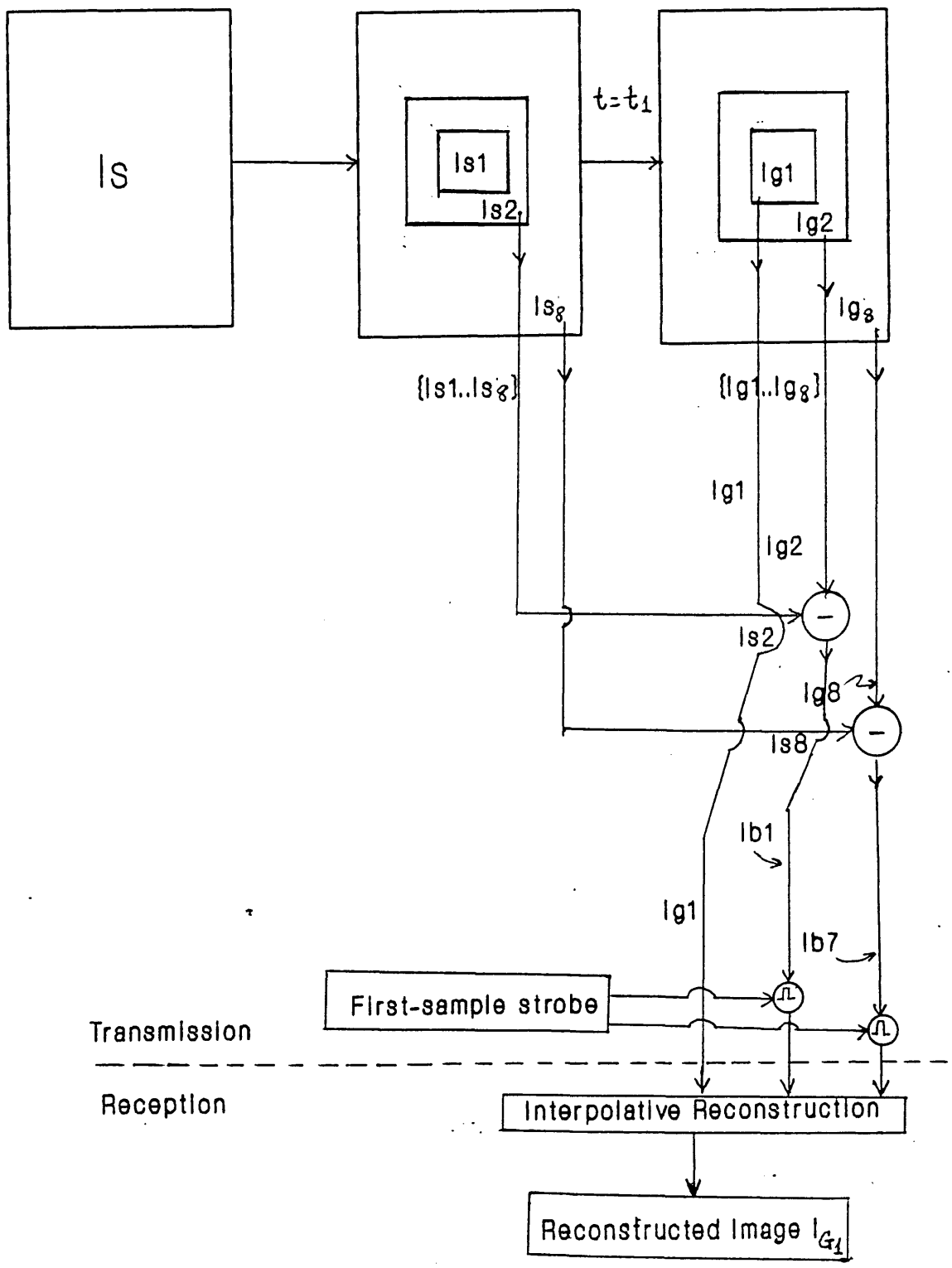
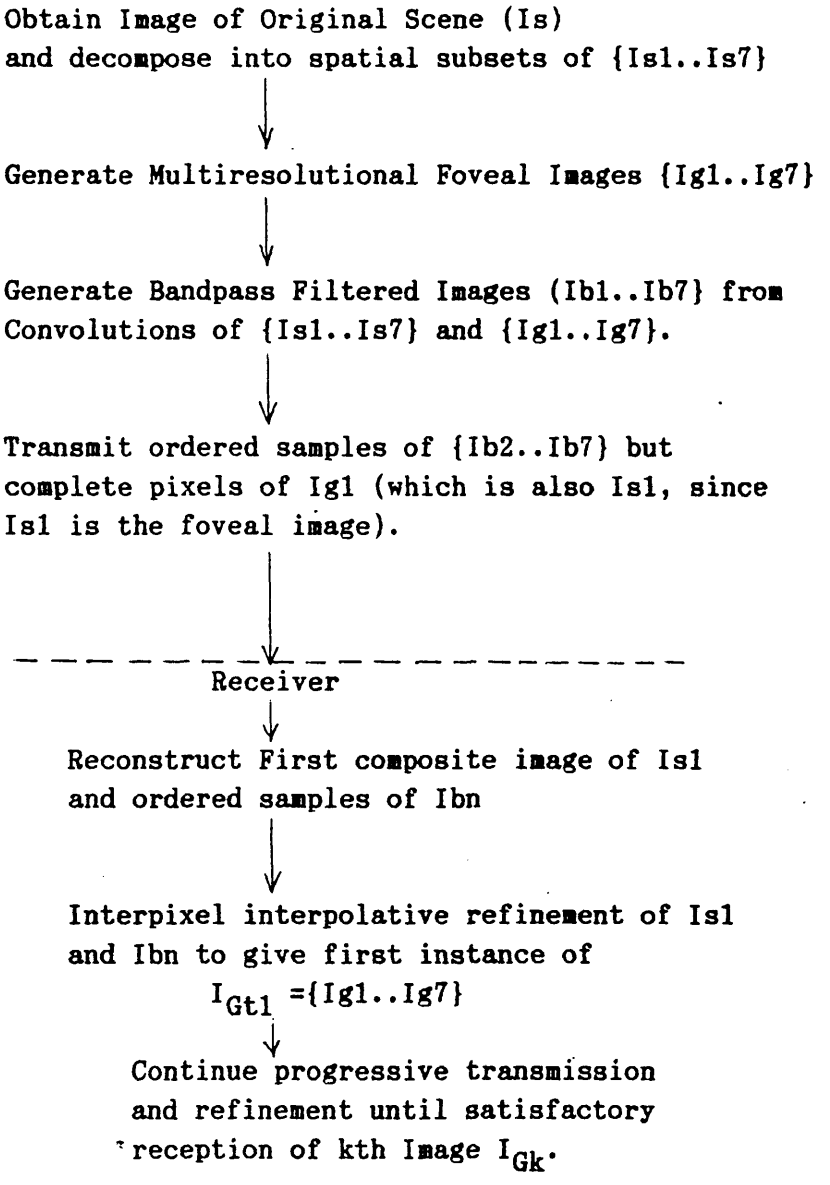


Fig. 8.3.1b

Algorithm for Progressive Transmission Using Multiresolutional Foveal Images



At time t1 the source image I_s is consolidated to obtain the set of $\{I_{g1},..I_{gn}\}$. However instead of sending the I_{gns} we send the difference signal I_{bns}. Therefore a set of bandpass images $\{I_{b2}..I_{bn}\}$ is obtained using eqn (8.3). The complete set of images comprising the foveal image I_{g1} **and** the bandpass images $\{I_{b2},..I_{b7}\}$ are then transmitted. At the receiving end the first instance of the progressively received image $I_{G_{t1}}$ is reconstructed with a simple interpolative method to reconstruct the subsampled perifoveal images using one immediate past sample only,

$$y_{n+1}=(x_n+x_{n-1})/n \quad \text{---(8.3.1) where} \quad y \text{ is the predicted pixel.}$$

The user, that is the receiving party can, at this stage, decide to proceed or abandon the transmission (if he sees sufficient evidence for continuation or cessation from the information in $I_{G_{t1}}$). If he decides to continue, a new set of images now of $\{I_{b1}..I_{bn}\}_{t2}$ is sent. This set is a *refinement* of the earlier set in that pixels samples not previously sent are only now sent. Various schemes for refinement can be used. The method chosen here is to send the extreme sample from the subsampled window. This was done hopefully that it will give better prediction performance in eqn (8.3.1).

This process continues until completion of the transmitted image set at $t=t_7$ where full recovery should be possible subject mainly to quantisation and interpolational errors. Fig. 8.3.1c shows the progressive transmission of the multiresolutional foveal image using the method described above. It is compared with the same image transmitted in simple line–sequential manner. Ignoring aspects of code efficiency from the difference-image transmission for the moment, we can see that if R_k is the cumulative total of pixels transmitted at time t_k then R follows the rule:-

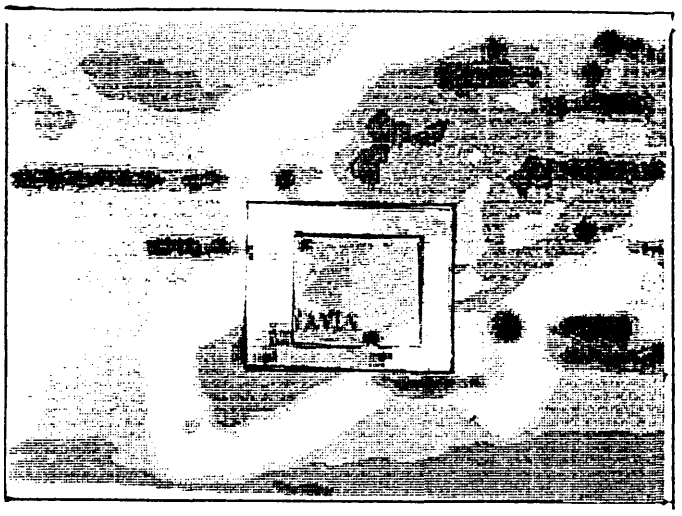
$$R_1=R_0 \qquad \textbf{(derived from table 4.4.1d)}$$

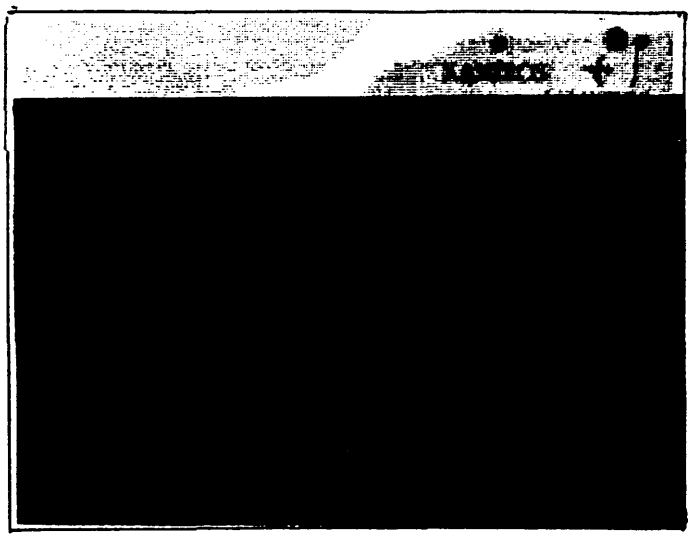
where R_0 is the total pixel number in the first MF image.

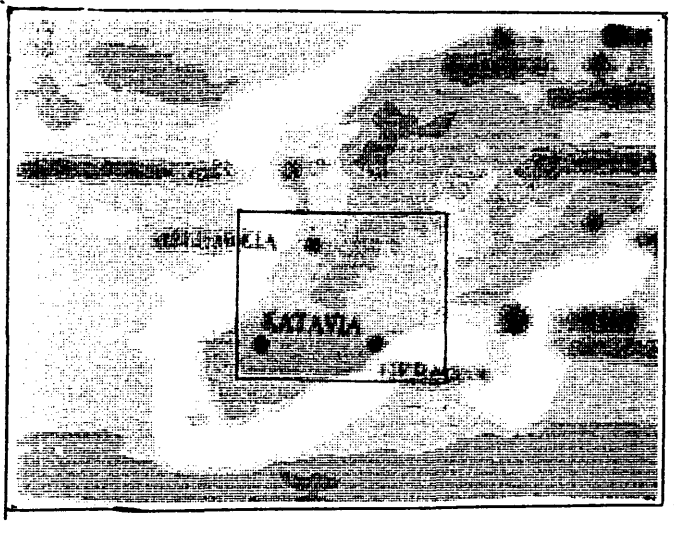
$$R_2=R_0+a_2/2+a_3/3+..a_7/7 \text{ or}$$

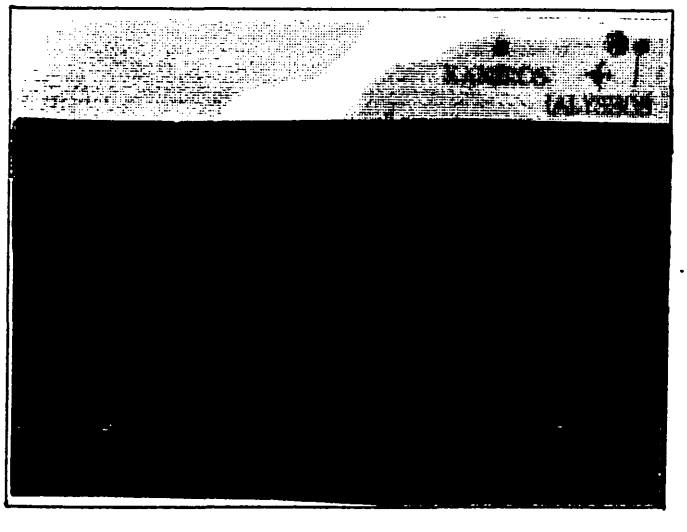
$$R_n=R_{n-1}..+..a_7/7 \text{ for the MF image.}$$

This means that the pixels that are progressively transmitted reduces with the instance $t=t_k$. This is shown in the graph of fig. 8.3.1d.









Multiresolutional Foveal Progressive Transmission

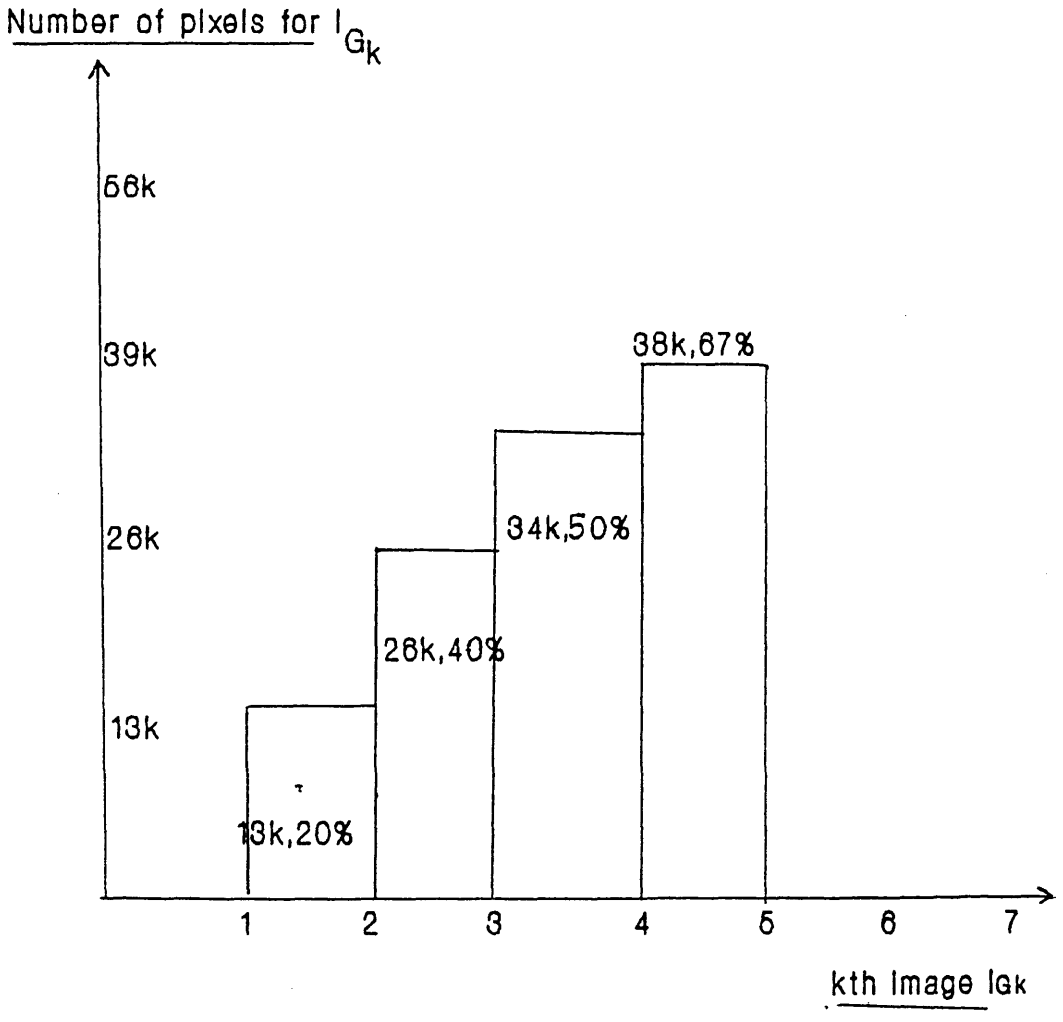
Top left = I_{G1}, 13k pixels
Bottom left = I_{G2}, 26 k pixels

Simple line-sequential transmission

Top right = 13k pixels
Bottom right = 26 k pixels

Fig. 8.3.1c Results of Transmission of 256x256-pixel Image 'Rhodes' using Progressive Transmission Method of the Multiresolutional Foveal Image at left and Simple Line-sequential Method at right.

Fig. 8.3.1d Number of Pixels transmitted during Progressive Transmission of 256x256 image



Compared to non-progressive transmission schemes, progressive transmission allows for a better mode of *interactive* transmission since the destination is better able to make use of available information per pixel transmitted, in order to decide on the further progress of the transmission. As an example, it is likely that a decision may be made at $t=t_2$ in the progressive transmission case, compared to perhaps $t=t_4$ in the simple scheme (non-progressive transmission scheme.)

8.3.2 Decorrelation in the bandpass images I_{bn}.

In the difference-signal of eqn (8.3) which is really a bandpass we note that the subtraction process creates pixels that are largely decorrelated. This arises from the fact that as I_{gn} is obtained through the consolidation process Π, over a window of the the input image I_s, where $I_{gn}=\Pi * \{I_s\}$, it follows that variance σ^2 of I_s and I_{gn} images (which should give a measure of the intra-frame pixel correlation) are related by

$$\sigma^2\{I_{gn}(x,y)\}=\sigma^2\{I_s(x,y)\}/\{n\} \qquad \text{--eqn (8.3.2)}$$

Eqn (8.3.2) shows that as the window-size increase, n increases and the resulting variability of the pixels in I_{gn} decreases. Put in another way, the pixels in I_{gn} are more correlated than the pixels in I_s. Consequently the pixels in the difference image

$$I_{bn} = (I_s) - (I_{gn}) \text{ would be largely decorrelated.}$$

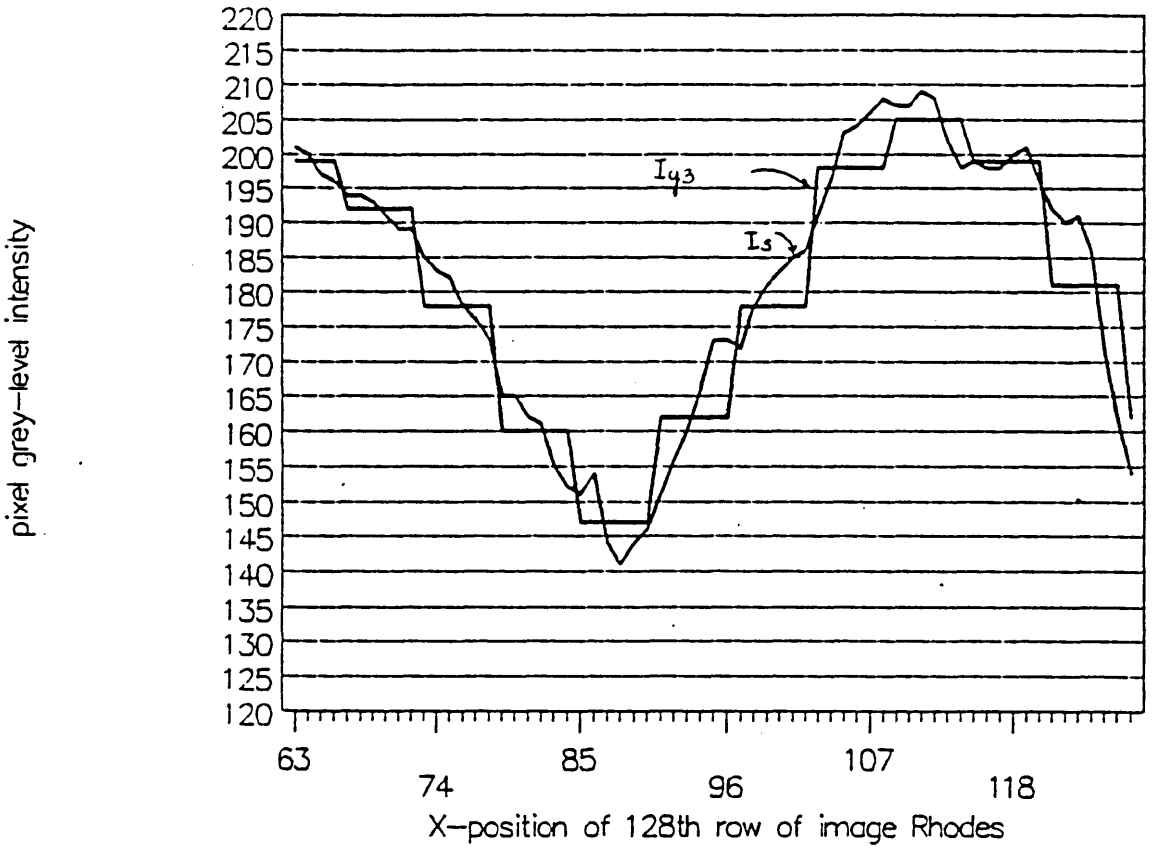
This is shown in fig. 8.3.2a for the middle row of the of I_S and the I_{g3} images. An alternative measure of the decorrelation of the pixels can be expressed in terms of the entropy of the image

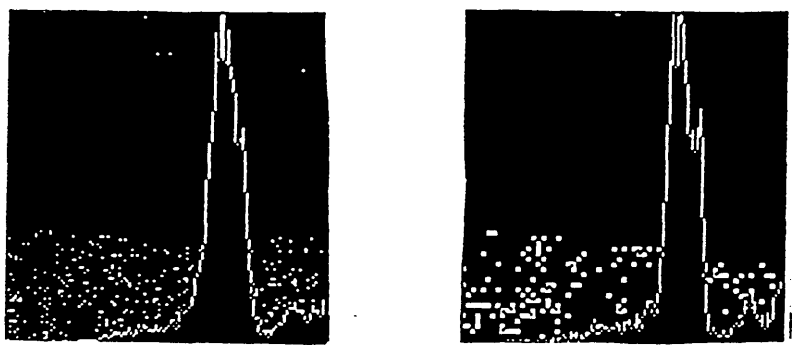
$$H=-\sum_{i=1}^{256} f(q)\log_2 f(q)$$

where f(q) is the probability of the pixel having intensity q.

The maximum entropy for the 256 level image would be 8 (if all pixels are equiprobable). If the variance increases, the entropy increases. The entropy measure is useful as it defines the minimum number of bits required to encode an image i.e the lower the entropy, the less bits per pixel for encoding.

The band-pass images of Fig. 8.3.2b would therefore be largely decorrelated and hence its entropy would be high. This consequently would allow for more opportunities of efficient coding for transmission of the image as shown in fig. 8.3.2c.





Histogram of I_s Histogram of I_{g4}

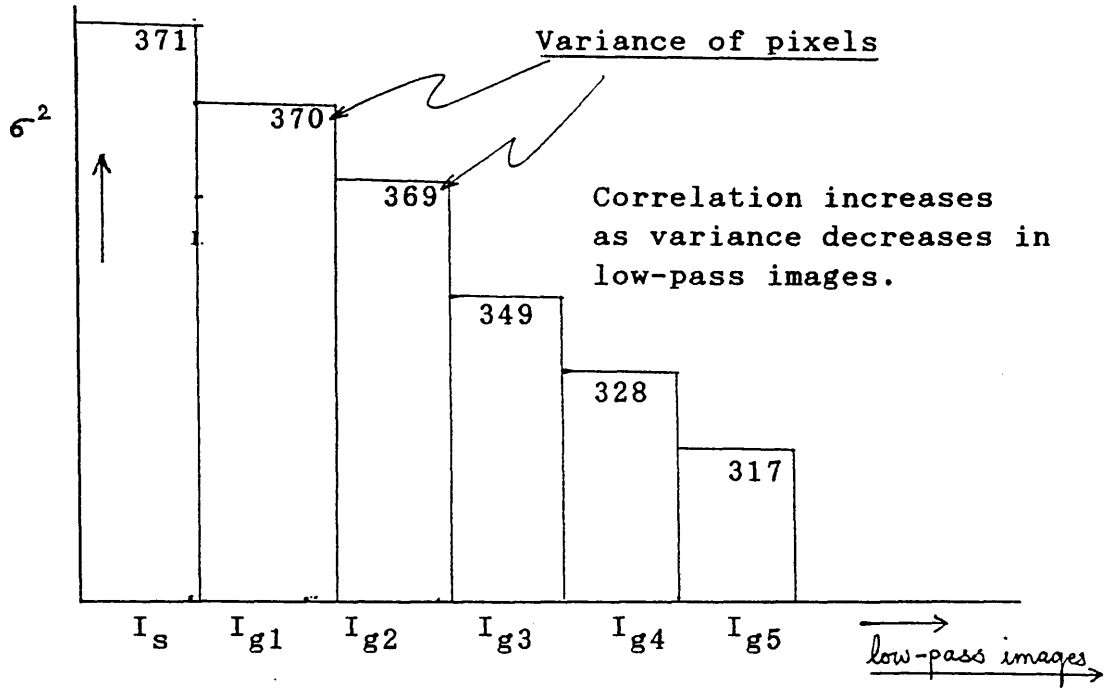
Fig. 8.3.2a

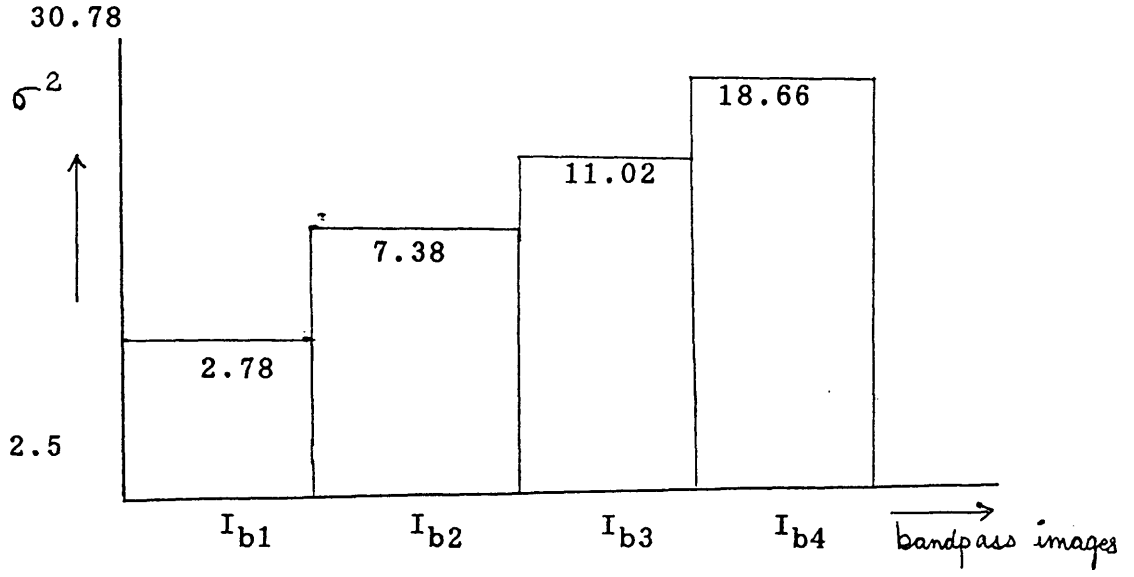
Grey-level Intensity of Is and Ig4
and their corresponding histogram.

Fig. 8.3.2b Printout of image grey level intensities of three typical images of 'Rhodes'

x-position	Original Image Pixel Intensity I_s	Low-pass Image Pixel Intensity I_{g2}	Bandpass Image I_{b1} Pixel Intensity ($I_s - I_{g2}$) (Intensity difference)
64	200	199	1
65	197	196	1
66	196	196	0
67	194	194	0
68	194	194	0
69	193	192	1
70	191	192	1
71	189	189	0
72	189	189	0
73	185	183	2
74	183	183	0
75	182	179	3
76	178	179	1
77	176	173	3
78	173	173	0
79	165	165	0
80	165	165	0
81	162	161	1
82	161	161	0
83	155	154	1
84	152	154	2
85	151	155	4
86	154	155	1
87	144	144	0
88	141	144	3
89	144	144	0
90	146	144	2
91	151	152	1
92	156	152	4
93	160	163	3
94	166	163	3
95	173	172	1
Variance	330.83	317.8	1.61
Mean	170.81	167.5	2

Fig. 8.3.2C Decorrelation of pixels of 'Rhodes' showing original image I_s, low-pass images $I_{g1..g6}$ and band-pass images $I_{b1..b4}$





Correlation of pixels in bandpass images showing increasing decorrelation with higher-order bandpass.

From the statistical properties of the bandpass images Ib_n, it would be possible to employ the simple Huffman Code as the basis for the variable-length code for each subimage Ib_n. As the entropy H increases with the order of n in Ib_n, meaning that the bandpass images further away from the fovea would have a higher entropy as shown typically in the table below for Ib_4 of the image 'Rhodes', one can naturally expect the coding transmission efficiency to increase with n.

There is also an added advantage that since the size of the perifoveal bandpass image increases with n (i.e the number of pixels within the bandpass image), this increased coding efficiency is used to a proportionally bigger extent when its coding efficiency is higher.

However this method also brings along with it, other encoding overheads, to provide for what is equivalently the supervisory codes to delineate the image boundaries, codeword lengths, codeword assignments, etc in order to fully use this Huffman scheme.

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0	4	7	0	-3	-1	5	5	-3	6	-5	5	1
1	1	-4	-4	1	-2	2	4	-1	7	-5	2	-3	3
2	-1	8	-3	-2	2	-4	-4	12	-1	9	-2	-2	3
3	6	-3	-11	6	-8	-9	7	-4	-3	-1	-2	-2	-4
4	1	4	3	-2	8	-3	7	1	2	3	-2	-1	6
5	10	-1	-3	0	3	-2	-1	-3	-2	3	-5	-2	1
6	11	-5	7	3	-4	-6	11	-3	-1	-4	2	5	0
7	-7	2	-4	-8	8	-6	1	-9	8	-3	-1	-1	-4
8	11	-1	5	-2	6	0	0	0	3	2	-2	4	1
9	2	-4	3	-3	4	1	3	0	-1	0	-1	6	-16
10	3	0	4	4	-2	3	3	-4	0	3	7	-5	3
11	-7	-6	0	-5	0	5	-14	1	3	-4	-3	0	-6
12	12	-10	3	6	-5	15	2	-4	7	-3	0	-1	-1

Bandpass Image pixels of top left corner (up to row 12 and column 12 of Ib_4 of Image 'Rhodes'. In this case, 16 distinct grey-level intensities exist, in addition to the sign-bit.

8.3.3 Quantisation advantage from decreased dynamic range of I_{bn}.

It is obvious that I_{bn} will have its pixel probability distribution function (pdf) concentrated near its 0 grey level, since it is the error of the input image and its predicted value. Instead of using the full 8 bits to represent the levels, we could use a smaller number of b bits to encode the j distinct levels by using the relation

$$b = \log_2 (j).$$

This should give a net increase in transmission efficiency though there are overheads involved. However, this advantage was not made use of in this experiment.

The number of distinct levels of the bandpass image I_{b1} being 4 in this example of 'Rhodes', would theoretically require only 2 bits plus a sign-bit to sufficiently encode the image pixel intensities.

An example of such bandpass images for I_{b1} and I_{b4} of 'Rhodes' is shown in the fig. 8.3.3a below.

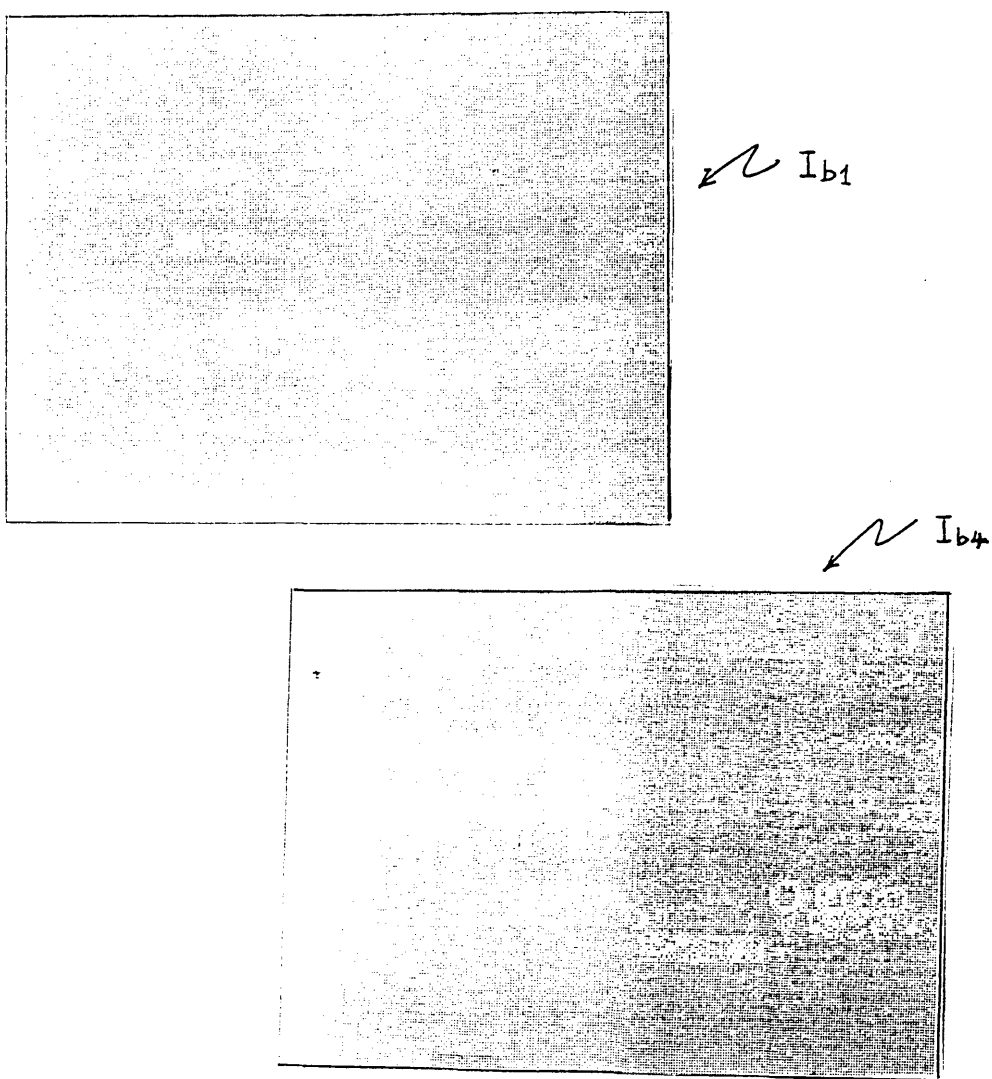


Fig. 8.3.3a Reduction of codeword lengths in the bandpass images.

8.4 Experimental Results.

Using a simple model for the progressive transmission of the Multiresolutional Foveal image (i.e without considering opportunities for further efficient coding using features of matched channels, optimal length codes (e.g Huffman codes), higher-order predictive coding, etc, the experimental results of just sending the bandpass images of the prescribed dimensions followed by a simple linear predictive reconstruction, as in eqn (8.3.1) show that this method of Progressive Transmission does have some distinct advantages.

This can be seen from an example of the image I_{G1} in fig.8.4a which is reconstructed using the bandpass images only. As the foveation centre was randomly chosen, falling at the lower middle part of 'Rhodes', the incidental appearence of the label 'GERADION' provided ample revelations about the image 'Rhodes'. This image required about 65600 bits of information for transmission, being approximately equal to 12% of the total information in the image 'Rhodes'. This can be contrasted with a similar 12% of the image sent using simple line-sequential transmission as seen in fig. 8.3.1c.

A detail calculation of the information bits used in this method of Progressive Transmission is shown below:

Image dimensions of 'Rhodes' = 256 x 256 pixels
Grey-level range of pixels = 256, requiring 8 bits for encoding.
Number of total bits in the image = 256 x256 x 8 = 524,288 bits

As we shall send I_{bn} as explained in para. 8.3, we consider the number of bits required to encode the bandpass images, obtained

in this case, by measurements using the real image 'Rhodes' as:-

Image	I_{b1}	I_{b2}	I_{b3}	I_{b4}	I_{b5}	I_{b6}	I_{b7}
Number of Distinct Levels in Bandpass Image=B	4	10	16	28	37	56	63

The number of bits required to encode and transmit the image in the Progressive Transmission method of the Multiresolutional Foveal Image, ignoring the details for signed bit and datum bit required for reconstruction of the image is shown below:-

Image Name	Area x B bits	Reduced Sample Transmission factor	I_{G1}	I_{G2}
I_{g1}	1024 x 8 = 8192			
I_{b1}	3072 x 4 = 12288	2	6144	6144
I_{b2}	5120 x 4 = 20480	3	6826	6826
I_{b3}	7160 x 5 = 35800	4	8950	8950
I_{b4}	9216 x 6 = 55296	5	11059	11059
I_{b5}	11264 x 6 = 67584	6	11264	11264
I_{b6}	15360 x 6 = 92160	7	13165	13165
Total (bits)	291800 bits		65600	123008

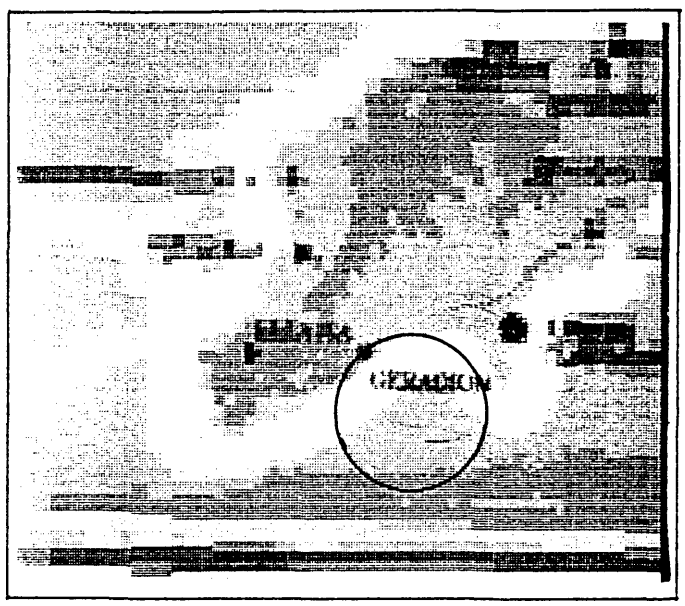


Fig. 8.4a Reconstruction of Ig1 from bandpass image pixels. Foveation centre randomly chosen as in circle.

From the above results we can see that the number of bits required to transmit and reconstruct the image I_{G1} and I_{G2} are

I_{G1} ----> 65,600 ----> 12% of pixels in 'Rhodes'
I_{G2} ----> 123,008 --> 23% of pixels in 'Rhodes'

By comparing the images in fig. 8.3.1c, we can see that there is a fairly large gain in efficiency using the Multiresolutional Foveal Image Transmission method, as opposed to the conventional line-sequential methods.

8.5 Comparison with related Progressive Transmission schemes.

There are three schemes for progressive transmission that are closely related to the method using multiresolutional foveal images. These are the method of Knowlton [Knowl80], Tanimoto [Tani79] and Burt [Burt83]. The method by Knowlton is one of the earliest novel methods to exploit the advantages of **progressive transmission**. As it does not explicitly use pyramidal structures, we shall instead make comparisons with the two hierarchical image structures of Tanimoto and Burt.

8.5.1 Comparisons with the Gaussian Pyramid.

In [Tani79], a progressive transmission scheme based on the Gaussian Pyramid of $\log_2(L^2-1)$ planes was described. Using an image of 2^L by 2^L matrix where L is any integer, an averaging pyramid is built up using 2^j by 2^j windows until the 2^0 by 2^0 plane is reached. During transmission, the first 2^0 by 2^0 plane p_0 is sent. Successive 2^1x2^1 pixels are sent and writes over the existing quadrant of the planes. This continues until the 2^Lx2^L plane P_k is constructed. This simple scheme is improved by sending selective parts of the quadrant to compute existing quadrants *non-destructively*. The third method mentioned in the paper is applicable to bilevel images. It makes use of the property of creating binary graphs that allow for the maximal block encoding of the images [Tani77].

8.5.2. Comparisons with the Laplacian Pyramid.

In [Burt83], the progressive transmission scheme makes use of the property of Laplacian Pyramids which is a difference signal of the input image and the Gaussian image planes (The Gaussian pyramid is so named because,instead of using the spatial averaging window as in the **Canonical Pyramid**, it uses an Gaussian-like function whose parameters are selected from the coefficients in the gaussian window on the spatial domain. However, in generating the difference image, an **EXPAND** operation is used to adjust for the decimated images in the gaussian planes. Decorrelation and quantisation of the difference images are used to achieve a high efficiency in this scheme.

In making comparisons between the multiresolutional foveal image PT scheme and that of Tanimoto and Burt, it is quite evident that the method of Tanimoto is a basic method, achieving efficiencies more from the hierarchical planes than from any decorrelation properties (which is why some maximal block encoding was found necessary). This must be expected since Gaussian Images do not provide the necessary decorrelation to employ encoding expedients.

The multiresolutional fovea scheme follows much of the ideas of Burt but with some differences namely:

a. There is no necessity to perform the **EXPAND** operation as in Burt since in the original MF image, a sample and hold was used to generate image planes that are undecimated,as

in chapter 5.

b. Unlike the Laplacian pyramid where the difference image is built for the **whole** of the $2^L x 2^L$ image through the **EXPAND** operation, the multiresolutional foveal images are constructed successively from the successive I_{bn} for the areas a_n. In the case of multiresolutional foveal images, progressive transmission is obtained from building up a hierarchy of $I_{G_{tk}}$ where each $I_{G_{tk+1}}$ is refined from previously received $I_{g_{tk}}$ in a non-destructive manner, unlike that in one of the methods of Tanimoto.

c. In the case of the MF images, there is always an instance of the foveal datum image being sent in the reconstruction of any $I_{G_{tk}}$. This datum being located in the central areas of the image is subjectively advantageous as it is generally true that many of the images have their more important parts near the centre. This is not true for both the Tanimoto model and the Burt model. It is not clear how important this is in progressive transmission but certainly it would not seem to be a impediment.

8.6 Summary.

Being able to make use of the multiresolutional image for progressive transmission of images is a desirable feature that is in a sense incidental to the hierarchical nature of this image representation. It should be useful in any robot vision development environment where template images perhaps need to be interactively transmitted between work stations. In using this scheme, we can point out the following limitations:

Computation overheads are required to generate the averaged and the bandpassed images though using schemes of pipeline processing as in examples achieved by Burt, this is not a significant disadvantage.

There is a need to send *supervisory* information to index the progression of the transmission (nth pixel of I_{bn} image of $I_{g_{tk}}$ of vector-quantised code).

In this simple form, it is not a truly information-preserving system and must be used in conjunction with the allowable error set by the HVS.

Overall, there is valid reason to suppose that the multiresolutional foveal images for progressive transmission will have a use where still-frame images need to be shared in the course of image processing work in a robot vision environment.

CHAPTER 9

Conclusion

The thesis has been concerned with a possible alternative image representation for use in robot vision. The emphasis has been on real- time robot-vision for tasks that were in the main related to object recognition. Certain regularities in robot vision work mainly in **structured lighting** were dispensed with in order to examine possibilities of robot vision in situations with the scenes are relatively more **visually-complex.**

In Chapter 1, important definitions relating to this thesis were introduced. The relationship between image representations and robot vision was discussed. The problem in robot vision was viewed in terms, not only of the hardware and algorithms but the **image representation,** defined as a suitable computational data structure to match the nature of the image. Some introductory aspects of the Multiresolutional Fovea Image (MF) were outlined. Theoretical and industrial approaches to solving the computer vision problems in general were reviewed. The theoretical parts covered the proposals of Barrow and Tennenbaum where intrinsic images were the critical link towards generalised robot vision. Marr's work in terms of the HVS as a model for low-level vision and the need for more global object-representation descriptors in terms of 2.5 dimensional structures were described. In seeing robot vision as being constrained by hardware, cellullar arrays and pyramidal hardware became a intense area of research in the last decade. The success in the WISARD architecture using conventional random access memories to work in a neural-like manner indicated that certain tasks in robot vision can be solved by novel techniques of pattern recognition. The motivations of the hierarchical manner in which humans perform cognition tasks led to the ideas of recognition cones and pyramidal computation structures. The wide coverage of computer vision techniques was seen in the syntactic approaches to visual recognition using adaptations of formal language and automata theory. Model-based approaches to industrial robot vision was described together with a brief discussion on the indispensability of model-based approaches in practical robot vision.

Chapter 3 was a review of image representations for computer vision. This review made a distinction between hierarchical and non- hierarchical, segmented and unsegmented images. For images that were already segmented, it is seen that there exist many techniques to represent the resulting bilevel images, or images with considerable number of similar regions, ranging from skeletonal methods to geometric moments. However for unsegmented images, compact data structures were not as easy to derive.

The idea of hierarchy in an image in terms of the differing weightages at different parts of the image was related to the multiresolutional images,for which the pyramid is one example. As the pyramid image representation is most related to the multiresolution foveal

some details of the canonical pyramid computational procedures were described.

Chapter 4 introduces the computational structure and characteristics of the multiresolutional foveal images. Since the MF image is very related to the human visual system, attention was drawn to the aspects of the human visual perception from the psychological, physiological and other allied fields.

Modelling the multiresolutional foveal images as images generated from a composite set of all-pass (fovea) and low-pass channels or filters, (perifovea) the possibility of spatial bandpass channels being a result of the collective actions of retinal receptive fields and backward inhibition, was postulated. Such spatial bandpass channels were shown to possess filtering characteristics very akin to the optimal filters in indeterministic scenes.

An application of a form of multiresolutional image known as the local multiresolutional (LMF) image, corresponding to local foveations of a smaller area in human perception was shown to be possible for the task of line or curve tracking in robot vision. The emphasis here,is that the use of the perifovea-support to make decisions on tracking a line that is described by semantic inputs (curvature, thickness of boundary etc.) makes the algorithm more robust. It was observed that false searches are not easily triggered due to the suppression of spurious information in the periphery. Similar line-tracking without perifovea-support was shown to be more uncertain.

One of the main applications of robot vision today is object recognition of which occluded object recognition is a comparatively more difficult task. Unlike the more conventional approaches to perform object recognition using binarised model features, an attempt was made to use the LMF to derive intrinsic images of the occluded object scene. This used three techniques to infer the various states of occlusion using textural and edge activity index. The performance and experimental results were compared to two methods that were proposed to solve similar task.

As a further study of the MF image in a possible situation where images need to be shared amongst users, the hierarchy of the MF image was observed to possess desirable properties for its being used in interactive, progressive-transmission schemes. The system description and theory of such a scheme, sometimes also referred to as a gross coarse-information transmission scheme was presented. Comparisons with the model proposed by Burt and Tanimoto showed that the MF scheme possess additional advantages.

The emphasis in the MF images for robot vision has been that it not only attempted to use images in a way akin to that of some aspects of human visual perception, but it has the possibility of being implemented in near real-time, without resort to special cellullar architectures

that will expectedly bring along with it, associated encumbrance of overly parallel processing. A suggestion on how such MF processing could be done using relatively conventional hardware of pipeline and digital signal processing chips with additional memory frames to store the various resolutions at near frame-rate was proposed.

The difficulty of studying robot vision in this manner using a more unconventional image was that it was always hard to make generalisations. Even knowing the current limitations of robot vision and how computer vision research is attempting to solve these problems, one always have the feeling, notwithstanding that these ways may not be the immediate answer to problems of industrial robot vision, they have provided much theoretical support for engineers in the industry. An example is that of 'image understanding', for which at times the engineers in the industry would find little relevance, not only to the domain of the factory but also such attempts of image understanding from a position of ill-posedness seem very limiting.

As was pointed out, the initial steps of robot vision in industry would be to structure the scene so that robot vision could proceed without having to resolve some dispensable unknowns, (e.g. shadows, occlusions, etc.). Although this procedure has managed to place robot vision as a aid in automation, it would definitely make robot vision more universal and versatile if it could be used in unstructured scenes.

It is likely that in future, active vision in its various forms would predominate over more passive methods. Such *active* methods may include active observers (for which stereopsis is a subclass) or multiple sensing. It is also likely that there will be more support in terms of real-time optics/electronics from the image-acquisition front-end since such information (e.g. state of focus of the scene, colour balance etc.) can be viewed as additional inputs to fuller automation.

One more dimension of information that is difficult to use, but which is very encompassing is temporal image information. The studies in optical flow have helped to understand the problem but it may be sometime before such techniques have an application in the industry.

It has been the main thought in this thesis that a solution to the universality of robot vision would not only be strong computational algorithms, but that the image, or data representation, or image representation should possess desirable characteristics of hierarchy. Such a hierarchy can then be exploited to make algorithms faster, more immune to spurious information, and better guided through coarse perifovea support. One of the obvious results of being able to process at higher speed is that some form of that so-important temporal information could then be available.

9.1 Suggestions for the Hardware Implementation of the Multiresolutional Foveal Images.

It must be expected that much of the speed increase in machine vision must necessarily come from faster hardware instead of just software techniques as typically demonstrated in the WISARD [IA84]. Further, as an engineer, one is always reminded of the need to consider implementable alternatives founded in engineering practice [IA90].

The suggestions for a hardware implementation of the multiresolutional foveal image generation and acquisition are made here for completeness of this study, though serious implementation of these schemes require extensive research.

The suggestions are based on two methods:-

a) Optical Hardware method.
b) Method using signal processing techniques of DSP or simpler real-time analogue signal processing (i.e video low-pass filtering circuits)

A) OPTICAL HARDWARE METHOD.

Generation of Multiresolutional foveal images from optical system lensing and focusing geometries.

This method is based on the principle that a defocused image naturally provides for blurrring which is equivalent to low-pass filtering of the optical image. Two changes in the conventional video camera front end would be required, namely:-

1) Use of a suitable geometrically curved CCD element in the camera instead of the conventional flat surface CCD array.

2) Use of a suitable set of counters to generate the windows of the multiresolutional foveal image so that a fixed regime of samples of the pixel elements are scanned out as MRF images.

In (1) the property of optical blurring due to being out of focus in the curved periphery of the specially structured CCD array means

that the focus point becomes the fovea and the region extending away from the fovea naturally becomes the perifoval regions. The curvature of the 2-D CCD array can be optimised to allow for the graded deblurring at the peripheral regions.

In (2), suitable time-base electronics must be implemented so that the sampling rate of the pixel-elements of the CCD array will follow the rule of the multiresolutional structure of the MRF image. Hence, the higest sampling rate will be done at the foveal area of the CCD array and a decreasing number of samples away from the centre of the CCD array. This is illustrated in fig. 9.1a.

B) SIGNAL PROCESSING METHODS.

Whereas, the method (A) requires some changes in the physical structure of the video camera and the active CCD array, the method in (B) retains the structure of the standard TV camera.

For the Multiresolutional Foveal Image to be generated in near real-time, maximum use could be made of the increasing power of digital signal processing (DSP) chips that have found applications as universal video image convolvers or signal filters [Brabner and Ritchings88].

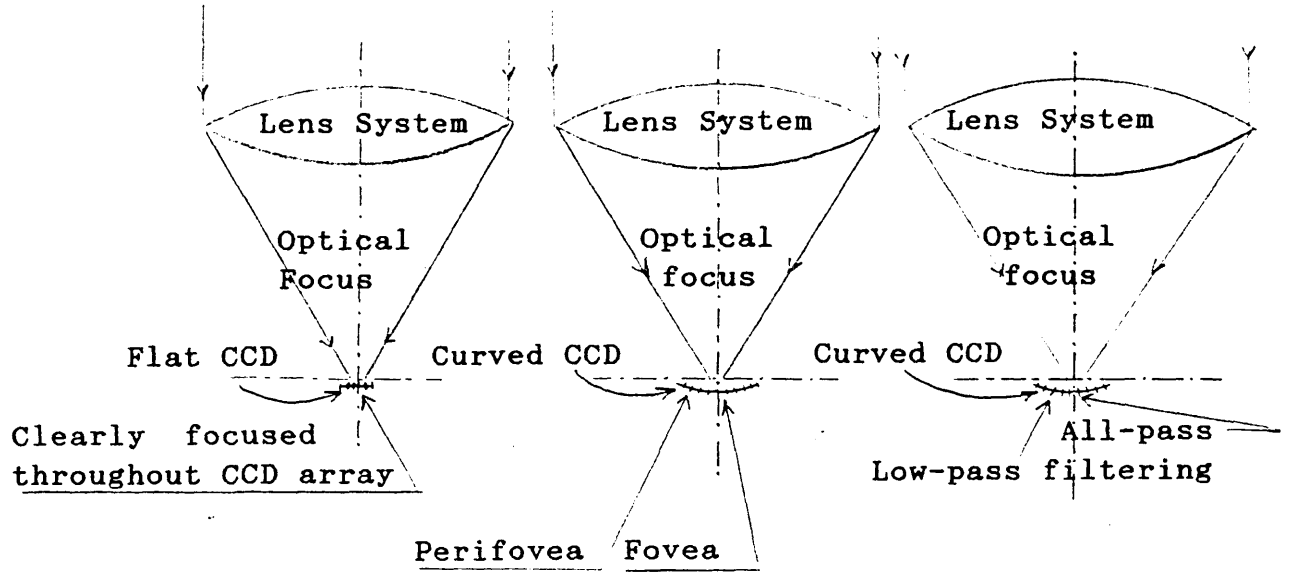
Alternatively, signal processing of video signals using conventional hardware-analogue filters (simple RC filters) which have been much ignored, seem to be potentially useful to generate the MRF image.

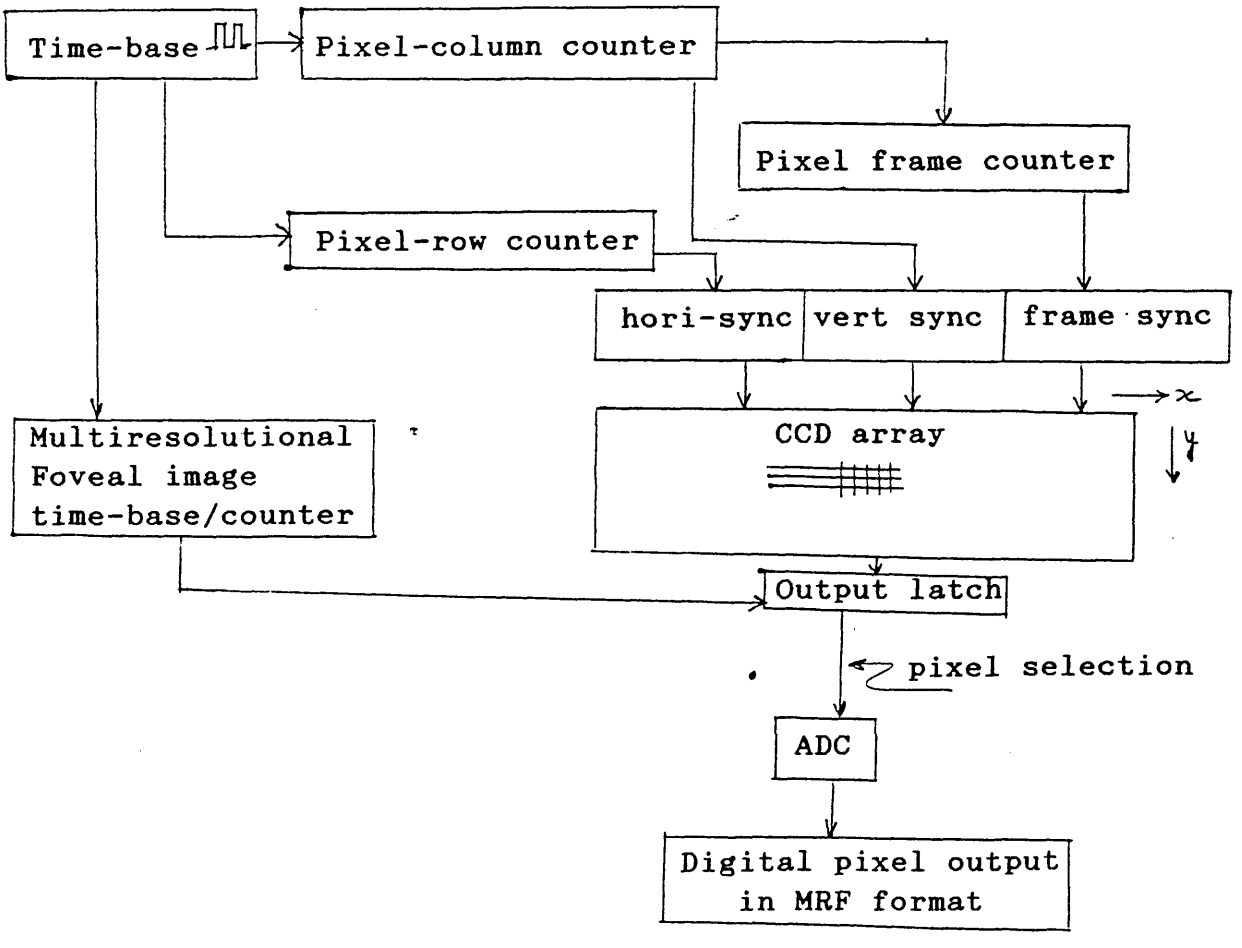
Consider the middle line of a standard video camera in Fig. 9.1b. The grey level intensity will be of equal resolution as it is scanned out from the scan electronics. However, if a suitable time-base/counter is included in this scheme so that those pixel samples that are away from the foveal area (corresponding to the mid-section columns of the line pixels) are passed to switched sections of low-pass filters of differing low-pass cut-off frequencies, and recombined, then the net signal produced will resemble that of a single line of the multiresolutional fovea image. Additional timebase/counter electronics and control could be included to assemble the MRF image frame.

The elegance in the method (B) is that there is a form of sequential processing much akin to the systolic-array [Kung82] or pipeline image processing methods, since each pixel is processed as it is being scanned out so that after a finitely small delay, perhaps not exceeding one frame-time, a new MRF image frame can then be made available.

The suggestions for the two possible methods of hardware implementation as in Fig. 9.1a and Fig. 9.1b are very preliminary. Further research on its implementation would be necessary.

Fig. 9.1a Defocussed CCD array as Multiresolutional Foveal Image.





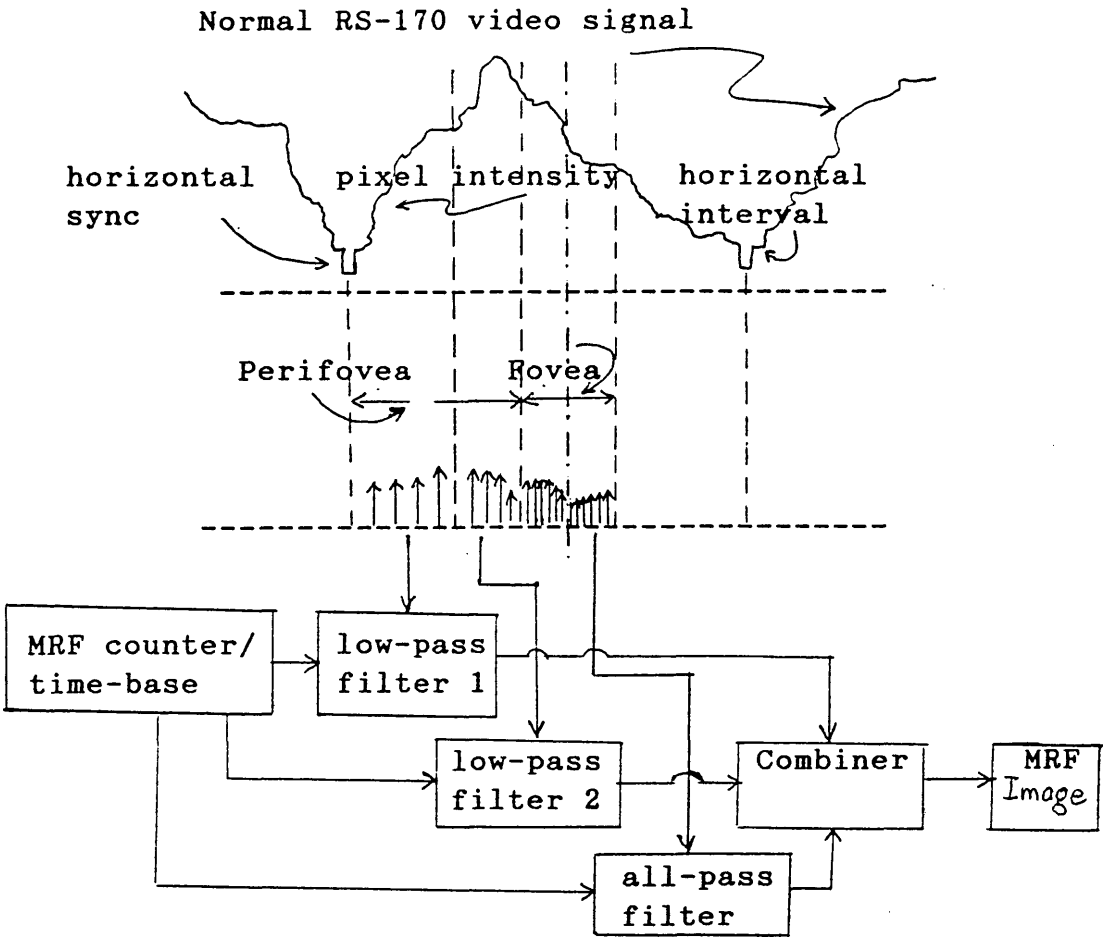


Fig. 9.1b

Generation of Multiresolutional Foveal images by MRF time-base controlled switching of filters.

9.2 Suggestions for Future Research.

The MF image proposed in this thesis has been quite primitive and exploratory. The dimensions of the various fovea and perifovea subimages have been experimentally chosen. The size of the fovea being a 32 by 32 in the case of the GMF seem to be too big so that foveal computations required approximately reading 1000 pixels. If the human fovea is equivalent to an arc of 1.5 degrees in a perimetric field of 180 degrees, then the equivalent fovea area in the GMF could perhaps be related to 0.008 of the image dimensions. The median generating kernel for the GMF image requires more computations since a sort of the neighbourhood window had to be made before the median is known. Further the median value may not even be a element of the neighbourhood since a new value may be generated. It is felt that perhaps results would have been the same, at least for real- world images, if the simple average or mean had been used.

In the model of the optimal filtering due to spatial bandpass channels, the characteristics of the perifoveal images could be treated as higher order filters than the simpler first order types. The study has not considered backward inhibitions from more than one source as in a multiple backward inhibition of Cornsweet. It would be also be more elucidating if some non-linearities in both the gain and frequency characteristics of the channels could be incorporated.

Further refinements incorporating non-uniform channel gains in the perifoveal channels would also be desirable.

In the application of the LMF for semantic line-tracking, the perifovea support, especially in detecting coherent edge angles to determine if candidate intensity-discontinuities should trigger a search, could perhaps incorporate better support calculations (e.g. both edge-angle coherence and edge-depth). This would have been more useful as semantic edges are also very much characterised by the

edge-depth as in the case of the RCA type plug where its edges have a different edge depth compared to the flatter coins. It would also be desirable to incorporate some weightages for the information from the two perifoveal regions than is done in the simple scheme here.

The set of *multiresolutional foveal intrinsic images* used for occluded object recognition might not have been a well-matched set. This is because the information using edge-strengths and edge-gradients are sometimes redundant, as they are both derived from a common resolution intensity- discontinuity operator. The problem that was not solved well was how to find the best initial region to search. In the computations in this thesis, a top-left region by region scan to the bottom right of the image was performed. This avoided the difficulty of keeping either a counter or a stack of the regions tracked, so that excessive visitations to the same area in case of a difficult search, could be avoided. This meant that the advantage of the global MFC was not used.

In the scheme for progressive transmission of MF images, the following areas could be improved.

a) The use of non-causal consolidation to reduce prediction errors from causal pixels only. It is known that in real-world scenes, the picture correlation varies with the Euclidean distance in the 2-D image and therefore the mean or median obtained from a better correlated set of pixels would be a truer value than from a more uncorrelated set. However, to resolve this, but still wishing to preserve the more gradual successive bandpass-image refinements, the simpler quadrant consolidation scheme may have to be modified.

b) The use of the median generating kernel could be improved by not sending a new median if the median value is not an element of the neighbourhood. This means that we could instead send a pixel sample belonging to the set of elements in the window so that the receiving end could *undestructively* use this sample, knowing that it is an

element of the set. This may require a different strategy to send the successive bandpass-image refinements but such a scheme can be expected to be more efficient since received samples need not be destroyed.

c) Interpolative reconstructions using a bigger neighbourhood would definitely give the progressively received images a better fidelity than the simple single-immediate neighbourhood pixels implemented in the thesis.

An important area of future research in the multiresolutional fovea images is perhaps the engineering implementation of such a MF image in real-time so that the multiresolutional foveal images could be generated from the real-world scene in near real-time. The two suggestions on the hardware implementation of the Multiresolutional foveal images need further research. It would seem that the approach using lens optics and curved CCD array would be a approach worthy of further research.

Finally, further research in *all* aspects of robot vision should continue in its varied ways, so that a wider base of knowledge is available to tackle what is currently considered to be one of the most difficult of computer tasks -- Computer Vision.

REFERENCE

Workshop in Spatial Reasoning and Multi-sensor Fusion, Sponsored by AAAI, Oct 1987.

Agin, G. J. and Duda, R. O., *SRI Research for Advanced Automation*, Proceedings of the 2nd USA Japan Computer Conference, Tokyo, pp. 113- 117, 1975.

Aleksander, I., *Adaptive Pattern Recognition and Boltzmann Machines: A Rapprochement*, Pattern Recognition Letters (6) pp. 113-120, 1987

Aleksander, I., *Neural Computing Architectures,* North Oxford Academic Press, 1989.

Aleksander, I., *Emergent Intelligent Properties of Progressively Structured Pattern Recognition Nets*, Pattern Recognition Letters, Vol 1, pp. 375-384, 1983

Aleksander, I., *The logic of Connectionist Systems* in 'Neural Computing Architectures', (ed) Aleksander, I., North Oxford Academic Press, 1989.

Aleksander, I., Stonham, T. J., *Guide to Pattern Recognition using Random Access Memories*, IEE Journal on Computers and Digital Techniques, Vol 2(1), pp. 29-40 1979

Aleksander, I., Thomas, W. V. and Bowden, P. A., *WISARD: a Radical Step Forward in Pattern Recognition*, Sensor Review, pp. 120-124, July 1984

Aleksander, I., *Artificial Vision for Robots*, (ed) Chapman and Hall, New York, 1983.

Aleksander, I., *Ideal Neurons for Neural Computers*, International Conference in Parallel Processing in Neural Systems and Computers, March 1989, Dusseldorf.

Aloimonos, J. and Bandyopadhay, A., *Active Vision*, First International Conference on Computer Vision, pp. 35-52, London 1987.

Baker, K. D. and Sullivan, G. D., *Multiple Bandpass Filtering in Image Processing* IEE Proc. Vol 127, Pt E, No. 5, 1980.

Ballard, D. and Brown, C. M., *Computer Vision*, Prentice-Hall 1982

Ballard, D. H., *Eye Movements and Visual Cognition* pp 188- 203 in Proceeding of the Workshop in 'Spatial Reasoning and Multi-Sensor Fusion', Sponsored by the AAAI, Oct 1987.

Barrow, H. G., Narasimhan, R., Rosenfeld, A., *Visual Pattern Analysis in Humans and Animals*, Science, no 4, pp. 177, August 1972.

Barrow, H. G. and Tennenbaum, J. M., *Recovering Intrinsic Scene Characteristics From Images*, in 'Computer Vision', (eds) A.R. Hanson and E.M Riseman, pp. 3-25, 1978.

Baynon, J. D. and Lamb, D. R., *Charge-coupled Devices and their Applications*, London, Mcgraw-Hill, 1980.

Binford, T. O., *Survey of model-based Image Analysis Systems*, International Journal of Robotic Research, 1, Spring pp 18-63, 1977.

Blum, A., *A Transform for Extracting New Descriptions of Shape",* Symposium on Models for Perception and Speech and Visual Form 1964, MIT Press.

Brebner. G. E., Ritchings, R.T., *Image Transform Coding: a case Study Involving Real-time Signal Processing*, IEE Proceedings, Vol 135, Pt. E, No. 1, pp. 41-48, Jan 1988.

Brown, C. M., *Inherent Bias and Noise in the Hough Transform*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 5, pp. 493-505, Sept 1983.

Burt, P. J. et al, *A Pipelined Pyramidal Machine*, in Pyramidal Systems for Computer Vision (ed) Cantoni, V. and Levialdi, S., Springer-Verlag, 1986.

Cantoni, V. and Levialdi, S., (eds) *Architectures in Pyramidal Systems for Computer Vision*, Springer-Verlag, 1986.

Capellani, V., Constantinides, A. G., *Digital Filters and their Applications*, Academic Press, London, 1978.

Campbell, F. W. and Robson, J. G., *Journal of Physiology*, London, Vol 187, pp. 517, 1968.

Cornsweet, T. N., *Visual Perception*, Academic Press, New York, 1970.

Davson, L. H., *The Physiology of the Eye* McGraw-Hill, New York, 1949.

Dowling, J. E., *The Retina: An Approachabble Part of the Brain* Cambridge, Harvard University Press, 1987, Chap 3.

Duff, M. J., *Image Processing System Architectures*, (eds) Josef Kitler and M. J. B. Duff , Research Studies Press, U.K 1985.

Duff, M. J. B., *Clip4: A Large-scale Integrated Circuit Parallel Array Processor*, Proc. Third Intnl. Joint Conference on Pattern Recognition, Vol 4, pp. 728-733, 1976

Durell, T. and Kwangyoen Wohn, *Pyramidal Based Depth from Focus*, IEEE 1988.

Elnas, S. E., Tzou, K. H., Cox, J. R. Cox, Hill, R. L. and Jost, R. G., *Progressive Coding and Transmission of Digital Images*, IEEE Trans. Med. Imaging. Vol M1-5, No. 2, pp 73-83, Jun 1986.

Enroth-Cugell, C. and Robson, J. G., *Journal of Physiology*, 187, pp. 517, 1966.

Faugeras, O. D., *Fundamentals of Computer Vision*, Cambridge University Press, pp 315-325 (1982)

Ferretti, M., *Overlapping in Compact Pyramids*, in Pyramidal Systems

for Computer Vision, (eds) Cantoni, V., Levialdi, S., Springer-Verlag, 1986.

Fischler, M. A., *On the Representation of Natural Scenes* in *Computer Vision Systems* (eds Riseman) pp. 47-52.

Flory, R. E., *Image Acquisition Technology*, Proceedings of the IEEE, Vol 73, pp. 613-637, April 1985.

Fong, A. M., *Multiresolutional Retina Images for Machine Vision*, INNS-IEEE International Joint Conference on Neural Networks, Washington, Jan 1990.

Fong, A. M., *Optimal Filtering using Multiresolutional Foveal Images*, International Conference on Automation, Robotics and Computer Vision (ICARCV 90), Singapore, Sept 1990.

Fu, K. S., *Syntactic Methods in Pattern Recognition*, Academic Press, New York, 1974.

Fu, K.S., Conzeles, R.C., Lee, C.S.G., *Robotics: Control, Sensing, Vision and Intelligence*, McGraw-Hill, 1987

Gabor, D. E. and Hill, P. *Television Bandwidth Compression by Contour Interpolation*, IEE Conf. Proceedings 168B, No. 39, pp. 303-315, 634, 1961.

Ganong, W. F., *Review of Medical Physiology*, Maruzen Asian Edition, 1981.

Georgeson, M., *Spatial Fourier Analysis and Human Vision* in 'Tutorial essays in Psychology' (ed). S. Sutherland, Earlbaum Associates, Vol 2, 1978.

Georgeson, M. A. and Sullivan, G. D., *Contrast Sensitivity: Deblurring in Human Vision by Spatial Frequency Channels* J. Physiology, 252, pp 627-656, 1975.

Graham, D. N., *Image Transmission by Two-Dimensional Contour Coding*, Proc IEEE, vol 55, no. 3, pp.336-346, Mar. 1967.

Hall, E. L., *Computer Image Processing and Recognition*, Academic Press, New York, 1979.

Hanson, A.R. and Riseman, E.M., *Processing Cones: A Computational Structure for Image Analysis*, pp 101-131,in Structured Computer Vision (eds) S. L. Tanimoto and A. Klinger (Academic Press) 1980.

Haralick, M. R., *Image Texture Survey*, in Fundamentals of Computer Vision, (ed) O. D. Faugeras, Cambridge University Press, 1983.

Haralick, R. M., Shamugam, R. and Dinstein, I., *Textural Images for Image Classification*, IEEE SMC-3, pp 610-621, Nov 1973.

Hi.S. Hill, Walker, S., Gao, F., *Interactive Image Query System using Progressive Transmission*, Computer Graphics, Vol 17, No. 3, pp. 323-330, July 1983.

Horn, B. K. P., *Robot Vision*, McGraw-Hill, 1986

Horowitz, S., *Peak Recognition in a Waveform*, in 'Syntactic Pattern Recognition: Applications'. (ed) K.S. Fu, pp. 1-31, Springer- Verlag. 1977.

Hsu, M., *Visual Pattern Recognition by Moment Invariants*, IRE Transactions on Information Theory , pp. 179-187, 1970.

Huang, T. S., (ed) *Picture Processing and Digital Filtering*, Springer, Berlin, 1979

Huang, T. S. and Tretiak, O. J. (eds), *Picture Bandwidth Compression*, New York, Gordon and Breack, 1972.

Hubel, D. and Wisel, T., *Receptive Fields, Binocular Interaction and Functional Architecture in the Cats Visual Cortexs* J. Physiology, 160, pp. 106-154.

Hueckle, M. H., *An Operator which Locates Edges in Digitised Pictures*, Journal ACM, Vol 18, pp. 113-125, 1971.

Huffman, D. A., *A Method for the Construction of Minimun Redundancy Codes*, Proc. IRE, vol 40, pp.1098-1101, Sept 1952.

IEEE Trans. Comm., Vol-25, Special Issue on Image Bandwidth Compression, Nov 1977.

Kelly, G., *Edge Detection in Pictures by Computers Using Planning* (M2,D,Mitchie 1971 pp397-409)

Klinger, A., *Data Structures and Pattern Recognition* Proceedings of the First International Joint Conference on Pattern Recognition, IEEE, New York, pp. 497-498, Oct 1973.

Klinger, A. and Dyer, C. R., *Experiments on Picture Representation Using Regular Decompostion*, Computer Graphics and Image Processing 5, 69-105 1976

Knowlton, K., *Progressive Transmission of Grey-Scale and BInary Pictures by Simple, Efficient, and Lossless Encoding Scheme*, Proc IEEE, Vol 68, No 7, pp 885-896, July 1980.

Kruse, B., *The PICAP Picture Processing Laboratory*, Proc. of the Third Intnl Conference on Pattern Recognition, Vol 4, pp. 875-881, 1976

Kruse, B., Gudmunson, B., Antonsson, D. in *multicomputers for Image Processing*, (eds) K. Preston and L. Uhr, pp. 31-45, Academic Press.

Kunt, M., Ikonopoulos, A., Kocher, M., *Second Generation Image Coding Techniques*, Proc. IEEE, Vol 73, pp. 549-574, April 1985.

Kunt, M. and Johnson, D., *Block-Coding of Graphics: A Tutorial Review*, Proc IEEE, Vol 68, No. 7, pp. 770-786, July 1980.

Levialdi, S., *Edge Extraction Techniques* in Fundamentals of Computer Vision, (ed) O. D. Faugeras, Cambridge University Press, 1983.

Levine, M. D., *Feature Extraction: A Survey* Proc. IEEE, Vol 57, pp. 1391-1407, Aug. 1969.

Li, R.Y. and Fu, K. S., *Error-Correcting Tree Automation for Syntactic Pattern Recognition*, IEEE Transactions on Computers, C- 27, no. 11 pp. 1040-1053. 1978.

Marr, D., *Representing Visual Information- A Computational Approach*, in 'Computer Vision Systems', (eds) A.R Hanson and E.M Riseman, pp. 61-80, 1978.

Marr, D., *Vision: A Computatational Investigation into the Human Representation and Processing of Visual Information*, Freeman, San Franscisco, CA, 1982.

Marr, D. and Nishihara, H. K., *Representation and Recognition of the spatial organisation of 3D shapes*, Proceedings of the Royal Society, B200, pp. 269-290, 1977.

Minsky, M. L., *Steps towards Aritificial Intelligence*, Proc. IEEE, Vol 49, pp.8-30, Jan 1961.

Mussman, H. G., *Predictive Image Coding*, in Image Transmission Techniques, (ed) W. K. Pratt, Academic Press, New York, 1979.

Mustafavi H. and Sakrison, H. J., *Structure and Properties of a Single Channel in the Human Visual System*, Vision Research, Vol 16, pp. 957-968, 1976.

Nahim, N., *The theory of the measurement of a Silhouette Description for Image Processing and Recognition*, Pattern Recognition, Vol 6, No. 2, pp.85-95, 1974.

Neveu, C. F., Dyer, C. R. and Chin, T. T., *Two Dimensional Object Recognition Using Multiresolutional Models*, Computer Graphics and

Image Processing, 34, pp52-65, 1986

Ngan, K. N., *Image Display Techniques Using the Cosine Transform*, IEEE Trans. Acoustics, Speech and Signal Processing, Vol 32,, No 1, pp. 173-177, Feb 1984.

Pavlidis, P., *Piecewise Approximation of Plane Curves*, Pattern Recognition, pp. 346-405

Perkins, W. A., *A Model-based Vision System for Industrial Parts*, IEEE Transactions on Computers, Vol C-27, pp. 126-143, Feb. 1978.

Porat, M. and Zeevi, Y. Y., *The Generalised Gabor Scheme of Image Representation in Biological and Machine Vision*, IEEE Transactions in Pattern Analysis and Machine Intelligence, Vol 10, No. 4, pp. 452- 467, July 1988

Pott, J. and Coiffet, P., *Comparison of 5 Methods for the Recognition of Industrial Parts*, in 'Artificial Vision for Robots' (ed) I. Aleksander, Chapman and Hall, New York, 1983.

Pratt, W. K. *Digital Image Processing*, New York, Wiley-Interscience, 1978.

Pugh, A., (ed) *Robot Vision*, IFS-Springer-Verlag, 1983.

Ratcliff, F., Hartline, H. K. and Miller, W. H., *Spatial Temportal Aspects of Retinal Inhibitory Interaction*, Journal of the Optical Society of America, Vol 53, 1, Jan 63, pp. 110-120, 1963.

Roland T. Chin and Dyer, C. R., *Model-based Recognition in Robot Vision*, Computing Surveys, 1986, Vol 1, No 1, March 1986

Rosenfeld, A., *Computer Representations of Digital Images and Objects* in *Fundamentals of Computer Vision*, edited by O.D. Faugeras 1982.

Rosenfeld, A., *Hierarchical Representation: Computer Representations*

of Digital Images and Objects, in Fundamentals of Computer Vision, (ed) O. D. Faugeras, Cambridge University Press, 1983.

Rosenfeld, A., *Intrinsic Images: Deriving three-Dimensional Information about a Scene from Single Images*, in Fundamentals of Computer Vision, (ed) O. D. Faugeras, Cambridge University Press, 1983.

Rosenfeld, A., *Quadtrees and Pyramids for Pattern Recognition and Image Processing* in Proc. 5th Internaltional Conference on Pattern Recognition, Miami Beach, Florida, pp. 802-811, 1980.

Rosenfeld, A., *Computer Vision*, Advances in Computers, Vol 27, pp. 265-308, 1988.

Rosenfeld, A., *Some Pyramid Techniques for Image Segmentation*, in Pyramidal Systems for Computer Vision (ed) Cantoni, V. and Levialdi, S., Springer-Verlag, 1986.

Rumelhart, D. E. and McClelland, J. L., (eds) *Parallel Distributed Processing*, I and II, Cambridge, MA, MIT Press, 1986.

Sachs, M. B., Nachmias, J. and Robson, J. G., *Spatial-Frequency Channels in Human Vision* pp 1176-1186, Journal of the Optical Society of America, Vol 61, No 9, Sept 1971.

Sakrison, D. J., *On the Role of the Observer and a Distortion Measure in Image Transmission* IEEE Transactions on Communications, Vol 25, No. 11, pp. 1251-1267, Nov 1977.

Shaw, A. C., *A Formal Picture Description Language scheme as a Basis for Picture Processing*, Systems Information and Control, vol 14, pp. 9-52.

Shirai, Y. and Tsuji, S., *Extraction of Line-drawings of 3- Dimensional Objects by Sequential Illumination from Several Directions*, Second International Joint Conference on Artificial Intelligence, London, Sept 1971.

Smith, W.H. Smith and Smith, H., *Adaptive Coding of Monochrome and Colour Images*, IEEE Transaction on Communications, Vol 25, No. 11, pp. 1285-1292, November 1977.

Spoehr, K. T. and Lehmkuhle, S. W., *Visual Information Processing*, W.H. Freeman and Co, 1982.

Sternberg, S. R. *Cytocomputer for Biomedical Image Processing*, Computer, Vol 16, pp. 22-34, 1983.

Sullivan, G., *Filtering in Image Understanding* in *Digital Signal Processing* (ed) N. B. Jones, IEE Publications 1982.

Takikawa, K., *Fast Progressive Reconstrcution of a Transformed Images*, IEEE Trans on Information Theory, Vol 30, No. 1, pp 111-117, Jan 1984.

Tanimoto, S. L., *Pictorial Feature Distortion in a Pyramid* Computer Graphics and Image Processing, Vol 5, pp. 333-352, 1976.

Tanimoto, S. and Klinger, A., (eds): *Structured Computer Vision: Machine Perception through Hierarchical Computation Structures* (Academic Press, New York, 1980.

Tanimoto, S. L., *A Pyramid model for Binary Picture Complexity*, Proceedings of the IEEE Conference in Pattern Recognition and Image Processing, Troy, New, York, pp. 25-28, 1977.

Tanimoto, S. L., *A Pyramidal Approach to Image Processing*, 10th International Symposium on Computer Architecture, Stockholm.

Tanimoto, S. L., *Image Transmission with Gross Information First*, Vol 9, Computer Graphics and Image Processing, pp.72-76, 1979.

Tanimoto, S and Pavlidis, T., *A Hierarchical Data Structure for Picture Processing*, Computer Graphics and Image Processing, Vol 4, pp. 104-119, 1975

Tanimoto S, Klinger A, *Structured Computer Vision*, Academic Press, New York, 1980.

Tescher, A. G.,*Transform Image Coding*, in Image Transmission Techniques, (ed) W. K. Pratt, Academic Press, New York, 1979.

Tsotsos, John K., *A Complexity Level Analysis of Vision*, IEEE Trans on Computers 1987.

Uhr, L., *Layered Recognition Cones that Preprocess, Classify and Describe*, IEEE Trans. Computers, 21, pp. 758, 1972.

Wallace, A. M., *Industrial Applications of Computer Vision since 1982*, Proceedings of the IEE, Vol 135, Pt E, No. 3, pp. 117- 135, May 1988.

Weiman, C. F. R. and Chaikin, G. M., *Logarithmic Spiral Grids for Image Processing*, Proc. 1979 Computer Society Conference on Pattern Recognition and Image Processing, Chicago, pp. 25-31, 1979.

Werblin, F. S., *The Control of Sensitivity in the Retina* Scientific American, Vol 1, pp. 21-29, 1973.

Yachida, M. and Tsuji, S., *A Versatile Machine Vision System for Complex Industrial Parts*, IEEE Transactions on Computers, pp. 882-894, vol 26, Sept 1977.

Yan, J. K. and Sakrision, D. J., *Encoding of an Image Based on a Two Component Source Model*, IEEE Trans. Comm,. Vol COM-25, pp. 1315-1322, Nov 1975.

Kung, H. T., *Why Systolic Arrays?*, IEEE Computers, Vol 15, No. 1, pp. 38, Jan. 1982.