

```
*****
*
* KNOWLEDGE-BASED VISUAL INTERPRETATION *
*
* USING DECLARATIVE SCHEMATA *
*
* by *
*
* ROGER A. BROWSE *
*
* Technical Report TN 82-12 *
* November 1982 *
*
*****
```

Department of Computer Science  
University of British Columbia  
Vancouver, B.C., V6T 1W5

This report was submitted as a thesis in partial fulfillment of the requirements of the degree of Doctor of Philosophy.

## Acknowledgement

I wish to thank Professor Alan Mackworth for his advice, support, encouragement, and inspiration. I am truly fortunate to have had my thesis supervised by such a helpful and understanding person.

I also wish to thank Professors Anne Treisman and Daniel Kahneman for guiding my involvement in psychology, and for helping me to understand what it means to be dedicated to the ideals of science.

I am grateful to Professors Richard Rosenberg, Ray Reiter, and Bob Woodham for their help and instruction.

For the many useful and enjoyable discussions I thank Jan Mulder, Randy Goebel, Hilary Schmidt, Jay Glicksman, Bill Havens, Bill Prinzmetal, Jim Little, and Marc Majka.

Finally, I wish to thank Deborah Brown for all her patience and love.

This work was supported by scholarships from Izaak Walton Killam Memorial Scholarships and The National Science and Engineering Research Council of Canada.

## Abstract

One of the main objectives of computer vision systems is to produce structural descriptions of the scenes depicted in images. Knowledge of the class of objects being imaged can facilitate this objective by providing models to guide interpretation, and by furnishing a basis for the structural descriptions. This document describes research into techniques for the representation and use of knowledge of object classes, carried out within the context of a computational vision system which interprets line drawings of human-like body forms.

A declarative schemata format has been devised which represents structures of image features which constitute depictions of body parts. The system encodes relations between these image constructions and an underlying three dimensional model of the human body. Using the component hierarchy as a structural basis, two layers of representation are developed. One references the fine resolution features, and the other references the coarse resolution. These layers are connected with links representative of the specialization/generalization hierarchy. The problem domain description is declarative, and makes no commitment to the nature of the subsequent interpretation processes. As a means of testing the adequacy of the representation, portions have been converted into a PROLOG formulation and used to "prove" body parts in a data base of assertions about image properties.

The interpretation phase relies on a cue/model approach, using an extensive cue table which is automatically generated from the problem domain description. The primary mechanisms for control of interpretation possibilities are fashioned after network consistency methods. The operation of these mechanisms is localized and separated between operations at the feature level and at the model level.

The body drawing interpretation system is consistent with aspects of human visual perception. The system is capable of intelligent selection of processing locations on the basis of the progress of interpretation. A dual resolution retina is moved about the image collecting fine level features in a small foveal area and coarse level features in a wider peripheral area. Separate interpretations are developed locally on the basis of the two different resolution levels, and the relation between these two interpretations is analyzed by the system to determine locations of potentially useful information.

## Table of Contents

Abstract .....	ii
List of Figures .....	vi
1 Introduction .....	1
2 Framework and Approach .....	6
2.1 Use of Model Knowledge in Vision .....	7
2.1.1 The Image Feature Access Approach .....	8
2.1.2 The Volume Access Approach .....	11
2.1.3 Discussion .....	12
2.2 Line Drawings in Computer Vision Research .....	23
2.2.1 Line as a Symbolic Level .....	23
2.2.2 Some History .....	25
2.2.3 Generalizing from the Blocks-World .....	27
2.2.3.1 Compressing Constraints to a Single Level .	27
2.2.3.2 Different Types of Features .....	30
2.2.3.3 Availability of Components .....	31
2.2.3.4 Structure Within Labels .....	31
2.2.4 Beyond the Blocks-World .....	32
2.3 Levels of Resolution .....	36
2.3.1 Resolution Levels in Natural Vision .....	37
2.3.2 Resolution Pyramids or Cones .....	40
2.3.3 Resolution Levels in Edge Detection .....	42
2.3.4 Knowledge Interaction with Multiple Resolution Levels .....	46
2.4 Location Selection in Vision .....	48
2.4.1 Saccadic Eye Movements .....	50
2.4.2 Non-Arbitrary Fixation Location and Duration ..	52



2.4.3	Piecing Together Fixations .....	53
2.4.4	Models for Saccadic Control .....	55
3	Research Overview .....	60
3.1	A Model for Perception .....	60
3.2	Declarative Schemata .....	65
4	A Computer Vision Implementation .....	70
4.1	Problem Domain and Image Generation .....	71
4.2	Knowledge Representation .....	77
4.2.1	Adequacy of Representation .....	89
4.3	Preparation for Interpretation .....	91
4.4	Feature-Based Operations .....	96
4.5	Model-Based Operations .....	103
4.5.1	Locally Legal Interpretations Issue .....	106
4.5.2	Applying Network Consistency .....	108
4.5.3	Incremental Consistency .....	112
4.5.4	Representing Relation Instances .....	115
4.6	Selecting Processing Locations .....	118
5	Working Examples .....	122
5.1	A Single Fixation .....	122
5.2	Example with Multiple Fixations .....	147
6	Related Issues .....	161
6.1	Grouping and Feature Integration .....	161
6.2	Picture Grammars .....	168
6.3	View Based Representations .....	174
7	Conclusions .....	177
	References .....	182
	Reference Notes .....	195

Appendix A Angles at Connections .....	196
Appendix B Body Form Knowledge .....	197
Appendix C PROLOG Body Definitions .....	238
Appendix D Interpretation Examples .....	250

## List of Figures

2.1.1.	The Multiple Access Model of the interactions of object knowledge in visual perception .....	13
2.1.2.	The Volume Access Model derived from the multiple access model .....	15
2.1.3.	The Image Feature Access Model derived from the multiple access model .....	16
2.2.1.	Two labelled vertices from the blocks-world .....	28
2.2.2.	(a) A mountain symbol (b) A sketch map .....	29
2.3.1.	Typical local-global stimuli: (a) incompatible (b) compatible .....	39
3.1.1.	Different levels in the image hierarchy cuing at different levels in the specialization hierarchy .....	62
4.1.1.	Some examples of body form drawings .....	71
4.1.2.	Complete collection of body part depictions .....	73
4.1.3.	Image features and their attributes .....	74
4.1.4.	Line drawing at (a) 1024x1024 initial line drawing. (b) 128x128 averaged image .....	75
4.1.5.	Line drawing at (a) 32x32 averaged image (b) the axis of each detected blob .....	76
4.2.1.	Component description of a view of a hand .....	78
4.2.2.	Image to scene mapping description for a view of the hand .....	79
4.2.3.	Body form in starting position .....	80
4.2.4.	Body part orientation relative to its rest position described as a triple $(\theta_x, \theta_y, \theta_z)$ .....	81
4.2.5.	A single depiction of an upper-leg used to represent three different orientations .....	82
4.2.6.	Left-arm schema description .....	84
4.2.7.	The component hierarchy for the fine layer of the body form representation .....	86

2.4.8.	The component hierarchy for the coarse layer of the body form knowledge .....	87
4.2.9.	The specialization/generalization hierarchy for the body form representation .....	88
4.3.1.	A simple set labelling structure .....	92
4.3.2.	A partial example of the set labelling data structure for the curvature of lines .....	94
4.4.1.	A typical fixation of an image. The area in 128x128 resolution indicates the periphery, and the 1024x1024 area is the fovea. The rest of the image is shown in 32x32 resolution .....	96
4.4.2.	Initial situation, showing three lines connected by image hierarchy to a blob feature. The model possibilities are shown beneath the line features .....	99
4.4.3.	After the application of grouping consistency ..	100
4.4.4.	The final situation after the inter-level consistency has been applied .....	101
4.5.1.	Locally legal, but globally illegal structures in body form problem domain .....	107
4.5.2.	A network constructed from a schema description ..	109
4.5.3.	A network constructed from a schema description with an entry made .....	109
4.5.4.	A network constructed from a schema description after several entries .....	110
4.5.5.	A network constructed from a schema description after several entries .....	111
4.5.6.	Incremental Consistency Algorithm .....	114
4.5.7.	Example of specifications for the evaluation of attributes for a relation .....	116
5.1.1	The body form line drawing to be used as the example .....	123
5.1.2.	Area of available fine layer features in the single fixation at point (350,325) .....	124

5.1.3.	Area of available coarse layer features in the single fixation at point (350,325) .....	124
5.2.1.	The first fixation (at location 449 192). The small squares indicate periphery, and the large squares show unprocessed areas .....	148
5.2.2.	The second fixation at 162 448. The previously processed area is also shown .....	151
5.2.3.	The third fixation at 160 228. The previously processed areas are also shown .....	153
5.2.4.	The fourth fixation at 270 832 .....	155
5.2.5.	The fifth fixation at 96 672 .....	157
5.2.6.	The sixth fixation at 448 832 .....	159
6.1.1.	Two displays of the type used in Feature Integration Theory experiments. (a) conjunction target R, (b) feature target R .....	162
6.1.2.	Analysis of display configurations shown in terms of model possibilities .....	163
6.1.3.	Group processing digit detection display .....	164
6.1.4.	Features available at two resolutions .....	165
6.1.5.	Low resolution objects detected and model possibilities assigned to high resolution features which are roughly located .....	166
6.1.6.	Features assigned to objects detected at a finer level of resolution, for one of the low level objects .....	166
D.1	The body form in rest position .....	248
D.2	A complete parse tree for a body form .....	249
D.3	A half body at large scale .....	251

KNOWLEDGE-BASED VISUAL INTERPRETATION  
USING DECLARATIVE SCHEMATA

by

ROGER A. BROWSE

B.Sc. McGill University, 1972

M.Sc. The University of British Columbia, 1977

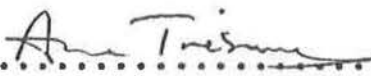
A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF  
THE REQUIREMENTS FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES  
DEPARTMENT OF COMPUTER SCIENCE

We accept this thesis as conforming  
to the required standard

  
.....

  
.....

  
.....

THE UNIVERSITY OF BRITISH COLUMBIA

November 1982

## 1. Introduction

Throughout human history there has been continual effort to develop tools and machines which can improve the efficiency and effectiveness of work. With the appearance of the digital computer, emphasis has shifted from a concern for the enhancement of physical capabilities to a concern for the development of machines which can accomplish tasks which otherwise require human mental activity. Artificial Intelligence is one discipline within the science and technology which has emerged to meet this challenge.

One of the basic goals of Artificial Intelligence is to develop a computational understanding of the powerful processes performed by the human mind through addressing tasks which are known to involve these processes such as natural language understanding, computer vision, problem solving and game playing.

Should a computational basis for these processes be understood to the extent that computers could be programmed to accomplish similar functions, profound modifications would be necessary in our concepts of intelligence and of existence. Over the first twenty years of Artificial Intelligence a number of areas of study have arisen as technologically useful byproducts of this research, such as knowledge representation, image analysis, expert systems, and robotics. Still it is unclear whether any progress has been made towards unravelling



the mysteries of the computational basis of human mental operations.

The research described in this document is concerned with the identification and pursuit of two principles which appear as important directions towards the accomplishment of this basic goal of Artificial Intelligence. These principles are followed in the context of a computational vision system which interprets line drawings of human-like body forms.

The first principle is a commitment to declarative structures, and to the separation of knowledge about objects and situations from the processes which employ the knowledge. If the computer is to remain an effective tool in research aimed at exposing computational mechanisms required for human information processing tasks, then it is essential that perspicuous definitions of the underlying knowledge structures be made available. Declarative structures are ideal for the explicit definition of both the task being undertaken, and the knowledge being employed, largely because of the close ties between declarative structures and the well known formal mechanisms of logic and grammatical representation.

A clear separation of knowledge and process provides the potential for verification and transferral of methods to other problem domains.

This idea has an intuitive appeal. In everyday human activity it appears that different processes access the same

knowledge structures: the same knowledge of objects seems to be employed in visual understanding, in forming mental images, in drawing, and in haptic manipulation. Furthermore, given the enormous array of objects understood in visual tasks, it is only reasonable that the complex operations of interpretation are not bound up separately with each object type, but rather reside as a unitary system which may operate with selected object knowledge.

This commitment to the separation of knowledge and process in vision has led to the development of a declarative schemata format for encoding knowledge of the problem domain of line drawings of human-like body forms. This declarative structure makes significant extensions to the earlier work in the use of grammatical representations for visual knowledge, and as well provides links to other popular approaches to line drawing interpretation.

The second principle being pursued in this research centers on the importance of considering the characteristics of human operations in the design of Artificial Intelligence systems. In the computational study of visual processing capabilities it is important to recognize that while some aspects of human vision may be little more than artifacts of the biological implementation, other aspects may reflect fundamental properties of the underlying processes. At the physiological level, for example, it has long been known that a stabilized image on the retina will quickly fade and disappear

because of the properties of the retinal receptors, but it would not seem useful to build this characteristic into computer based imaging systems. On the other hand, properties of the the spatial organization of receptors on the retina corresponds directly to the characteristics of one of the most successful computational edge detection operators (Marr and Hildreth, 1980).

Consideration of the characteristics of human vision may also be productive in other respects. The body drawing interpretation system has been structured to incorporate several aspects of human vision, and within reasonable limits, an attempt has been made to provide connections and analogies between the system's operation and results obtained through Cognitive Psychology experimentation.

Chapter Two provides a framework and background to the approach which has been taken. Particular care has been taken to describe the interpretation labelling approach to computer vision, which places an importance on a variety of features and their role in suggesting models for objects depicted in the image. This approach is described in the context of research in line drawing interpretation.

Chapter Two also furnishes a computational perspective on the use of multiple resolution representations, and on the related topic of the selectional processes of vision. These areas form the basis of the system's consideration for human

visual processing.

The third chapter is an overview of the research, giving a model for perception, and a description of the mechanism for encoding visual knowledge, called declarative schemata. This overview is presented without reference to the specific problem domain.

Most of the important ideas behind the computer implementation require examples for their description. Chapter Four presents the operations of the system, going through each stage in detail.

Chapter Five is an abbreviated demonstration of the working system. An example was chosen for the resulting clear demonstration of the processes described in Chapter Four. An appendix provides other examples.

During the design and implementation of the body form interpretation system, several issues came to light which relate to previous research, both in Computational Vision, and in Cognitive Psychology. Chapter Six discusses these issues. Chapter Seven concludes with a summary and suggestions for future research.

## 2. Framework and Approach

The research presented in this thesis is related to a variety of established avenues of investigation into the nature of visual processes. This chapter singles out four aspects which require discussion in order to develop a groundwork of ideas and terminology. The first topic is concerned with the point at which knowledge of specific objects might enter into the process of visual interpretation, with emphasis on the potential role of two dimensional image features in cuing such knowledge. The second presentation describes research centered on line drawings, exploring some of the limitations of the early problem domains, and following the progression of results to the development of "schemata-based" interpretation methods. The third issue is the use of multiple resolution levels in visual interpretation, which includes proposals for interpretation-based interactions among levels. The fourth topic is the selectional processes involved in visual interpretation. Research in saccadic eye movements is explored, with an attempt to uncover some of the computational bases of selection in human vision.

There are several other areas of investigation which relate to the current research. Appropriate discussion of these topics is deferred until after the elaboration of the implemented computer vision system provided in chapter four.

### 2.1. Use of Model Knowledge in Vision

Computational vision may be distinguished from its ancestral disciplines by its concern for the variation in appearance of objects when imaged. The two major contributors to this variation are: (1) the many possible viewing conditions, including the diversities of lighting, and (2) the possible variations in the objects themselves, including their deformation and arrangement. The techniques of correlation matching and of feature-vector classification have been discarded in favour of the development of methods which incorporate explicit models of these factors influencing the image.

Early computer vision research may be broadly categorized as an attempt to match image feature information against the features predicted by models of objects and thereby develop representations of imaged scenes. Later research has centered on the use of image features in the construction of more complete context-free representations to be later matched against models of specific objects. During the transition, the idea of accessing models early in the visual process on the basis of image features has fallen somewhat into disfavour.

This section begins by examining these two approaches, with emphasis on underlying perspectives on the process of perception. A model of visual perception is then presented which has provision for both approaches. Finally, a presentation is made of some arguments against the currently more

popular view that elaborate representations of scenes are constructed before knowledge of specific objects is involved.

### 2.1.1. The Image Feature Access Approach

One of the earliest computer vision systems was developed by Roberts (1965) to interpret photographic images of simple polyhedral objects. This research established three important approaches to computer vision.

(1) The problem domain of polyhedral objects known as the blocks-world became a focus of much research which followed. The domain provides possibilities for variations of view, configuration, and shape of objects which may be simply modeled geometrically.

(2) The system operated in two stages. The first step was to develop a line drawing from the digitized image by grouping intensity discontinuities. The second step matched geometric models of known objects against the line drawings. The notion of an intermediate line drawing stage in computer vision has been popular ever since.

(3) The matching phase exploited the fact that there are topological invariances in the projections of the polyhedral object models over simple transformations and changes of viewpoint. Thus object models could be suggested by the detection of image features. For example, a parallelogram suggests either a wedge or cube. This is the basis of the



image feature access approach to computer vision: that simple image features invoke the examination of more complex object models which may then be verified or rejected. In various forms, this approach has received a great deal of attention in computer vision research.

The research which followed Robert's work focused on the development of structural descriptions of blocks-world scenes on the basis of features extracted from line drawing images (Guzman, 1968; Clowes, 1971; Huffman, 1971). These systems used lines and vertices as image features, and exploited the relation between these features and scene properties. Soon after, effective methods were developed for understanding blocks-world line drawings. This research explored the use of local consistency methods (Waltz, 1972) and gradient space (Mackworth, 1973) in computational vision.

Section 2.2 describes a number of line drawing interpretation systems which use the image feature access approach. It includes a discussion of some of the issues which have advanced the research to other problem domains, and introduced more elaborate techniques.

One important aspect of image feature access systems is that knowledge of objects is introduced early to guide the interpretation process. This poses a significant problem: If models are to guide interpretation, how can the system employ the correct models until the scene has been interpreted? This

has been referred to as the "parsing paradox" (Palmer, 1975) and as the "chicken-and-egg problem" (Mackworth, 1978).

One solution to this paradox is to consider perception as a cyclic process rather than linearly staged. Mackworth (1975; 1978) has proposed such a cycle of perception consisting of four steps; cue discovery, model invocation, model verification, and model elaboration. The idea is that the cycle may start either with or without hypothesis of models, and gradually, as the cycles are completed, develop more refined correspondences to existing models, and thereby accomplish recognition.

A similar model for perception has been presented by Neisser (1976). This model is centered on representations of anticipations about visual information called "schemata". These schemata are modified by accumulated inputs and in turn direct exploration of the visual field for further input relevant to the schema's objectives. The cycle may be initiated either by stimuli or by anticipations.

This idea that model knowledge becomes more and more specifically useful as interpretation progresses is inherent in most knowledge-based vision systems (Mackworth and Havens, 1981; Hinton, 1981; Brooks, 1981; Browse, 1982).

Another type of solution, proposed by Palmer (1975) and by Havens (1976), is to develop knowledge structures and techniques which enable simultaneous hypothesis-driven and data-

driven searches.

### 2.1.2. The Volume Access Approach

A different approach to computer vision has emerged on the basis of the work of Horn (1975) and Marr (1976). Horn introduced the use of a mathematical formula relating physical characteristics of a scene (such as surface orientation and light source position) to the array of light intensities which results. This work has inspired attempts to recover knowledge of such physical characteristics on the basis of an intensity array by making additional assumptions about properties of the objects (Woodham, 1978; 1981; Witkins, 1981; Stevens, 1981).

Marr argued for modularity in the construction of vision systems with distinct intervening representations. This has also had widespread acceptance within computational vision research (see Brady, 1982). One of the basic premises of the work of Marr is that a large and complex computation (such as vision) must be split up into small, nearly independent specialized sub-processes (Marr, 1976; Marr and Nishihara, 1976). The major justification for this view is that such an organization is necessary to evolve a complex system; that otherwise an evolutionary change to improve one aspect would degrade another.

Modularity requires sub-processes which interact minimally with one another, and implies strong representational structures through which the modules may transfer

information. The most obvious mode of operation for such a linear-stage system is a strict bottom-up processing paradigm which defers the involvement of specific model knowledge until a complete three dimensional context free representation has been developed (see also Nishihara, 1981). This volume access approach is consistent with the recovery of physical characteristics of the scene through the use of the image formation equation (Horn, 1975; Barrow and Tenenbaum, 1978). Taken together, the result is an approach which relegates the use of specific model knowledge to a point after the development of elaborate context-free scene representations.

### 2.1.3. Discussion

Figure 2.1.1 is a schematic drawing of a simplified model for visual perception which reconciles some of the differences between "image feature-access" and "volume-access" approaches to computer vision. In this model, a series of processes transforms an image through intermediate representations. The early stages are image feature based, the later are volume and surface based. At each stage in this progression, some general knowledge and assumptions are necessary, depending on the type of transformation. These are depicted on the right hand side in figure 2.1.1. On the left hand side is shown the structures encoding knowledge of the specific objects and object categories. This knowledge has access to every level in the progression, and may influence any of the transformations.

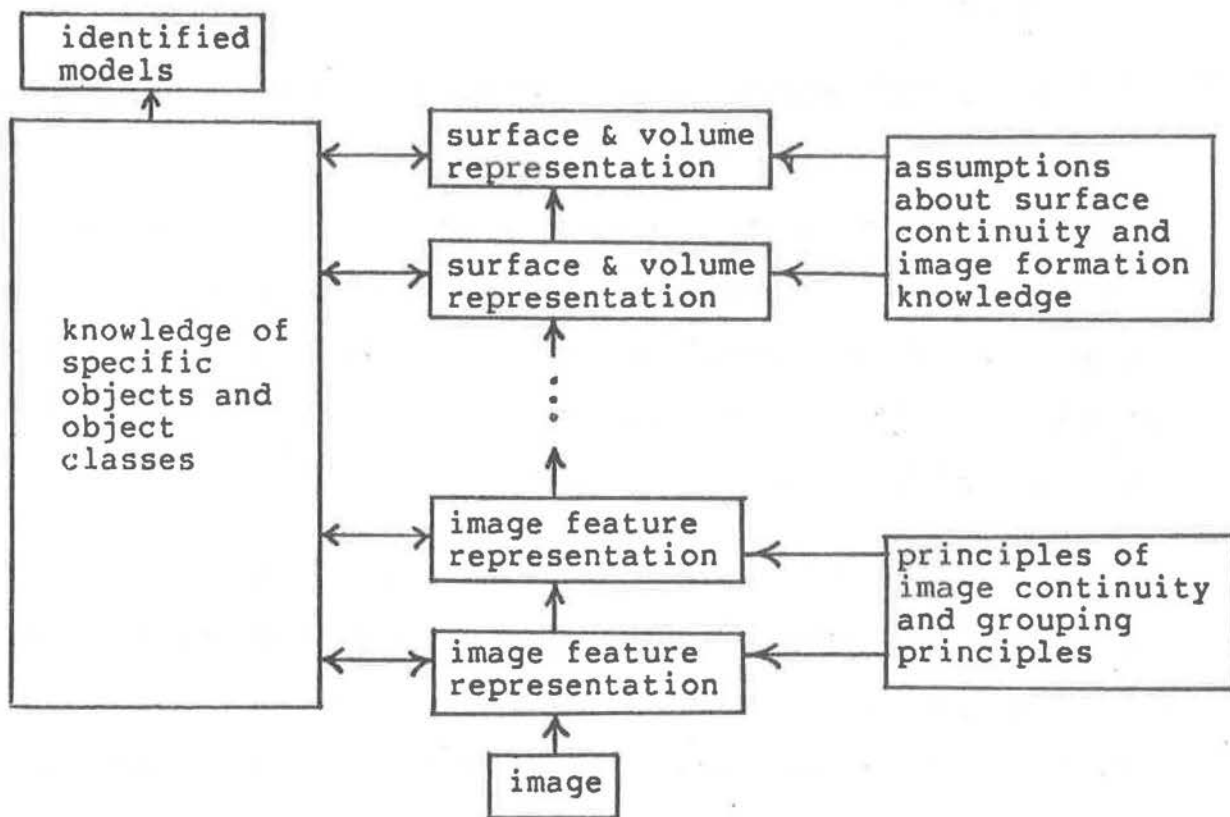


Figure 2.1.1. The Multiple Access Model of the use of object knowledge in visual perception.

The operation of this model involves the principle of least effort: visual processes will form a correspondence to known specific objects as quickly as possible, on the basis of any representation which can provide support. This means that in the case of line drawings and impoverished image situations, two dimensional cues extracted at the image feature level will be used to invoke models which will fill in the details at the level of surfaces and volumes. Under optimal viewing conditions, for unfamiliar objects, the most expedient route might be through a context-free surface-and-volume

representation.

During the course of perception, interactions will take place with the specific model knowledge. Features extracted early may cue these models which then make preparations for the development of representations at higher levels. Thus the model implements a cycle of perception similar to that of Mackworth (1975) and Neisser (1976), while retaining linearity and modularity of representation.

It is easy to see that this "multiple-access" model can reduce to the "volume-access" model (see Barrow and Tenenbaum, 1981) by removal of all connections between image feature representations and specific model knowledge. Figure 2.1.2 depicts this model. The research relating to the "volume access" model is centered on the examination of the role of the knowledge relating to each level, and so it is a reasonable step to not consider these lower connections.

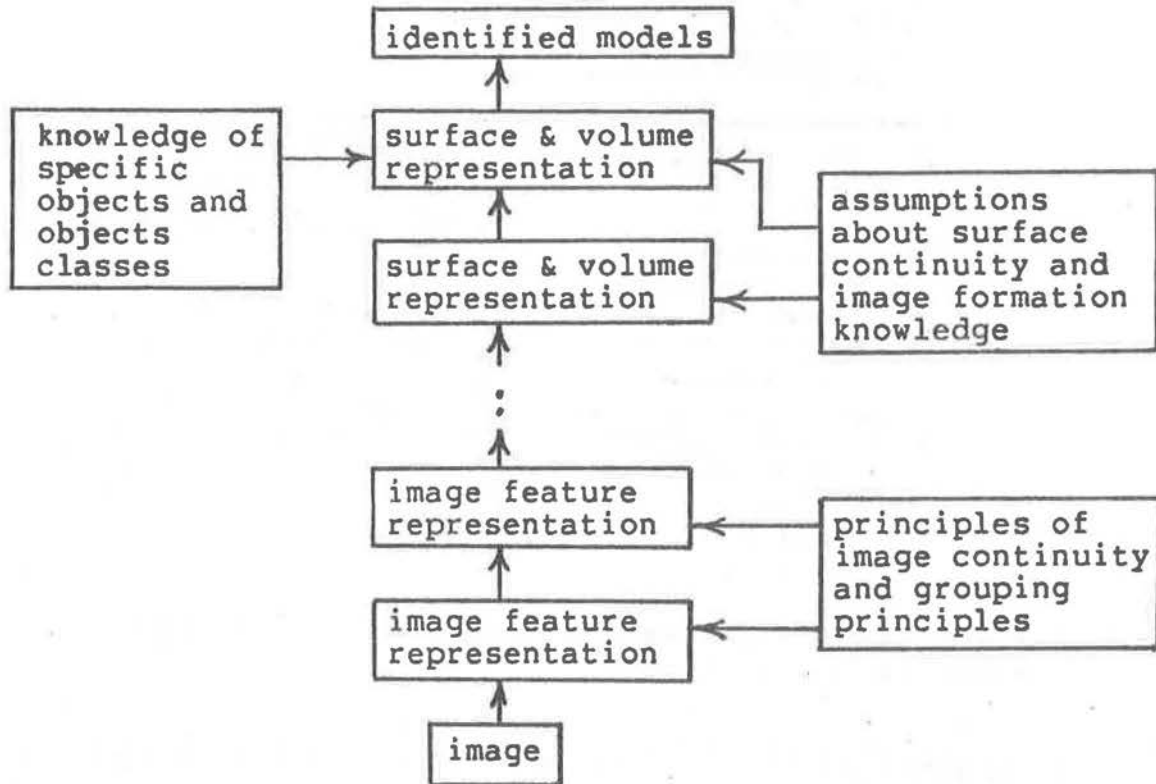


Figure 2.1.2. The Volume Access Model derived from the multiple access model.

The "multiple access" model will convert to the "image feature access" model by delaying the development of volume-based representations until after objects are identified (see figure 2.1.3). This step is required in the examination of images which are impoverished so as to not contain enough information to enable development of volumetric representations without the use of object knowledge. In this case, the description of the scene in terms of surfaces and volumes is viewed as a part of the correspondence to the object models.



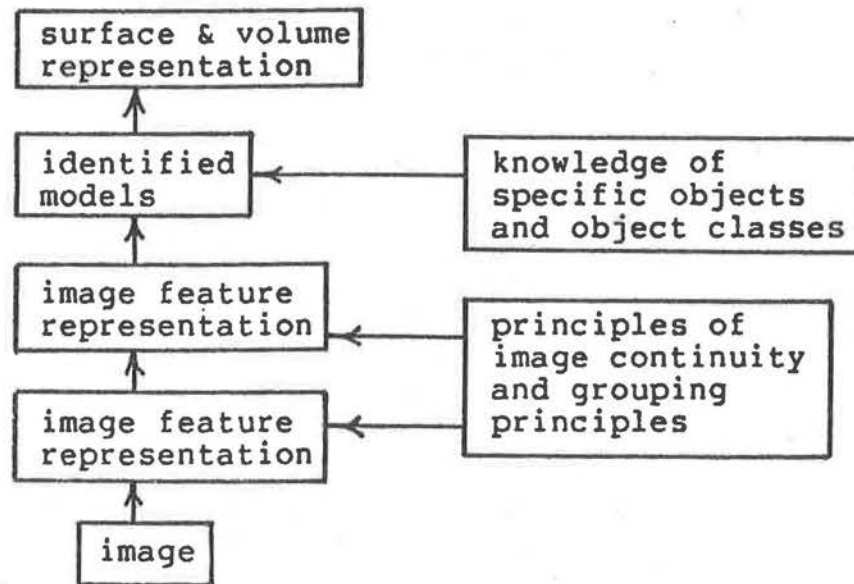


Figure 2.1.3. The Image Feature Access Model derived from the multiple access model.

Proponents of the "volume-access" model have cited examples of the perception of unfamiliar objects such as microphotographs of pollen (Barrow and Tenenbaum, 1978). The argument is that since few of us have specific models for such objects, and since we do seem to "understand" the images in terms of the surfaces and volumes, therefore such representations play a vital role in human perception. Of course, this type of demonstration only shows that human vision is capable of developing a model-free three dimensional representation, not that it must. Further, it is possible to argue that such representations are formed with the aid of models of analogous objects. At any rate, the demonstrations do not preclude the involvement of specific object models, invoked from the level of image features, influencing the development of three dimensional structure during the course of normal perception.

Another argument for the "volume-access" model is centered on the research of Warrington and Taylor (1973; 1975). This research has shown that some patients who have suffered parietal lesions are able to understand the shapes of objects even though they cannot name them or explain their use. From this Marr (1982) has concluded that shapes may be determined, even in difficult cases, without the intervention of specific models. There are two points of caution in formulating conclusions on the basis of this type of research: (1) no two lesions are the same, and it is difficult to generalize from the characteristics of the condition, (2) it is possible that the patients are impaired in their ability to report the object, even though the visual model structured around the object is still being used in visual perception. There are such cases, for example, in which the patient is unable to report seeing a "telephone", but uses terms such as "dial" in its description (Schmidt, note 1).

The structure of the human eye casts serious doubt about the possibility of inferring three dimensional scene properties on the basis of a single retinal image. Beyond the small foveal center, visual acuity rapidly diminishes toward the periphery[1]. This effect is a result of the organization of

---

[1] There is a variety of measures for visual acuity. Riggs (1965) describes a typical result: at 10 minutes of a degree off the fovea, acuity is reduced by 25%, and at one degree, by 60%. This reduction follows the pattern of decreasing density of cones in the retina.

the retinal receptors, and as well a result of the scattering of light by the lens and cornea (Haber, 1978). Thus in only a small portion of the visual field is there available the type of high detail input necessary to discern surface orientation context-free. This means that a number of fixations, consuming about a third of a second each, would be required before most objects subtending extensive visual angles could be identified. Yet, there is conclusive evidence that the progress of interpretation based on specific models of objects influences the selection of fixation locations (Mackworth and Morandi, 1967; Antes, 1974; Parker, 1978). Loftus and Mackworth (1978) have demonstrated that even the first saccade is highly dependent on the results of interpretive processing.

In a related experiment, Friedman (1979) has shown that subjects do not fixate as long on objects which are more consistent with the entire scene, and as a result have less detailed recollection than for objects which are unexpected within the context. These results argue for the early use of knowledge, not only of individual objects, but also of entire scenarios. Many other Cognitive Psychology studies support this view that preliminary interpretations, based on global and coarse image properties are utilized in the extraction of fine details. (Weisstein and Harris, 1974; Palmer, 1975; Biederman, 1981).

The interference effects discovered by Bruner and Potter (1964) are also interesting evidence of the human visual

system's willingness to form an early hypothesis about the nature of the scene. Subjects were asked to identify the content of scenes depicted in slide presentations. The slides were shown initially out of focus, but gradually becoming more clear. At a specific point this focussing process was stopped, and the subjects were asked to identify the scene. The length of time the subject was exposed to the defocussed image was varied, and it was found that the more exposure to the defocussed image, the less likely to identify the scene correctly. The accepted interpretation of this surprising result is based on the idea that while viewing the defocussed image, a number of tentative, conflicting hypotheses are developed about the scene. These hypotheses then interfere with the formation of an understanding of the more clearly focused image. This indicates that models and scenarios are invoked to assist in the development of interpretations when only very impoverished image information is available.

It is difficult to explain the findings of Gilchrist (1977; 1980) in terms of a strict linear stage process. His results show that the perception of brightness is interrelated with perceived spatial arrangement and orientation of surfaces. This is not so much an argument against building surface representations on the basis of more primitive aspects of the image, such as intensity, but rather it is an argument in favour of the inclusion of a mechanism which enables different representational levels to influence one another.

It is not surprising that many cognitive psychology results are consistent with the idea of accessing specific models on the basis of two dimensional features. Most experimental situations involve two dimensional presentations, often in line form, with specific response selections required of the subjects. However, the established psychological validity of this view of perception paves the way for research in the computational structures of vision which may benefit from the large collection of clues inherent in the experimental results.

It is an unfortunate fact that there is no unequivocal definition of the task of vision. Yet computer programs must have clearly defined inputs and outputs. There is little controversy over the nature of the input to vision, but the output specifications of each computer vision system constitutes a commitment to the objectives of vision. The "image feature access" approach implies that vision is the formation of correspondences between images and known objects and situations, and as such the representations are concerned with what is necessary to compute. The "volume access" approach implies that vision is the development of objective context-free representations of the depicted scene, and as such is concerned with what is possible to compute from the image.

The result of differing task definitions is different simplifying assumptions. The "volume access" approach avoids the "chicken and egg" problem with a simplified overall

control structure in the retention of a modular, and linear stage view of perception through elimination of early involvement of specific objects and scenarios. The "image feature access" approach often utilizes the simplification of a clean line drawing input in order to facilitate the formalization of the (usually cyclic) interactions within a limited realm of models of known objects. Neither simplification results in a "general vision" system. The specific set of lighting and surface conditions that must be obtained for context-free volume and surface representations are no more likely to occur in a scene than some specific object. The "multiple access" model makes the nature of these assumptions clear within the context of a more realistic view of perception which includes both the linear stages in the development of representations and the cyclic interaction with specific models.

The arguments presented in this section have been biased towards the "image feature access" simplifications, partly because the research outlined in this thesis follows that tradition, and partly because the approach has recently been in disfavour. The arguments should not be taken as attempts to demonstrate the correctness of one approach, but rather as an effort to further the search for a means of combining approaches towards a coherent model for perception.

One final analogy is irresistible. The study of natural language experienced a great influx of ideas with the introduction of phrase structure and transformational grammars

(Chomsky, 1957; 1965), which produced widespread and diligent computational study within Artificial Intelligence. The conclusions were that structure based on context-free general categories offered a useful dimension in language analysis, but that the real key to understanding language use requires the study of semantic and pragmatic knowledge of concepts and scenarios.

For computational vision, the stringency of the constraining assumptions necessary to operate without specific knowledge, and the evidence based on cognitive psychology studies of vision point to the requirement for the use of detailed knowledge of objects and organizations of scenarios.



## 2.2. Line Drawings in Computer Vision Research

The particular class of objects around which this research is centered is line drawings of human-like body forms. This section is concerned with line drawings in general, and the computational vision research which has been aimed at their interpretation.

### 2.2.1. Line as a Symbolic Level

Computer Vision and Natural Language Understanding are two areas of Artificial Intelligence which can be viewed as attempting to attain a computational understanding of some aspect of human intelligence by studying perception. Natural Language Understanding has one advantage in that there exists a clear symbolic level, the level of words[2], which may be assumed in order to study the involvement of human intelligence and experience in language use[3].

There is certainly no clear counterpart in computational vision research. This is perhaps because there does not exist an appropriate level, but on the other hand, the level of line drawings augurs well as a candidate. As objects are represented in line form, the aspects which are less important to be aware of in a scene, such as light source location,

---

[2] It may be argued that morphemes are a better choice.

[3] The understanding of language influences the perceived input to a lower level than that of words, but it is accepted that little context-free processing takes place past this level.

surface texture, and shadow are discarded just as a representation in words discards intonation and inflection.

The line holds a place of particular esteem in human activities. Line drawings appeared in caves around 10,000 B.C., progressed into hieroglyphic communication, and finally formed the characters of writing. Many visual communication devices, such as maps, text book illustrations, flow-charts, and circuit diagrams are largely line-based. This tendency toward the use of line may be related to the human vision system's well known sensitivity to intensity boundaries. The mental structures which encode and operate on visual information may themselves be tuned to line-like structures (see Marr, 1976).

Computational Vision research based on line drawings has several potential benefits:

- (1) Inasmuch as line drawings are involved in human communication, it is of both practical and theoretical interest to study their interpretation (see Mackworth, 1977b).
- (2) Even if line drawings are not adequate intermediate representations for the human vision system, it may be expected that studies which develop methods for the application of model knowledge in the interpretation of line images will probably provide insight into the methods required to process on the basis of some more refined, and perhaps more realistic intermediate

representation.

### 2.2.2. Some History

Following on the work of Roberts (1965), Guzman's (1968) program used heuristic and symbolic methods in an attempt to interpret line drawings of blocks-world images. A classification of image vertices was devised, and regions bounded by arms of the vertices were studied for the possibility of their belonging to the same object. This information was used to group regions which composed individual blocks. Although many organizations of blocks were interpreted correctly, there were many that the system could not handle (see Winston, 1972).

Clowes (1971) and Huffman (1971) recognized that a variety of edge types (convex, concave, occluding) in a blocks-world scene are all depicted as lines in the image, and that the classification of vertices provided by Guzman reflected a variety of corners and abutments of blocks in the scene. This distinction between the image and scene domains was carried further in the realization that only certain line interpretations (as edges) were possible for the lines composing each vertex type. Thus each line was assigned a set of interpretation possibilities (or labels), and the vertices could be used to enable a search for the appropriate label for each line.

Waltz (1972) expanded on this theme by considering more labellings, including those for cracks and shadows. Vertices

were viewed as nodes of a network, with each node having an associated set of label possibilities. Using the uniform constraining relation that straight lines must have consistent interpretation over their extent, a filtering operation removed impossible labels towards a much reduced, and often unique interpretation.

This approach will be referred to as the interpretation labelling approach. There are two fundamental ingredients:

- (1) Some local image elements (such as lines) are assigned lists of labels, indicative of the roles that the elements might play in the structure of the scene.
- (2) Relations among elements are identified which serve to constrain the label lists, and techniques are devised to propagate these constraints.

The constraint propagation technique described by Waltz (1972) has been generalized to network consistency algorithms by Mackworth (1977c), who also argues for their general usefulness in tasks such as computer vision. The idea behind network consistency is that a constraint satisfaction problem is specified as a network, whose nodes are variables with associated domains of possible discrete values. The relations required between variables are represented as directed arcs of the network. In order for a network to be arc consistent, all variable values must be locally possible: for example, for the relation  $P_{ij}(x,y)$ , for each "x" in the domain of values at

node  $i$ , there must exist a "y" in the domain of values at node  $j$  such that  $P_{ij}(x,y)$  is true. This does not guarantee the existence or uniqueness of a complete solution, but an arc consistent network may be searched for solutions with an expected reduction in thrashing behavior (see Mackworth, 1977c). In developing an arc consistent network, the arcs are examined one at a time, and revised by deleting domain values which are not locally possible. After an initial pass through the arcs, only those arcs that lead into a revised node must be reconsidered in a relaxation process.

### 2.2.3. Generalizing from the Blocks-World

There are some specific aspects of the blocks-world which make its interpretation labelling formulation particularly simple. To extend the use of these concepts to other problem domains requires significant alterations of the techniques. The following outlines four such aspects of the Clowes/Huffman/Waltz blocks-world solution, and serves as preparation for a subsequent examination of some schemata-based systems, whose objectives include addressing the more general problems of applying knowledge of objects in more complex domains.

#### 2.2.3.1. Compressing Constraints to a Single Level

In the blocks-world solution outlined above, constraining relations from different types of information are represented in the same form: as vertices with legal interpretation

possibilities. Consider, for example, figure 2.2.1a. This shows a cube, viewed in such a way that its upper protruding corner appears as a "T" vertex. One valid labelling for the vertex is therefore as two occluding edges (indicated as arrows) and a central convex edge (indicated by a plus sign). This is entered into the pool of legal configurations for a "T" vertex, and reflects a property of an individual block in isolation. Figure 2.2.1b shows a similar vertex formed by two adjacent blocks. Thus another legal configuration for the "T" vertex is established[4], but this time on the basis of the way blocks interact.

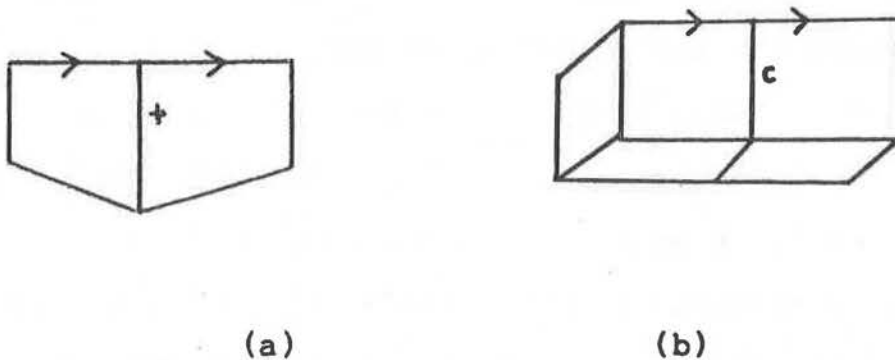


Figure 2.2.1. Two labelled vertices from the blocks-world.

In general, constraints based on different aspects of the structure of a problem domain must be expressed separately. Consider a problem domain such as that of geographic sketch

[4]The "c" label indicates a crack.

maps (Mackworth, 1977b), as shown in figure 2.2.2. The formulation of the idea that two lines must meet in a specific way to become a mountain symbol can be accomplished at the line level. To specify the requirements for mountain symbols combining to make a mountain range one requires more complex objects and their attributes. Another level still is necessary to indicate how a river combines with a mountain range to form a river system.

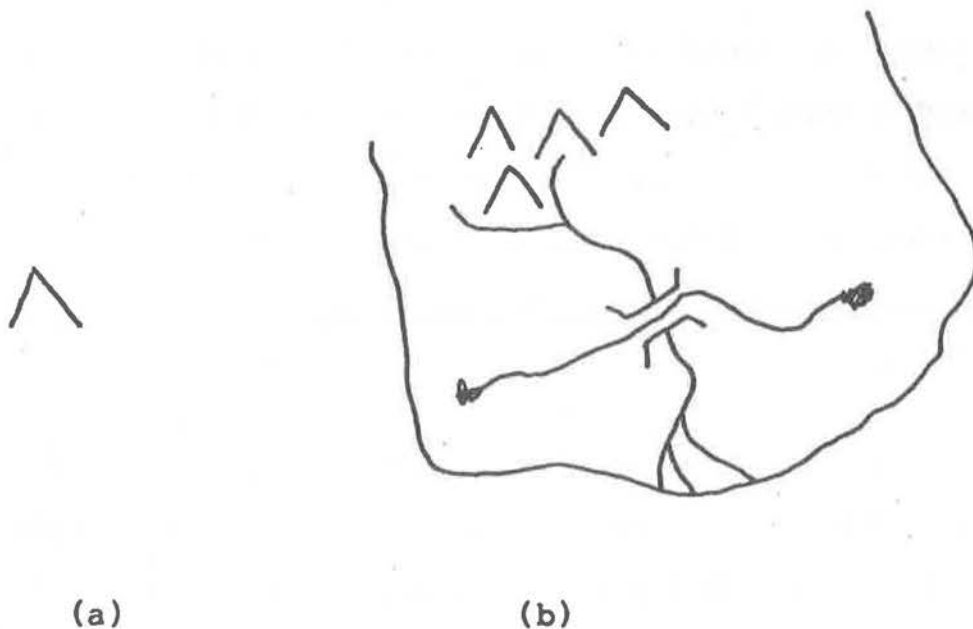


Figure 2.2.2. (a) A mountain symbol (b) A sketch map.

As a consequence of the compression to a single level, it is possible to use a uniform vertex-finding method to locate all relations among lines. In the more general case the relations among more complex objects must follow the discovery of the primitive relations among lines.

#### 2.2.3.2. Different Types of Features

It is particular to the blocks-world that the lines which act as basic features, have no structure or attributes which suggest interpretation[5].

In natural images, there is a rich assortment of information available. Marr (1976) makes the point that it is important to represent a variety of feature types, and to specify their attributes. This attitude is reflected in the nature of the "primal sketch", which encodes several different edge types with attribute values for such aspects as length, width, and orientation. This same view is inherent in psychological studies aimed at the identification of "feature dimensions" along which feature values may vary (Garner, 1974). Lines and vertices of the blocks-world are the only image elements of concern.

Even in terms of line drawing images, it will generally be the case that individual lines may be assigned attribute values. For example, curvature, orientation, and length may be important aspects of these features in some other domain. In the blocks-world curved edges do not exist, and neither orientation nor length have any constraining force on the roles that the line may play in its representation of the scene.

---

[5] There is an exception in Waltz' system which considers an attribute of shading edges (which side is darker).



In the more realistic situation of many different types of features, each with its own attributes, the issue of selection becomes important. What features are necessary to interpretation, and which attributes provide the strongest constraints? Certainly in human vision, attentional mechanisms operate towards resolving these problems.

#### 2.2.3.3. Availability of Components

In the blocks-world, relations among lines are uniformly available: If two lines are found to connect, it is a simple matter to check for other connecting lines and thereby complete the relation. In the general case, all components of a relation may not be available, either because the missing component is itself a more complex object, or because the relation cannot specify the means of obtaining it from the image. Even incomplete knowledge of a relation, however, might be enough to serve as a constraint upon the interpretation of features entering into the relation.

#### 2.2.3.4. Structure Within Labels

In the blocks-world, edge labels are assigned to lines as interpretation possibilities. The labels exhibit a structural organization, though most blocks-world interpretation systems do not exploit it (see Mackworth 1977a). Generalizations over groups of possible labels can be either filtered or retained as a group through the consideration of a single relation, rather than considering the relation over each element. This

inherent organization of possible labellings is more apparent in problem domains such as sketch maps (see Mackworth and Havens, 1981).

#### 2.2.4. Beyond the Blocks-World

The previous sub-section has reviewed four issues in the use of model knowledge in the interpretation of line drawings, which are not inherent in the blocks-world problem domain. These issues are not unique to line drawing domains, but through the assumption of the availability of clean line drawing input makes the issues emerge as addressable in the context of other problem domains.

Computer vision systems have been implemented to examine the more subtle aspects of applying specific model knowledge to visual processing. Together they are often termed "schemata-based" systems because they embody some ideas behind the variety of psychological models of cognition which go by the same name (Bartlett, 1932; Piaget, 1967; Neisser, 1976).

There are three main ingredients of a schemata based vision system:

- (1) object centered knowledge.
- (2) use of the natural structure of the domain.
- (3) recursive cuing mechanism.

The following brief review of such research will be aimed at the explanation of these concepts.

Mackworth (1977b) extended the basic idea of interpretation labelling to a system to interpret geographic sketch maps. One important innovation was that the features were not uniform: both line chains and regions acted as features. The interpretation possibilities assigned to these features were common objects of the problem domain. For example, a line chain could have any of the interpretations {road, river, mountain, bridge, shore}. The line chain is then said to act as a cue for any of these interpretations. The system accomplished interpretation through a two-step segmentation and network consistency cycle.

The movement towards using common object types as the basis for encoding knowledge about the problem domain was carried even further in the recognition model devised by Havens (1978). He devised a programming language "Maya" in order to represent the knowledge necessary to accomplish model-based vision. These procedural schemata held together everything known about individual objects in a way similar to the "frames" proposed by Minsky (1975).

The structural framework for encoding object knowledge is the natural structure of the objects themselves: the component and specialization hierarchies. In Havens' model, the component hierarchy defines a recursive cuing mechanism. This

means that just as a basic image element may cue an intermediate structure, the confirmation of that intermediate structure acts as a cue for some more complex structure. In the sketch-map domain, this means that the "line-chain" has as its possible labels {road, river, ..}, and that "river" has as its possible label "river-system" which in turn cues "geosystem".

This system, MAPSEE2 (Mackworth and Havens, 1981) also provides a means of grouping labels according to the specialization hierarchy of the problem domain. For example, the relations between regions on either side of a "shoreline" may be evaluated with respect to the labels "landmass" and "waterbody". Only later on is it necessary to specialize these regions to "island" or "mainland" and "lake" or "ocean".

The use of the component hierarchy in computer vision is quite straightforward. It has been used in numerous models of perception, providing a clear indication of its benefit. The specialization hierarchy poses more difficult problems. This hierarchy may be structured on the basis of distinctions such as functional similarity, visual similarity, or criterial property. It is not clear which criteria are suitable for encoding visual knowledge. Further problems are found in trying to establish the role of specific entities, which may be viewed as the leaf nodes of the specialization hierarchy (see Mulder, note 2).

All schemata-based systems described thus far are procedural in nature. These procedures encode both the requirements for objects, and the actions to be taken to obtain an instance of themselves. This procedural approach is productive in experiments aimed at discovery of the basic principles of how knowledge should be structured for vision because, it is easy to modify and test small segments when they are represented as procedures.

One step in the development of schemata-based systems is to move towards a more declarative knowledge base. That is, to separate the knowledge of the objects from the knowledge of the processes that effect interpretation. Such a development would have a number of advantages, which are described in section 3.2.

### 2.3. Levels of Resolution

In the development of an image, a planar projection of reflected light from objects and surfaces is always represented in discrete terms. A number of individual picture elements cover the area of the image. These elements could be the light sensitive silver halide crystals used in photographic material, the array of responses of the retinal cells of the eye, or the coordinates of imposed grids in digitization processes. In each case there is always a resolution associated with an image: the number of picture elements per unit area.

There is a variety of evidence in favour of approaching vision as a process which operates over several different, but related levels of resolution. Neurophysiology, Psychology, and Computer Science all contribute towards this approach. Naturally there is some disagreement, particularly in terms of the level of processing at which information from different resolution levels interacts. For some, multiple resolution is a tool in the discovery of context-free image features such as edges. Others believe that the structure and organization of object knowledge is related to the availability of several levels of detail. This section reviews and contrasts some of these ideas.

### 2.3.1. Resolution Levels in Natural Vision

Measurements of cell responses in the early portion of the primate visual system have demonstrated selective sensitivity to a variety of retinal field sizes. The orientation-independent responses of the center-surround fields encountered at the ganglion and geniculate cells, and the more specifically sensitive simple and complex cells located a few synapses away in the primary visual cortex, are both examples of receptors which exhibit a variety of field size response (Hubel and Weisel, 1979).

As retinal eccentricity increases, average field size systematically increases. This effect is attributable to the varying density of retinal and ganglion cells and the variation in convergence of signals between them. This relates to, but does not completely explain the change in visual acuity with eccentricity (Westheimer, 1982). At a single point on the retina, there is an overlap of receptive fields of different sizes.

The spatial extent of the retinal center-surround fields determine the types of edges which may be detected. For example, a wide receptive field will not respond to closely spaced lines, and small receptive fields will not respond to gradually changing intensities. The different field sizes may be viewed as encoding intensity discontinuity information based on different resolution levels because of the associated vari-

ation in the number of retinal receptors.

Psychophysical experimentation has developed an analogy between spatial frequency analysis and the variation in receptive field size. A large receptive field size corresponds to a low spatial frequency channel in the sense that, in either case, sensitivity is greatest for gradual intensity changes.

Experiments have been performed which rely on this analogy. Subjects who observe sinusoidal gratings for a few minutes exhibit an elevated contrast threshold to subsequent test gratings of similar spatial frequency (and otherwise identical), but show no such effect for test gratings of dissimilar spatial frequency (Pantle Sekuler, 1968; Blakemore and Campbell, 1969). This type of result is explained in terms of the selective desensitization of frequency specific channels at each retinal location in the human visual system. Wilson and Bergen (1979) have proposed four channels, each with a center surround profile described by a difference of two Gaussian distributions. Others suggest as many as seven channels (see Watson, 1982).

The spatial frequency analogy has also been useful in identifying two types of cell responses: sustained and transient. Generally, low spatial frequencies are transient and have been proposed as specialized for detection of temporal and global aspects of a scene, whereas the sustained high frequency channels are believed involved in form and pattern per-



ception (Breitmeyer and Gantz, 1976).

Another important result has been obtained through research in summation at threshold for spatial frequency channels. Stimuli composed of sinusoidal gratings of several different frequencies are only slightly more detectable than the most detectable of the composing gratings. This result is independent of the relative phase of the gratings (see Graham, 1981). The small enhancement of detectability is attributed to a probability summation model of detection: that each channel has an independent probability of detecting the pattern, and hence the potential detection by several channels increases the overall probability of detection. Given the small size of the enhancement, this model is preferred over one which enables the combination of information from different resolutions at an early stage in the vision system.

There is also a line of Cognitive Psychology research which is concerned with different levels of resolution. The issue centers around the order of processing at the different levels.

The traditional constructivist view of perception proposes the development of holistic properties on the basis of the results from fine resolution processing (Neisser, 1967). The opposing view is that high-order forms are processed initially, followed by the finer details (see Kahneman, 1973). Kinchla (1974) established what was to become one of the the

main paradigms in the investigation of this issue: subjects are shown a display consisting of a large letter, which is made up of many instances of a smaller letter (see figure 2.3.1). By varying the task between reporting the identity of the small or large letters, and by varying the compatibility between the letters at the two levels, researchers were able to address the questions of local-global interaction and ordering.

N N N N N N N N N N N N N N N	H      H H      H H      H H H H H H      H H      H H      H
(a)	(b)

Figure 2.3.1. Typical local-global stimuli: (a) incompatible  
(b) compatible.

Navon (1977) showed that in attending the large letters, the small can be effectively ignored, but that the presence of the large letter always influences the reaction time to identify the small, and thus established the concept of global precedence in perception. Others have demonstrated that such factors as absolute size, relative density, and quality of the letters will influence the results (Kinchla and Wolfe, 1979; Martin, 1979; Hoffman, 1980).

Miller (1981) altered the task somewhat to require subjects to detect specific letters, whether they appear at the

local or global level. Strong facilitation was found in the compatible condition. This result requires a model of perception in which information from both levels of resolution feed into a single decision process which integrates the results. Miller suggests that the integration may be based on attentional shifts between levels, with initial emphasis on the global level because of the guidance it is thought to afford in normal perception. This idea is consistent with other areas of psychological research which will be discussed in section 6.1.

### 2.3.2. Resolution Pyramids or Cones

Computational vision research has also been concerned with information at different resolutions. Kelly (1971) introduced the idea in a system to detect the outline of a human head in an image of background contours. A second image was developed consisting of one pixel for every 8x8 area of the original digitized image. Thus this extra image was much smaller, and did not have as much detail. Edge segments in the small image were compared to the coarse requirements of an image of a head, and then this information was used to guide search among edges in the original image to construct the detailed outline of the head.

This idea was extended to the notion of a recognition cone, or image pyramid (Uhr, 1972; Hanson and Riseman, 1975; Tanimoto and Pavlidis, 1975) which represent an image as

several interrelated layers constructed at different resolutions. The base level is the regular digitized image, and the upper layers are successively smaller images, with pixel values derived by some averaging operation on four (or more) pixels at the level beneath it. A number of processing schemes have been devised to use these structures to aid in the detection of image features. The basic idea behind the use of these pyramids is that indications of the existence of a feature may be found in a simple search of a smaller, coarser resolution version of the picture, which can then be used to direct the extraction of the feature from the finer levels (see Tanimoto, 1980). This idea has been generalized to systems which permit specification of parallel algorithms which operate with transferal of information in both directions in the image hierarchy, as well as laterally within a level (Hanson and Riseman, 1978; 1980). Levine (1980) describes a computer vision system which integrates information from separate pyramids used to encode a variety of image features.

### 2.3.3. Resolution Levels in Edge Detection

A similar idea is found in the work of Marr (1982; Marr and Hildreth, 1980). An image, smoothed with a variety of Gaussian filters, is convolved with the Laplacian operator. The zero-crossings of these convolutions are representative of the intensity changes in the image within different spatial frequency channels, dependent on the value of the space

constant of the Gaussian distribution. The response characteristics of these operators are similar to the difference of Gaussian operator proposed by Wilson and Bergen (1979).

Oriented zero-crossing segments are detected, and represent candidate edges. The results in different channels are then combined to produce a single representation of the image as the raw primal sketch, consisting of symbolic descriptions of segments, providing location and a number of other properties (see Marr, 1976). The process of combining the results from the different channels relies on the idea that zero crossings at the same location at different scales are probably a result of the same underlying physical phenomenon. So whenever the segments obtained at two or more (contiguous) channels agree in both position and orientation, an edge is hypothesized. Subsequent operations group these edge tokens according to several similarity measures in order to obtain tokens for larger scale areas of continuity and boundary.

We must question the use of a single location-based representation for tokens consisting of a variety of properties. In particular, we must question the early combination of information from several channels. It is quite a reasonable alternative to retain separate representations for each receptive field size, interconnected through convergence of location as in the case of image pyramids, the difference being that instead of containing averaged image intensities,

the pyramid would encode the more elaborate structure of zero-crossing segments.

There is a variety of reasons why this approach is more useful and more realistic:

(1) The area of vision investigated by Wilson and Bergen (1979) included only 4 degrees of eccentricity, which constitutes less than one percent of the visual field. Even within this area, the spatial extent of each visual channel doubles toward the periphery. In such a system of varying receptive field size, the outcome of combining results would be different at each eccentricity for the same stimuli. This would confound the task of detecting variation associated with changes in surface orientation.

(2) An important task during changes in fixation location is to form a correspondence between what is already known and the newly available information. With each change in location, there is a switching of the foveal and peripheral resolutions. If low resolution channel results for the fovea are maintained and elaborated separately, and not collapsed to a token at the finest possible level, then structures will be available to facilitate the establishment of correspondence.

(3) Basic to the idea of the primal sketch are representational tokens which tie together a number of properties (such as orientation and size) in a single retinotopic array. Recent experiments indicate that such combinations of

properties are not available in parallel over the visual field, but must rather be constructed through the sequential application of focal attention (Treisman and Gelade, 1980)[6]. Also, the perceptual grouping necessary to form boundaries is more difficult when based on conjunctions of properties rather than single properties (Treisman, 1982). These results favour the maintenance of a number of retinotopic arrays which can be used as the subject of grouping operations and may be accessed as necessary to consider the coincidence of features at particular locations[7]. It does not seem reasonable to take the step of consolidating several aspects of the available information into a single array and then apply grouping operations which must sort through the tokens in search of similarity.

(4) The threshold summation results previously described argue against the early combination of outputs from several channels, at least in terms of enhancing detection.

(5) The research using image pyramids has established that the computational advantage to using a variety of resolutions lies in the idea that coarse elements need not be precisely located, and so can be maintained in smaller arrays. The representation of zero-crossing segments could gain this

---

[6] A more complete discussion of this experimentation is found in section 6.1.

[7] Zeki (1978) has demonstrated that for the Rhesus monkey, projections from the primary visual cortex to prestriate areas are divided into retinotopic areas of separate features.



processing advantage through separation in terms of receptive field size.

#### 2.3.4. Knowledge Interaction with Multiple Resolution Levels

Most of the computer vision research described thus far has a common goal in the use of multiple resolution levels: to more accurately and efficiently extract features from an image. There have also been a number of studies which attempt an interpretation-based interaction between levels of resolution, using knowledge of the class of objects which comprise the problem domain.

The original work by Kelly (1971) falls into this category. The coarse level features are analyzed in the context of what was expected for the outline of the head, thereby ignoring the other prominent edges produced by the background.

Catanzariti and Mackworth (1978) applied a similar idea to the task of classifying regions of ground cover type from satellite images. A pyramid structure is developed from the image, and information from maximum-likelihood classifiers are passed across the levels of the pyramid.

Rosenthal and Bajcsy (1978; Bajcsy and Rosenthal, 1980) have extended the interaction between world knowledge and image hierarchy in an inquiry-driven computer vision system. The natural hierarchical relations of the problem domain are



explicitly encoded. These include the part-of relation, the class inclusion relation, and a size ordering relation. In investigating a query for a specific object, the system first devises a series of contexts from the objects found towards the root node in the part-of hierarchy. A search for these context objects is made at resolutions determined by the size relations. Since the part-of relation implies that the part lies within the spatial extent of the whole, each successful context search reduces the candidate search area at the finer resolution levels.

A model for perception has been presented by Palmer (1975; 1977) which is very similar in its theoretical position. The model proposes a structural hierarchy based on the whole-part relation, forming a network. Each node expresses its component structure as part-of links upon which further relational requirements may be imposed. Each level expresses holistic properties in terms of features at different resolutions, becoming lower for concepts towards the root node.

#### 2.4. Location Selection in Vision

As described in the previous section, the human perceptual system has the characteristic that receptive field size increases towards the periphery, resulting in a graded acuity. With a fixed number of receptors, this configuration provides both a wide field of view, and the capability for high resolution extraction of detail. The saccadic eye movements which accompany visual perception, are the actions which enable selective high acuity vision throughout the field of view.

At one point in the evolution of human vision, it is possible that the sole purpose of saccadic eye movements was to produce this enhancement of acuity. In fact, there remains a reflex action to fixate upon moving objects detected in the extreme periphery, even though we cannot be aware of their movement (Gregory, 1966). However, it seems a reasonable hypothesis that the structure of human intelligence has developed to be attuned to the sequences of high resolution input obtained through eye movements. It is also reasonable that the structures which aid in the understanding of scenes and objects contain the knowledge necessary to guide the process of selection to areas which will provide useful information.

There are a number of identifiable aspects to the selection processes that take place during human vision. Saccadic eye movements are among the most obvious and accessible to

analysis. Even when the eyes are not moving[8], other selection operations are in effect. There is the spatial allocation of an attentional mechanism which enables or enhances the extraction of visual information (see Posner, 1978; Treisman and Gelade, 1980). This spatial attention may be moved much more rapidly than the eyes, and has the property of being variable in its extent (Eriksen and Hoffman, 1972). Other, perhaps related, attentional mechanisms provide selective activation of memory structures which attune visual processes to the reception of particular image properties (Laberge, 1976; Shiffrin and Schneider, 1977), and still other mechanisms are thought to be involved in the selective preparation of responses (Kahneman, 1973).

The purpose of this section is to single out saccadic eye movements as representative of the selectional actions of perception. The basic characteristics of saccades will be reviewed with the objective of emphasizing the non-arbitrary nature of the selection of fixation locations. The steps involved in selecting and moving to fixation locations will be outlined, with the objective of exposing the computational requirements. Finally, some theories and computer simulations of saccadic eye movements are discussed.

---

[8] When fixated, the eyes undergo a number of small shifts, drifts, and tremors.

### 2.4.1. Saccadic Eye Movements

During normal viewing of a picture, humans move their foveal vision over angles up to 15 degrees about 3 times per second (see Yarbus, 1967; Gould, 1976). Of this viewing time, about 90% is spent in fixation (Yarbus, 1967). The actual eye movement is caused by the application of the full force of the eye muscle, where the duration of the application determines the distance covered (Alpern, 1972). The saccade is "ballistic", in that it cannot be corrected once initiated<sup>[9]</sup> (Westheimer, 1954). A minimal amount of information is picked up during the saccade itself (Latour, 1962). These two facts indicate that during a fixation, the visual system must be both extracting visual information, and preparing for the next movement. The following is a scenario of the steps which might be required, starting from the point of the eyes coming to rest at a location:

- (1) The first problem is to determine if the saccade was effective in placing the fovea at the desired location. Such errors are likely detected within the ocular muscular system, and may result in a small, corrective saccade (Yarbus, 1967:134).
- (2) If one accepts the notion that some internal model of the visual field is being maintained, then an updating of

---

[9] This is not the case with other forms of eye movements such as convergences.

that representation must be accomplished to establish the continuity of perception.

- (3) The new periphery must be analyzed, and results compared with previous interpretation results. There is evidence that the periphery is processed from the outside in (Lowe, 1975), and there are suggestions that it is done before the fovea is analyzed (Parker, 1978).
- (4) Foveal feature information is extracted and used in the enhancement of the ongoing scene interpretation.
- (5) The next location must be selected for fixation, and the exact muscle "program" must be developed. It has been shown that the more precise the saccade must be, the longer the latency to the eye movement, and so presumably the longer it takes to compute the parameters of the movement (Leushina, 1965).

Research into the nature and determinants of eye movements raise two interesting issues from the point of view of the development of a computational understanding of vision:

- (1) What affects the location and duration of fixation?
- (2) How is integration across fixations accomplished?

The two main sources of information about these questions are research in reading and picture viewing. The two tasks are quite different and results may not always be generalized from

one to the other (Rayner, 1978:641).

The great volume of research literature pertaining to eye movements prohibits the inclusion of a comprehensive review in this document. The interested reader could pursue the excellent review provided by Rayner (1978), or the series of three books by Senders, Fischer, and Monty (1978; Monty and Senders, 1976; Fischer, Monty and Senders, 1981).

#### 2.4.2. Non-Arbitrary Fixation Location and Duration

The subjective experience of saccadic eye movements is somewhat deceptive. We may believe that we are tracing a smooth path along a line while in fact our eyes execute a series of irregular shifts. We may not be aware of the exact locations upon which we fixate, only the objects which we detect. We may be aware of the gross influences on our fixation, such as sudden movement, but we are generally unacquainted with the subtle factors.

One common misconception of the role of eye movements in reading is that the ocular-motor system executes rhythmic or random movements across the line of text. This notion has been dispelled by research such as that of Just and Carpenter (1978) who showed that the semantic connections between sentences is a good predictor of the amount of time the agents of the sentences are fixated. Certain types of grammatical structure produce more frequent fixations (Klein and Kurkowski, 1974).

Buswell (1935) and Yarbus (1967) noted that when viewing pictures, subjects are more likely to fixate certain areas of the image. Mackworth and Morandi (1967) devised empirical methods to determine that the areas of pictures which subjects consider to be more informative are more likely to be fixated, both early and late in the viewing. Loftus and Mackworth (1978) have shown that objects which are unexpected in a scene are more likely to be fixated early, demonstrating the influence of cognition and expectation on eye movements, and showing the usefulness of the interpretation which takes place in the periphery.

Gould and Schaffer (1965) report the influences of task specifications on the duration and selection of eye movements during visual search (see also Gould, 1976). Loftus (1972) describes the relation between memory and fixation choices. Objects which were remembered in a scene were fixated by the third fixation 95% of the time.

We must accept the intricate influences of image properties, visual task, expectations, and progress of understanding in the determination of processing locations.

#### 2.4.3. Piecing Together Fixations

The studies of fixation determinants are quite descriptive, and do not generally suggest mechanisms. Research into the possible ways that information is pieced together from several fixations often includes processing models.

Parks (1965) moved a slit in a piece of cardboard over an image of a pattern and noticed that even though only a single element of the pattern could be seen at a time, that an understanding of the whole pattern emerged. This led him to conclude that individual glimpses can be assembled into a complete perception. Hochberg (1968) extended this experiment to include line drawings and introduced the idea of a "schematic map" which is used to synthesize successive glimpses, along with eye movement information.

Hochberg (1978; Hochberg and Brooks, 1978) emphasize the importance of underlying expectations in the development of a coherent structure of results from many fixations. Arguments in favour of this approach rather than translation of visual field on the basis of feedback from the eye movement system are made on the basis of the ease of understanding film clips which shift perspective and scale without predictability.

Reading studies indicate that effects of integration across saccades can be simulated without actual eye movements. Using an "on-line" eye movement monitoring and display generation mechanism, researchers are able to take advantage of the "ballistic" property of saccades, and by being able to determine fixation locations before the eye comes to rest, the displays may be altered during eye movements. Rayner (1975) showed that naming words on which fixation falls is easier



when the actual word appeared in the previous parafovea[10] (rather than a similar string of letters). Rayner, McConkie and Ehrlich (1978) demonstrated that the same effect can be obtained when the subject maintains a fixation and the displays are modified exactly as if a change in fixation were taking place. McConkie and Rayner have suggested that parafoveal and peripheral material are stored as an "integrative visual buffer", which is used as the basis for the incorporation of information from subsequent fixations.

#### 2.4.4. Models for Saccadic Control

Noton and Stark (1971a; 1971b; 1971c) proposed a representation for knowledge about objects which consisted of rings of alternating features, and motor traces to permit moving the eyes to the next feature. Eye movements were considered to be the following of "scan paths", as provided by the representation. These repetitive sequences of saccades were shown to develop early in the viewing of an image, and to recur in subsequent perceptual tasks with the same image. As a theory of eye movement control this notion of "scan-paths" has two weaknesses:

- (1) No central role is provided for the use of peripheral vision.

---

[10] The area just outside of the foveal center is often referred to as the parafovea.

- (2) The representation is strictly view-oriented. It does not consider that if the locations to which fixation is to be drawn rotate with an object, then the motor traces necessary to effect the saccades will have to change.

Farley (1976) presents the description of a computer implementation of an eye movement system. The general form of the model is derived from Noton and Stark's ideas, but it includes a hierarchical organization of the objects being viewed. The strategies for effecting eye movements consist essentially of breadth-first and depth-first search of the space defined by the object models, and the lines given in the model are followed to look for expected vertices. The basic directive for changing processing location in Farley's system is suggested by the expected directions of the corners of the objects. This is similar to the concept employed by Shirai (1975) in his knowledge-based line finding program. Shirai's knowledge of scene domain corners was encoded as corresponding image domain vertex information.

Didday and Arbib (1975) also report an eye movement computer implementation which is based upon the Noton and Stark model. They conclude that eye movements are based on properties of the image (features) and not on motor traces. This suggestion requires a more complete representation of the scene models than is provided by Noton and Stark. In addition, peripheral vision would be required to form hypotheses about the location of features which need more careful

examination. This is the basis of a model proposed by Walker-Smith, Gale, and Findlay (1977), using studies of eye movement paths over images of faces as supportive evidence.

Parker (1978) also argues for the importance of peripheral vision in the control of eye movement behaviour. Parker's model is based on Neisser's (1976) perception cycle: expectations about the type of information that will be provided for an object are encoded as sequences of features to be fixated. The "exploration" phase of the cycle involves the detection of these sequences.

The conclusion drawn on the basis of psychological experimentation that several diverse influences act towards the determination of fixation location is consistent with the tendency for computer implementations to emphasize one isolated factor. Roy and Sutro (1982) describe a system which selects a sequence of fixation locations in an image on the basis of the expected amount of edge. A rough measurement is made at each location, and the processor follows an ordered list of the expected amount of edge at each location[11]. Funt (1976) developed a system to analyze the stability and structure of a group of imaged objects. The operations included the movement of a graded resolution retina across the image in response to

---

[11] The paper also includes suggestions as to how the Nott and Stark model might be extended into three dimensions, following on some of the work of Marr and Nishihara (1976) and Oshima and Shirai (1981).

the requirements for problem relevant information such as the locations of points of contact between objects.

Pylyshyn, Elcock, Marmor and Sander (1978a; 1978b) implemented a perceptual-motor system which includes, as one component, the application of an area of high resolution availability across drawings of geometric figures. Of particular interest in this operation is the idea that just because objects or features fall within the fovea, does not mean that they are automatically fully processed. An attentional mechanism must be applied, and features collected to enable a matching with nodes of a memory network. This idea is very similar to that expressed by Kahneman and Treisman (1982), in their "object file" model for visual attention.

From a computational point of view, the basic requirements of a system which can intelligently select processing locations are:

- (1) The ability to exploit the result of more extensive, lower resolution peripheral analysis.
- (2) The capability to direct processing to area on the basis of expectations or conflict within an ongoing interpretation process.
- (3) The recognition of areas of image detail which are intrinsically more likely to provide important information.

Hochberg and Brooks (1978:312) have proposed that a dual system could best accomplish the direction of eye movements:

".. a fast component which brings the eye to those peripherally visible regions that promise to be informative or to act as landmarks, and a more sustained component that directs the eye to obtain more detailed information about the main features that have already been located."

### 3. Research Overview

The research presented in this thesis has two major objectives.

(1) To develop formal representation methods for the definition of simplified problem domains, and to devise generalized operations which can utilize these representations to effect interpretation of images representative of the problem domain.

(2) To implement these strategies within the framework of a consistent and realistic model of visual perception.

This chapter provides an outline of the methods developed without reference to the computer implementation or the specific problem domain used. As a result, the outline is sketchy and incomplete. It should only be viewed as providing an overall structure for the detailed accounts with reference to the implementation found in chapter four.

#### 3.1. A Model for Perception

Within this model of perception, component hierarchy information is made explicit in a knowledge structure, with non-decomposed elements represented in terms of image features available at the finest level of resolution. These image constructions are prototypical views of objects which are flexible enough to cover a wide range of actual viewing angles. Other representations of object knowledge might coexist with this view-based structure, but the capabilities of this

structure are considerable, particularly in the understanding of convention-based line drawings. (see Section 4.2).

The representation of objects directly in terms of image features provides the possibility for an interpretation labeling approach in which features are assigned lists of object models that use the features in their descriptions. This cuing structure is extended to the more complex objects and thereby develops a recursive cuing mechanism, which encodes potential relations among objects and their depiction in images. The result is structures which provide the capability for both top-down or bottom-up analysis (or both), without making a commitment to any particular strategy. The resultant descriptive structures might equally well be used in the generation of drawings of the objects.

This structure accounts for features at a fine resolution level. In a line drawing domain, these features would be the lines themselves, and their properties of length, and curvature. In addition, other, less detailed structural descriptions of objects are maintained based on the types of features available at a coarse resolution level (such as blobs). In some cases there are relations between concepts of the fine resolution structure and of the coarse resolution structure. These relations coincide with a specialization hierarchy and thereby form a natural part of the concepts of objects.

This adds a new dimension to the cuing structure. Not only does the component hierarchy give the structure for a recursive cuing mechanism, but the specialization hierarchy provides a direct relation to the hierarchical relations within the image, as shown in figure 3.1.1.

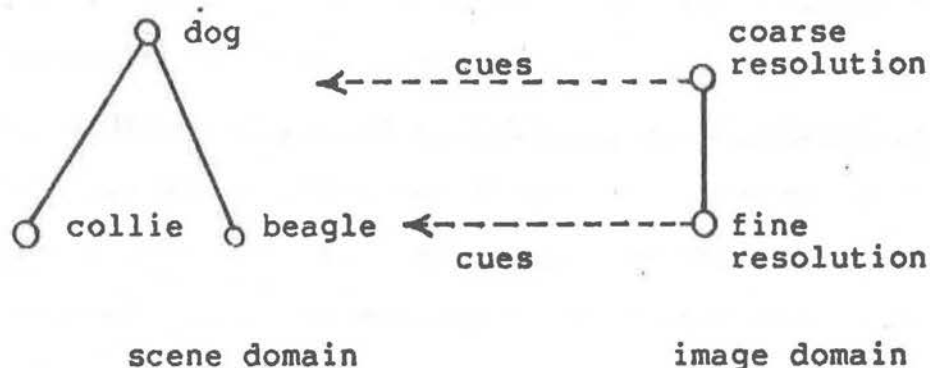


Figure 3.1.1. Different levels in the image hierarchy cuing at different levels in the specialization hierarchy.

Central to the operation of this model of perception is the idea that fine detail features are available in only a small portion of the image at a time, and that this area of availability coincides with a larger area of availability of coarse level features. Within a single such fixation, features are collected from their appropriate areas. Each feature has associated with it a list of the possible object models to which it may belong.

The first interpretation processes act towards the reduction of these model possibilities by the formation of groupings of features interrelated by the image hierarchy structure. These groupings allow the reduction of model



possibilities through the enforcement of the requirements for consistent interpretations within groups (see section 4.4). These reduction operations are assumed to be parallel within groupings, and rely on set intersection as their primary operator[12].

The remaining possibilities must be examined in more detail with consideration of the specific relations required among features to verify models. Whenever these requirements are similar across a class of objects, model descriptions are compressed into generalized forms, thereby creating a criterion for a second type of specialization hierarchy.

Any object which is found to be adequately supported in the image is asserted, and then can act as a cue for the more complex structures of which it may be a component.

The results which are possible on the basis of a single fixation location may be quite limited, and so other areas of the image must be processed. Intelligent selection of fixation locations will expedite the interpretation. This selection relies on the correspondence between foveal-based detailed results, and the results obtained in the coarse level periphery as follows:

---

[12] See (Fahlman, 1979) for a discussion of the use of set intersection as a unit operator in parallel systems.

- (1) The coarse level results act as a framework for the integration of the successive high detail results.
- (2) The locations to process are selected so as to maximize the propagation of detailed interpretation into the periphery.

In addition, there is provision for consideration of both the structure of the image and the task at hand in the selection of new processing locations.

After a number of fixations[13], the entire image is understood in terms of the coarse level models, and an understanding on the basis of the fine level models, obtained locally at the fixation centers, has been adequately propagated to the coarse level interpretation such that a fine level understanding of the entire scene is possible without actually having scrutinized each location with the fovea.

---

[13] The number depends largely on the settings for the radii of the fovea and periphery.

### 3.2. Declarative Schemata

The characteristics and advantages of a schemata-based approach in the application of model knowledge in computer vision has been outlined in section 2.2. There are a number of potential advantages to the development of mechanisms which can encode such knowledge in a declarative way.

(1) A declarative description of the problem domain, without reference to the means of interpretation provides an explicit statement of the system's capabilities and requirements.

(2) Declarative domain representations may be used in conjunction with simplified control structures in order to verify the model knowledge before subjecting it to the typically more complex control required to use the model knowledge in vision. The recent interest in logic-based programming systems such as PROLOG has provided adequate tools for the accomplishment of this testing.

(3) Separation of problem domain knowledge from the interpretation methods permits simplified expansion or modification of the models, or even transfer to another domain which can be represented within the syntax of the declarative schemata. This separation also facilitates experimentation with a variety of interpretation control methods.

(4) Procedures may be developed to analyze declarative schemata towards the end of automatic generation of a cuing struc-

ture. Explicit statements about relationship among domain elements and their depiction in images means that the knowledge may be inverted to obtain a cuing structure.

(5) In laying bare the structure of the problem domain, provision is made for the analysis of objects in terms of the relative importance of their attributes. Of the many attribute values which may be developed for an object, some are criterial to its playing a part in the support of a more complex structure, while some may be relatively unimportant. Thus the interpretation processes may be tuned to first deal with those features which are important to recognition. The structural correspondences among representations based on features at different resolution levels may also be made explicit and available to analysis.

A declarative schemata system has been developed which is, in the strictest sense, a grammar of the problem domain and its depiction in the image. The terminal symbols of the grammatical description are the primitive elements of the image. The productions of the grammar will be referred to here as descriptions. This term is more appropriate because of their truly descriptive nature, and in order to avoid the connotation of a "production system" (Newell, 1973) which would be inappropriate because the descriptions involve no provision for interpretive action.

Phrase structure grammars are normally used in the representation of classes of objects which are essentially one-dimensional[14]. The implicit concatenation of vocabulary symbols in a production or sentential form is representative of adjacency in the input. For a class of two-dimensional image representations, or for a class of three dimensional scene objects, the notion of adjacency is more complex, and must be made explicit.

A system of assigning attributes to non-terminals has also been incorporated as a means of specifying the semantics of the domain. Values for the attributes are passed on and developed through a mechanism reminiscent of "attribute grammars" described by Knuth (1968) and Marcotty Ledgard and Bochmann (1976).

A simple example of a description for an isosceles triangle will serve as a good demonstration of the way these extensions have been introduced.

Each description has the underlying form of a phrase structure grammar production

$$X \rightarrow A B C$$

---

[14] There are techniques which can reduce two-dimensional image elements into one dimension, such as tracing around the perimeter and recording the changes in direction in a list, which is then treated as input (see Ledley, 1964).

where X is being defined in terms of the more basic elements A, B, and C.

```
triangle --> {line line line}
```

In order to expedite the assignment of relations among the basic elements, each is given a label by which it may be referred. This label also establishes its uniqueness within the description.

```
triangle --> (($1 line)($2 line) ($3 line))
```

relations indicate the elements over which they apply:

```
triangle --> (($1 line)($2 line)($3 line))
              (($4 connect ($1 $2))
               ($5 connect ($2 $3))
               ($6 connect ($3 $1))
               ($7 equal-length ($1 $2)))
```

Of course, this does not specify an isosceles triangle in two respects: the lines may not be straight, and they may overlap. The use of the attributes of the image features, as well as attributes for the relations can provide for the specification of these constraints.

```
triangle --> (($1 line (curve 0))
              ($2 line (curve 0))
              ($3 line (curve 0)))
              (($4 connect ($1 $2) (angle (1 179)))
               ($5 connect ($2 $3) (angle (1 89)))
               ($6 connect ($3 $1) (angle (1 89)))
               ($7 equal-length ($1 $2)))
```

Some higher level description may use this "triangle" as one of its basic elements, requiring conditions to be placed upon applicability through reference to its attributes, so it will be part of this description to specify what attributes are available and how they might be developed out of the attributes of the elements and relations composing the "triangle". This specification is easily added:

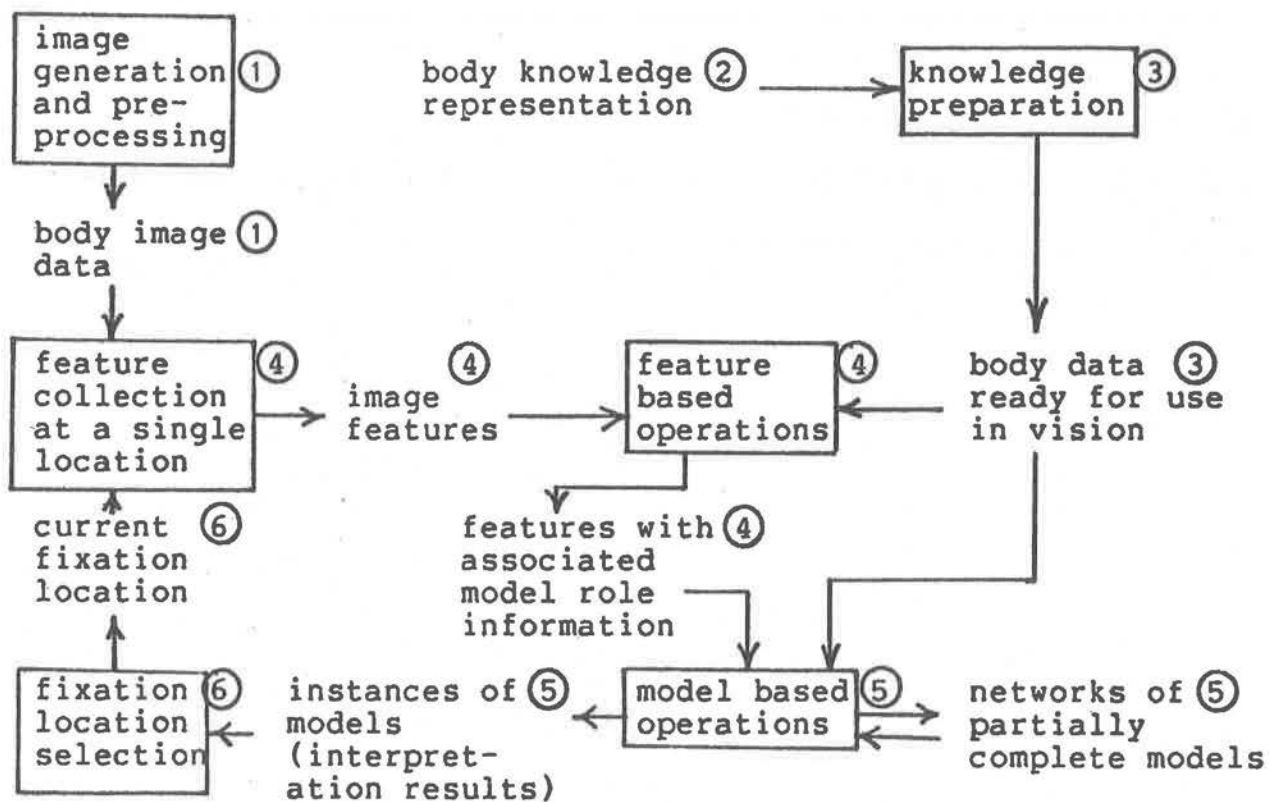
```
triangle --> (($1 line (curve 0))
              ($2 line (curve 0))
              ($3 line (curve 0)))
              (($4 connect ($1 $2) (angle (1 179)))
              ($5 connect ($2 $3) (angle (1 89)))
              ($6 connect ($3 $1) (angle (1 89)))
              ($7 equal-length ($1 $2)))
              ((base-length <- (length $3))
              (orientation <- (slope $3))
              (height <- (times (arctan (angle $5))
                                (divide (length $3) 2))))
```

With very few further modifications, this form is capable of encoding the entire test problem domain, without requiring that the specifications of relations or attribute generation methods become much more complicated.

#### 4. A Computer Vision Implementation

This chapter provides a detailed account of the computational vision system designed to implement and experiment with the ideas given in the previous chapter. With the help of examples, the ideas are elaborated considerably, and several discussions of related research issues are included.

Below is shown an overview diagram of the computer implementation. Processes are enclosed in squares while data structures are not enclosed. Each structure or process has beside it a number which indicates the number of the section of this chapter which deals with it. Examples of the system in operation are provided in chapter five, and as well in Appendix D.





#### 4.1. Problem Domain and Image Generation

The class of images interpreted by the system is line drawings of human-like body forms. The drawings are derived from those used by Eshkol and Wachmann (1958) to illustrate their dance notation, and as would be expected, they are very expressive of human body positions. Each body drawing is represented by 16 or 18 closed-line image constructions, depending on the perspective view. Some examples are provided in figure 4.1.1.

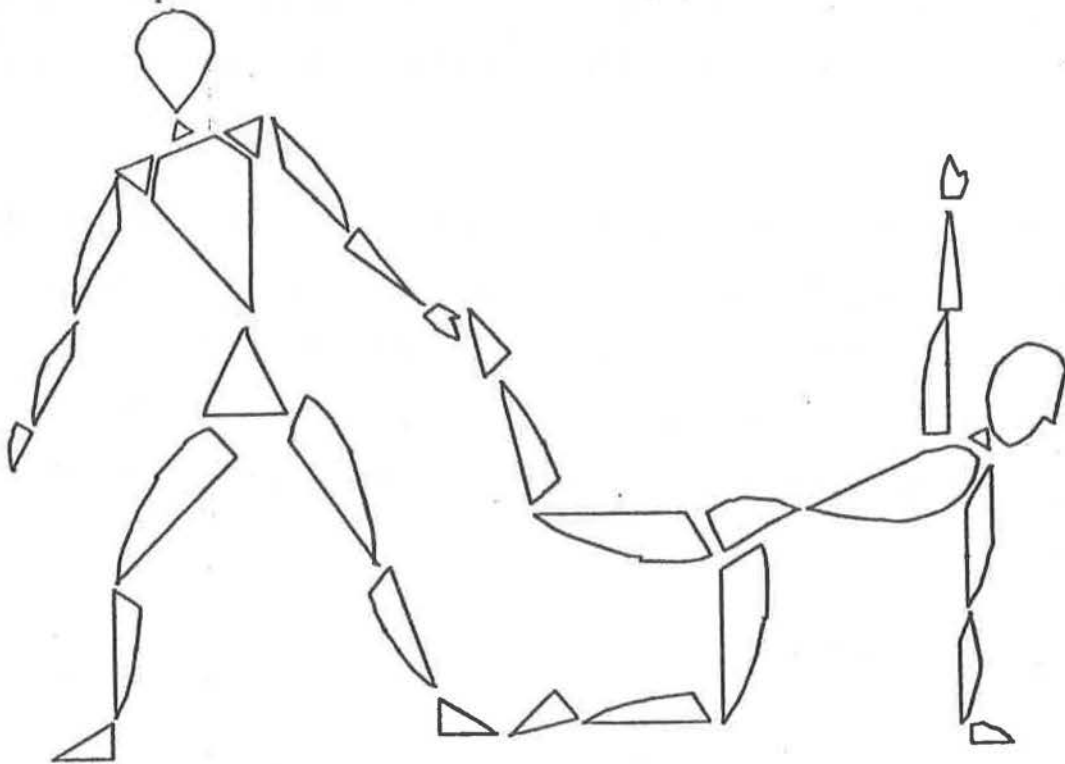


Figure 4.1.1. Some examples of body form drawings.

The drawings do not depict either foreshortening or occlusion, as the processes necessary to deal with such aspects of a scene require more detailed information about surface orientation and range, which are not easily deduced from the simplified image forms. Furthermore, these issues are not central to the goals of the research.

There is a very large number of drawings which fall within the problem domain. Requiring 45 degrees to distinguish between angular positions of body parts, ignoring the fact that several image constructions may depict a single view of a body part, and ignoring overall orientation and scale, there are still, by conservative estimate, about 100 million different drawings in the class.

The images are constructed through the use of a menu-driven program[15], which permits positioning, scaling and rotating of body part depictions selected from an inventory of image representations as shown in figure 4.1.2. The preliminary result is a list of straight line segment end-points on a 1024x1024 grid.

---

[15] This portion of the system runs on a PDP-11/34 using a VT-11 graphics generator, and a VR17 display tube. The interaction is accomplished with a light-pen.

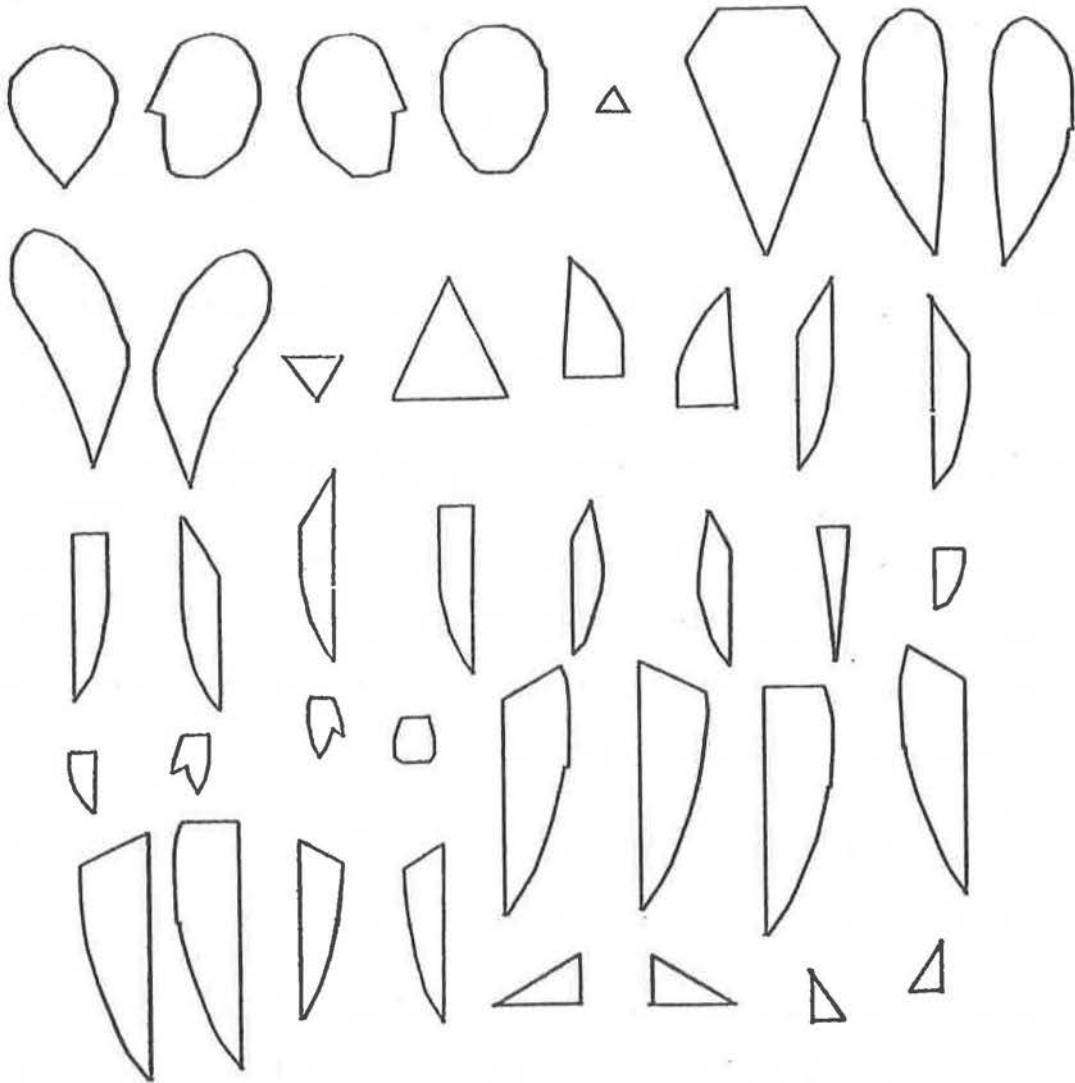


Figure 4.1.2. Complete collection of body part depictions.

Next, a series of programs operates on this list of endpoints in order to produce a data-base of features which will be made available to subsequent interpretation systems. The features to be returned are line segments and blobs with attributes assigned as shown in figure 4.1.3.

<u>Feature type</u>	<u>Attribute</u>
line	end-points curvature[16]
blob	center of gravity end-points of long axis end-points of short axis

Figure 4.1.3. Image features and their attributes.

The Psychology literature supports the notion of curvature as a feature involved in human perception (Riggs, 1973).

In addition to these features, some image hierarchy information is computed. For each line segment, a list is returned of the blobs with which it overlaps in space, along with a measure of the amount of overlap.

This information is extracted by a series of FORTRAN programs which performs the following steps:

- (1) Trace connected chains of straight lines looking for points of departure in curvature, and mark the segment boundaries, measuring the segment's curvature.
- (2) Develop a 128x128 representation of the image, with each pixel encoding the length of line segments found in an 8x8 window in the 1024x1024 image. Thresholding the

---

[16] The angle formed at the intersection of the tangents at the end-points.

value of "on" pixels, blobs are determined by expanding outward from the unfilled centers.

- (3) Determine the blob attributes. The axes are computed computed by averaging the orientations of the pixels nearest and most distant from the center of gravity.
- (4) Calculate the segment and blob overlap for the image hierarchy information.

Figure 4.1.4 and 4.1.5 show an example line drawing at different stages of processing.

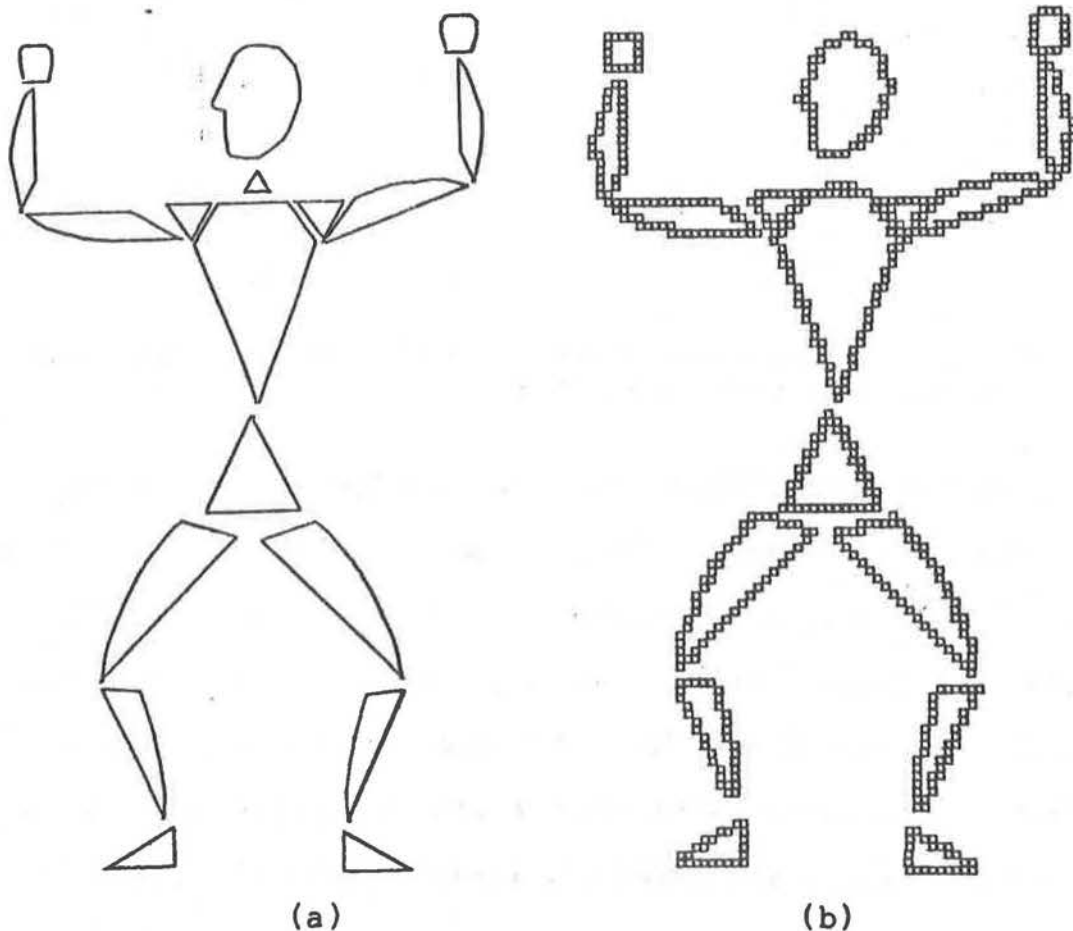


Figure 4.1.4. Line drawing at (a) 1024x1024 initial line drawing. (b) 128x128 averaged image.

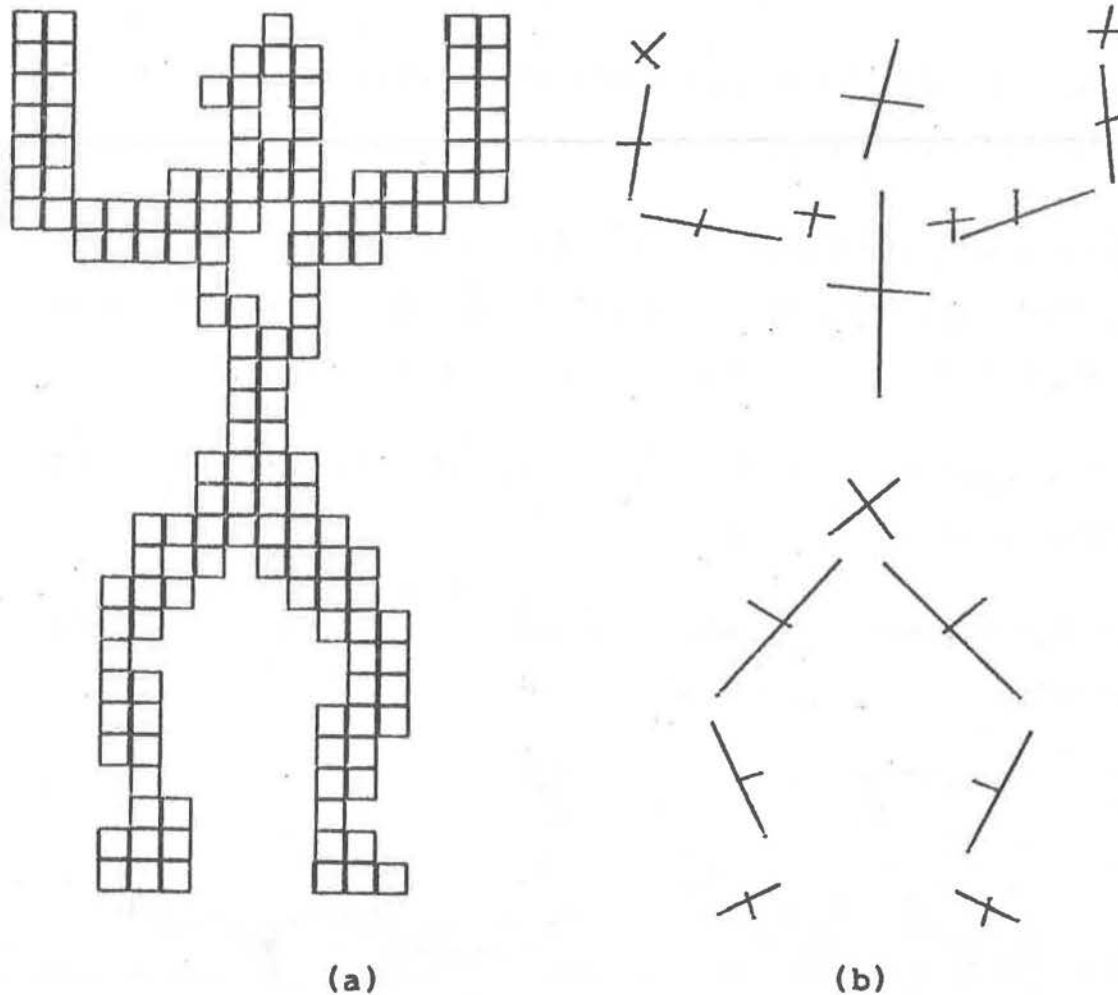


Figure 4.1.5. Line drawing at (a) 32x32 averaged image (b) the axes of each detected blob.

It must be understood that the results of this initial processing are intended only as a base of features, representative of the type of information which might be available to a vision system. For this reason, the inner workings of the programs described here have not been elaborated. Chapter 5 provides a complete working example which includes a description of the information made available from the image as basic features.

#### 4.2. Knowledge Representation

Body form knowledge is represented in a declarative schemata [17] system, consisting of three different types of descriptions:

- (1) those which develop image constructions from the basic features.
- (2) those which map between image constructions and basic scene objects.
- (3) those which develop complex scene objects.

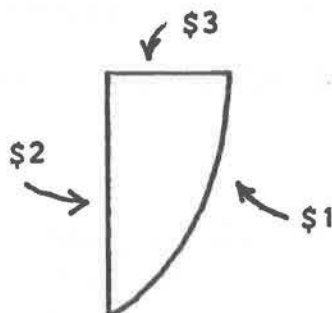
The following characterizes each type in turn with the help of examples from the body form knowledge.

Each of the valid structures in the line drawings is indicative of a particular perspective of a body part. The task of the image construction descriptions is to indicate how these views may be composed of image features. The descriptions are intended as prototypes, or ideal view representations, but as will be seen later in section 4.3, there is actually a wide variation which is acceptable, determined by the setting of global parameters. Consider the example of "line-hand-1" shown in figure 4.2.1. This view is the starting, or base view of the left hand (see figure 4.2.3). As

---

[17] The basic concepts behind the declarative schemata system were first described in Browse (1980).

indicated, this is a component description which depicts the hand view in terms of three lines, one of which is curved.



```
(line-hand-1
 (component
  (($1 line (curve 53))
   ($2 line (curve 0))
   ($3 line (curve 0)))
  (($4 connect ($1 $2) (angle 134) (ratio 108))
   ($5 connect ($2 $3) (angle 90) (ratio 225))
   ($6 connect ($3 $1) (angle 92) (ratio 41)))
  ((size <- (length1 $1))
   (a2d <- (slope (location $5) (location $6)))
   (proximal-end <- (midpoint $3))
   (location <- (middle (location $4) (midpoint $3)))
   (distal-end <- (location $4))))))
```

Figure 4.2.1. Component description of a view of a hand.

Connections are always such that the angle between lines is a deflection to the right of magnitude between 0 and 180 degrees, thereby eliminating some ambiguity. The angle at the connection is a local angle, based on the end-point to end-point angle, and the curvature of the lines (see Appendix A). The prototypical ratio of the lengths of the lines is also provided as a constraint on the attributes of the connections[18].

[18] Though not shown in this example, a connection may also take on a "ctype" attribute, which serves to point out the infrequent occurrences of concave line connections. Examples of this construct may be seen in the description of "line-head-4" in Appendix B.



In this example, as with all line construction descriptions, the attribute "a2d" indicates the two-dimensional orientation of the view as determined by the orientation of one of its composing lines. The attributes "size" and "proximal-end" are also important in later uses of this description.

Figure 4.2.2 gives an example of a description which maps an image construction into a basic scene element. There is only one element in the description, that is "line-hand-1", and there are no required relations. The description simply transfers attribute values from the image domain into the scene domain. There is a similar description for each of the topologically different views of the hand.

```
(hand
  (image
    (($1 line-hand-1)) nil
    ((posture <- open)
     (location <- (location $1))
     (proximal-end <- (proximal-end $1))
     (distal-end <- (distal-end $1))
     (size <- (size $1)) )
    ((side <- left)
     (a3d <- (list 0 0 (neg (a2d $1)))))
    ((side <- right)
     (a3d <- (list 0 180 (a2d $1)))))
```

Figure 4.2.2. Image to scene mapping description for a view of the hand.

In the attribute portion of the description, there is first of all a list of attributes which are to be passed directly. There are also additional sets of attributes which are referred to as elaborations. Each elaboration consists of

a cluster of attribute alternatives to be assigned to the scene element. In this example, if the hand turns out to be a right hand, the two-dimensional orientation of the "line-hand-1" object will influence the three-dimensional orientation ("a3d") of the hand in a different way than if it is the left hand.

The convention used to denote the three-dimensional orientation of a body part indicates the amount of orientation variation there is from the body part's starting position. Figure 4.2.3 shows the body in the starting position.

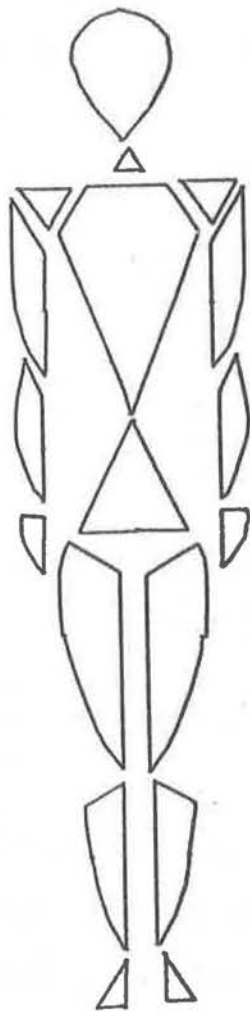


Figure 4.2.3. Body form in starting position.

The orientation is given as a triple indicating the left-hand rotation in the range  $[0, \pi)$  about each of the body-centered cartesian axes (see figure 4.2.4).

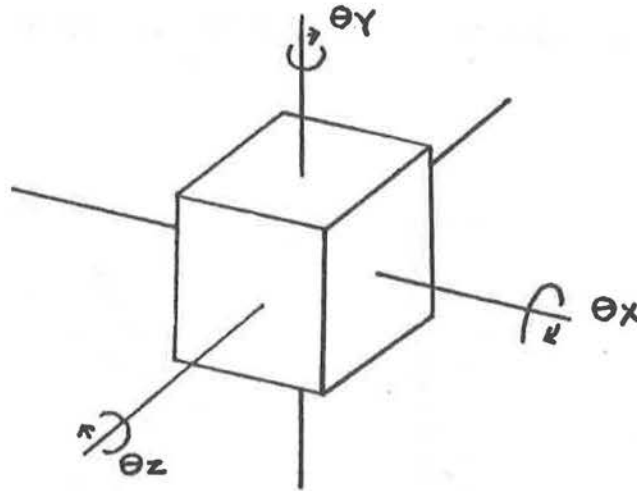


Figure 4.2.4 Body part orientation relative to its rest position described as a triple  $(\theta_x, \theta_y, \theta_z)$ .

Providing that the three component angles are always considered in the same order, each orientation triple is unique in its representation of the orientation of the body part.

There are instances for which the three-dimensional orientation will be the same regardless of whether the scene object turns out to be right or left, and there are instances for which several alternative three-dimensional orientations are possible for each side. These representations are sensitive to the depiction possibilities allowed for an image structure. Consider figure 4.2.5. If this image construction

is allowed to depict an upper leg in such a way that the curved bulge may be either the back of a leg or the outside of the leg, then each of these orientation elaborations must be included in the mapping to the scene domain. If the depiction were extended so that the bulge could represent the triceps, then an additional elaboration would be necessary.

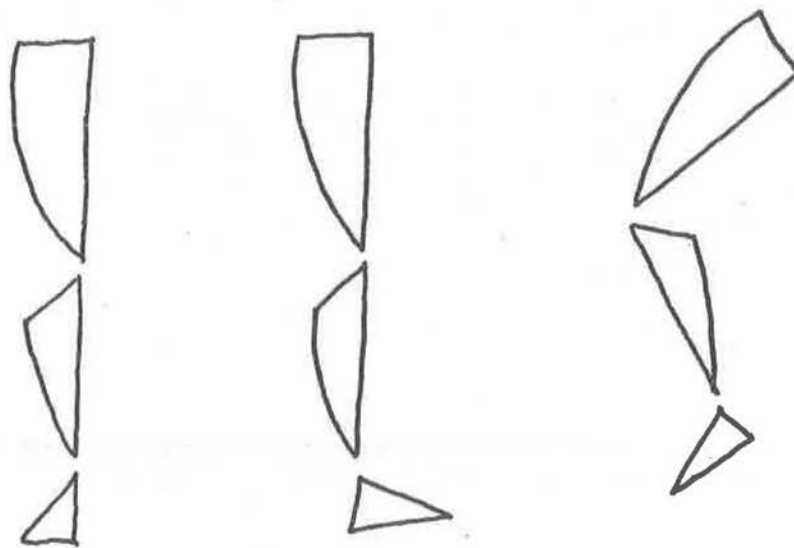


Figure 4.2.5. A single depiction of an upper-leg used to represent three different orientations.

The philosophy behind the use of this type of representation is based on a concept of objects as collections of attributes. Each of the potential mappings from the image constructions to the scene objects, differs only in the way it develops some of the attributes. For the example shown in figure 4.2.5, the attribute values of "size" and "distal-end" will be the same for each of the six possible mappings (three

per side). Only "a3d" and "side" vary across the different meanings of the view. It does not seem reasonable to specify each mapping separately, but rather, because of the similarity, it is best to develop a generic mapping from which several specialized versions, called "elaborations" may be obtained.

The third description type composes more complex body parts out of the basic ones. The structure of the description is almost identical to the image construction descriptions except that the components are now other body parts instead of lines, and the "connect" relation has been replaced by a "near" relation. The "near" relation also specifies two elements for which it must hold, and specifies point attributes of those elements, which must be within a proximity of one another as determined according to the overall size of the parts involved. For any such "near" relation, the orientation of the more distal part relative to the proximal may easily be computed. This orientation triple is broken down into three separate attributes of the relation: "angle-x", "angle-y", and "angle-z". Ranges in which these values must fall are provided in the relation specification, and indicate the range of motion capabilities at the joints of the body[19].

---

[19] The rest (or starting) position chosen is identical to that used by Eshkol and Wachmann (1958) in their dance notation. It is also identical to the rest position used by the American Academy of Orthopedic Surgeons in their handbook "Joint Motion: Method of Measuring and Recording" (1965). So the angles of joint movement provided in that handbook could be inserted directly into the body model.

Figure 4.2.6 provides, as an example, the "left-arm" description[20].

```
(left-arm
  (component
    (($1 hand (side left))
     ($2 lower-arm (side left))
     ($3 upper-arm (side left)))
    (($4 near ($1 $2 proximal-end distal-end)
      (angle-x (-30 20))
      (angle-y (0 0))
      (angle-z (-90 90))
      (ratio 43))
     ($5 near ($2 $3 proximal-end distal-end)
      (angle-x (-150 150))
      (angle-y (-90 90))
      (angle-z (-150 150))
      (ratio 79)))
    ((a3d <- (a3d $3))
     (proximal-end <- (proximal-end $3))
     (size <- (times 2.1 (size $3)))
     (location <- (location $5))
     (elbow-location <- (distal-end $3))
     (elbow-posture <- (diff (caddr (a3d $3))
                           (caddr (a3d $2)))))) ))
```

Figure 4.2.6. Left-arm schema description.

Other descriptions, of course, develop even more complex body parts, such as "lower-body" in terms of "leg" and "hips", and finally the distinguished symbol of the grammar "body" is defined. The entire body knowledge grammar is provided in Appendix B.

---

[20] The hand was described as a single object with a side attribute of either left or right. The arms and legs, however, have separate descriptions for their sides. This was an arbitrary and intentional decision made so that investigation could be made into the use of both modes. The final model was left with one of each.

The complete grammar for the body knowledge is made up of two layers[21]. The examples used in this section are all from the fine layer, which gives a detailed account of the body on the basis of line features. The coarse layer provides a rough account of the body on the basis of blob features. Each uses descriptions of the same form, and really is a separate grammar in its own right. Most non-terminals do, however, have counterparts in the other layer, specified explicitly through the use of generalization/specialization hierarchy linkage. For example, an object "limb" in the coarse layer grammar, has links to both "arm" and "leg" in the fine layer, while "extremity" links to both "hand" and "foot".

Within each layer of the grammar, then, is specified a component hierarchy of body parts. Across layers is specified the generalization/specialization hierarchy. Figures 4.2.7 to 4.2.9 show the complete structure involved in these hierarchies.

---

[21] The term "layer" is chosen here rather than "level" to avoid any confusion with the notion of a two-level grammar (van Wijngaarden, et al 1969), which is an entirely different concept.

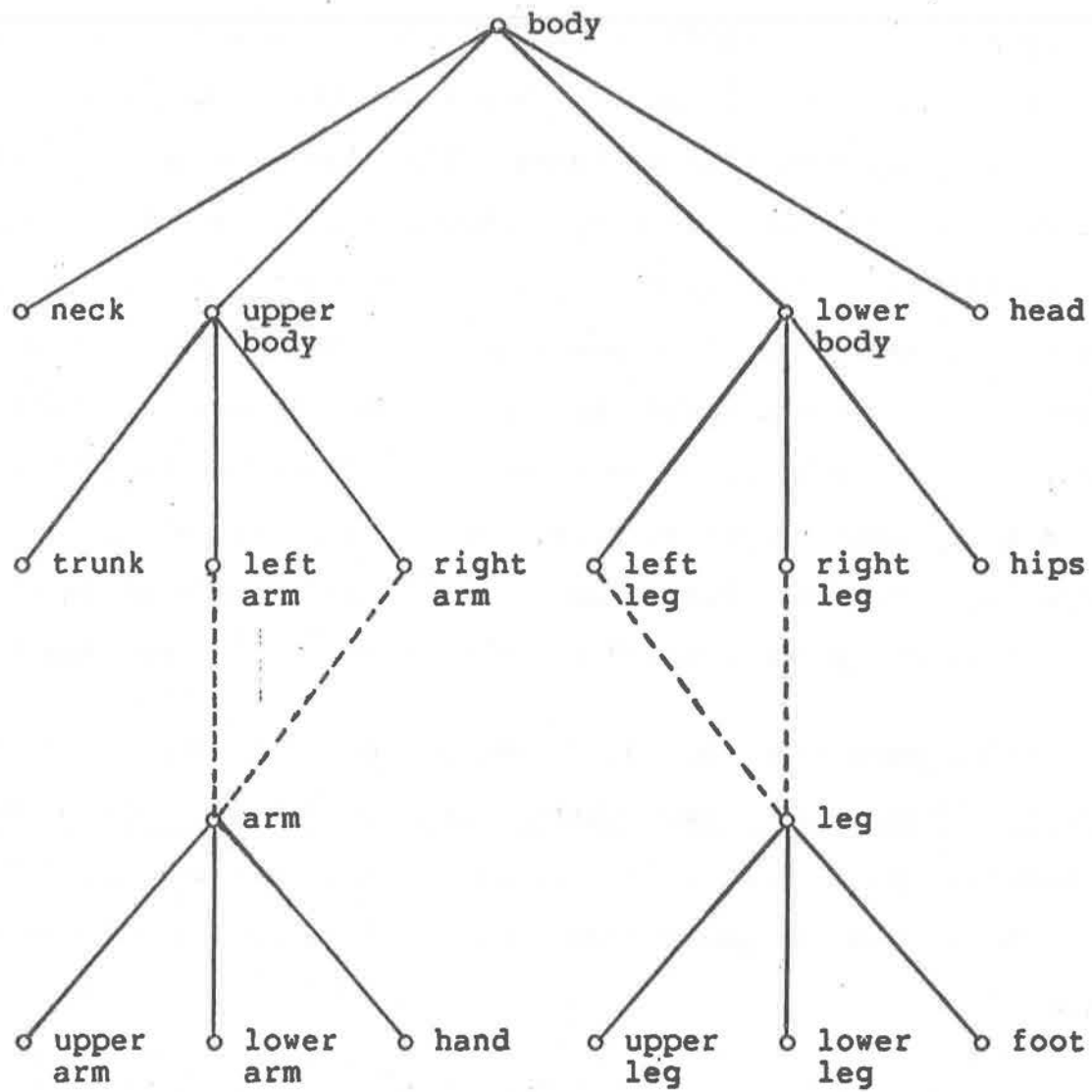


Figure 4.2.7. The component hierarchy for the fine layer of the body form representation.



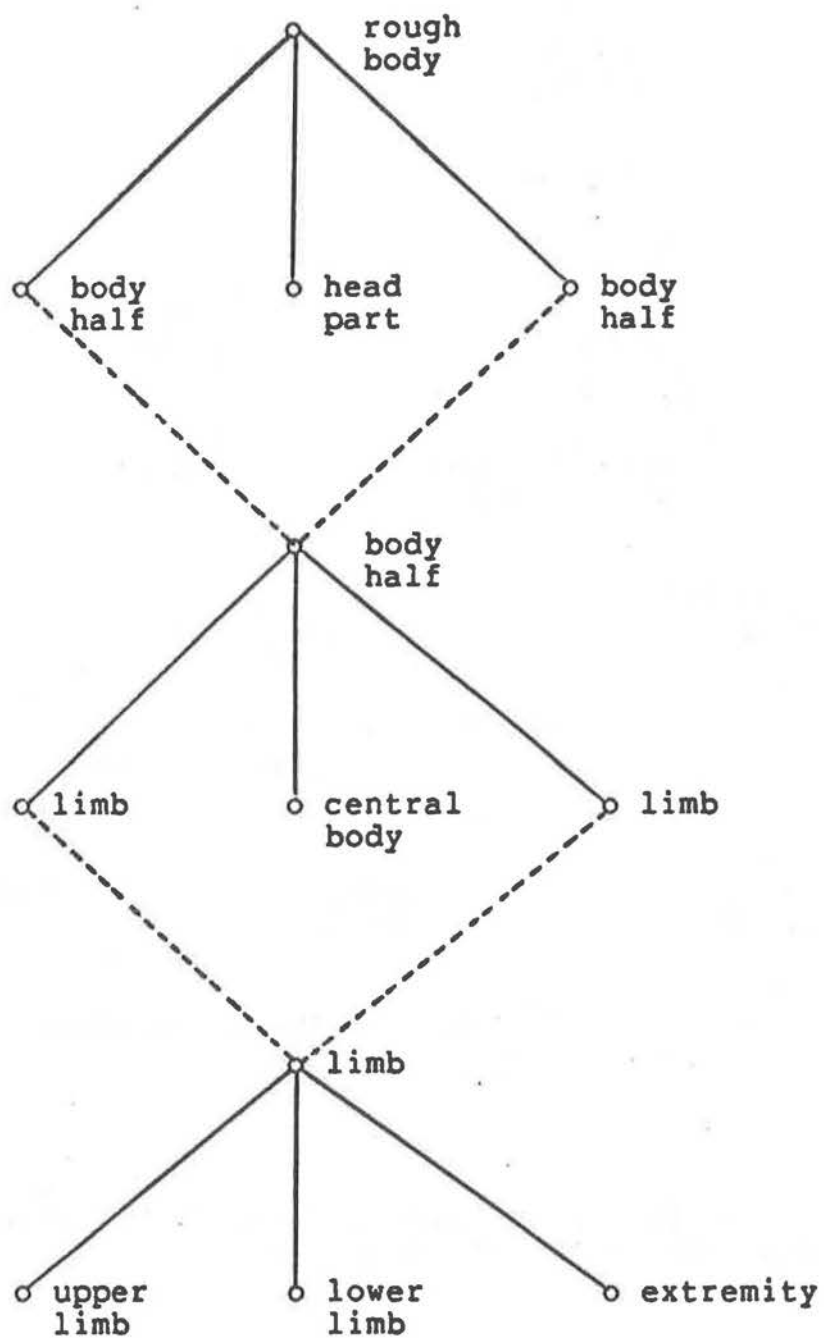


Figure 2.4.8. The component hierarchy for the coarse layer of the body form knowledge.

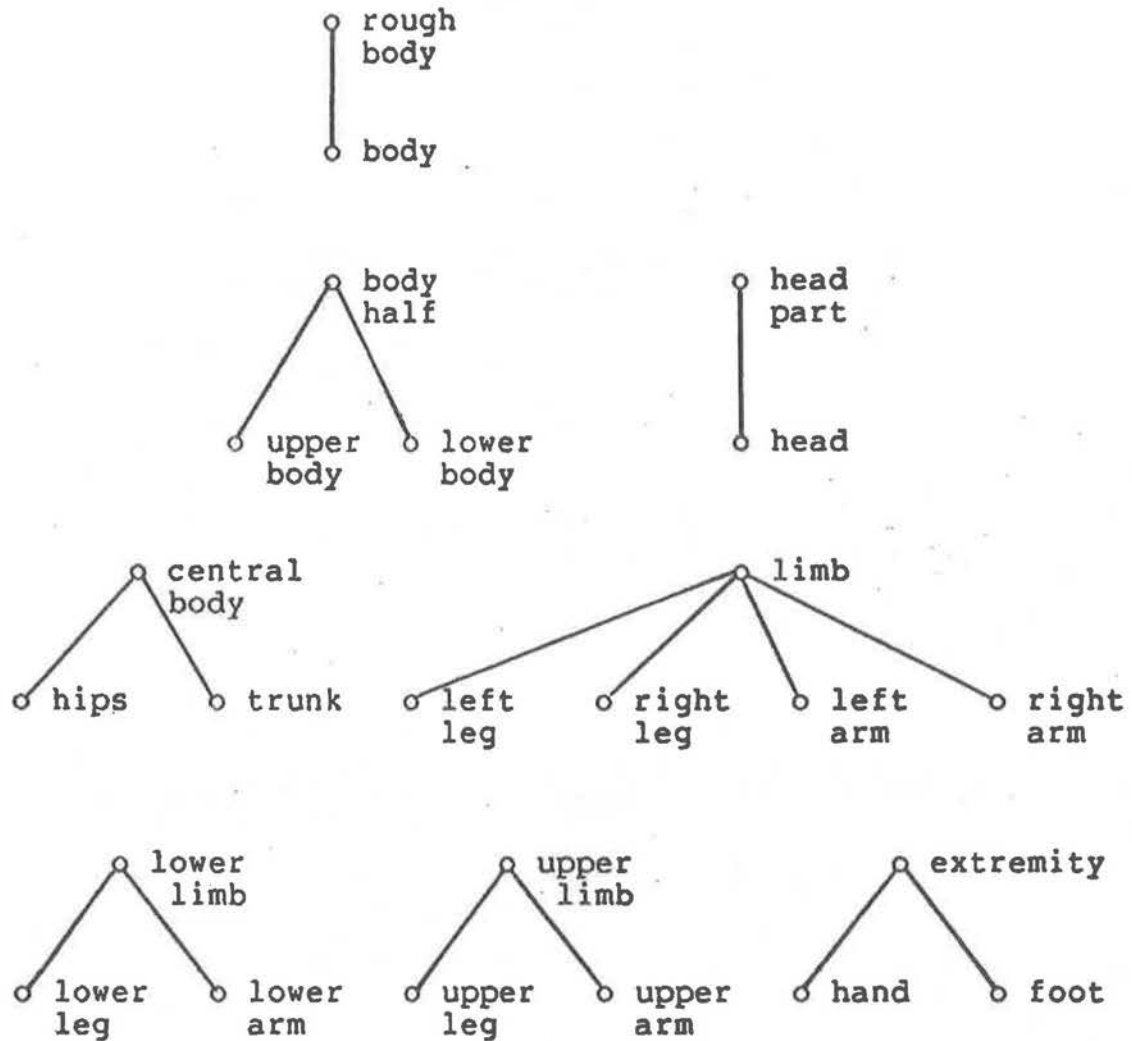


Figure 4.2.9. The specialization/generalization hierarchy for the body form representation.

#### 4.2.1. Adequacy of Representation

As a means of testing the adequacy of the body form knowledge, a portion has been translated into the programming system PROLOG[22], and used to "prove" body parts in a data base of assertions about lines. Each object in the PROLOG system is identified by a "tag" which is made up of its name followed by an integer (for example, "line7"). Attributes are then axioms asserted which involve the tag. For example, the assertion of "line1" is:

```
point(1,line1,432,876)
point(2,line1,383,950)
length(line1,88)
curve(line1,14)
```

The processing first detects all connections in the image and asserts them. Next, the existence of a complex scene object is entered as a goal:

```
<- leg(*tag,*side)
```

Through a straightforward process, it was possible to devise theorems for the body parts, based on the grammatical descriptions[23]. All that was necessary to support the use of these theorems, was to encode the relations which are specified in

---

[22] This portion of the system was implemented on an Amdahl V8, running MTS operating system.

[23] Appendix C contains printouts of some of the PROLOG theorems for the body parts. They may be compared to the grammatical descriptions found in Appendix B.

the descriptions (for example, "near", "connect").

This PROLOG system was useful in two respects:

- (1) It was determined that the body form representation is adequate to permit interpretation.
- (2) The system could be used as a means of debugging the knowledge representation, without the complication of interpretation processes being involved.

One of the basic goals of this research is to experiment in methods of controlling and directing interpretation of images using the declarative structures which define the problem domain. The following sections will outline these procedures. Because of the difficulty of implementing local consistency methods and because of its inherent commitment to backtrack search, PROLOG was not used as the language of implementation.

### 4.3. Preparation for Interpretation

The idea behind the "recursive cuing mechanism" of schemata-based vision systems is that each object (or perhaps relation) in the problem domain may act both as a feature and as a model. This way of viewing the structure of a domain seems applicable to the body knowledge representation because there are explicit ties among the elements throughout the descriptions.

A closer examination, however, reveals a problem: The fringe nodes of the component hierarchy, the features such as a "line", cue a large number of models - in fact every image construction in the fine layer. Similarly for connections between lines.

The solution to this problem is found through the use of a technique for incorporating the attribute structure of objects into the mechanism for maintaining model possibilities. We shall refer to this technique as set labelling.

The idea behind "set labelling" can be easily expressed in the context of a simplified problem domain. Consider the domain of vehicles (as used in Havens, 1978). Assume four vehicles; sports car, bicycle, cart, and truck, each of which has a description based on its components. Each will have "wheel" as a component, but each will express different attribute values which must hold for the "wheel":

```

truck -->   . . .
              . . .
              ($1 wheel (type solid) (width thick))
              . . .
              . . .

bicycle --> . . .
              . . .
              ($1 wheel (type spoked) (width thin))
              . . .
              . . .

```

Set labelling provides a cuing structure which is contingent on attribute values of features. A simple structure is set up, as shown in figure 4.3.1.

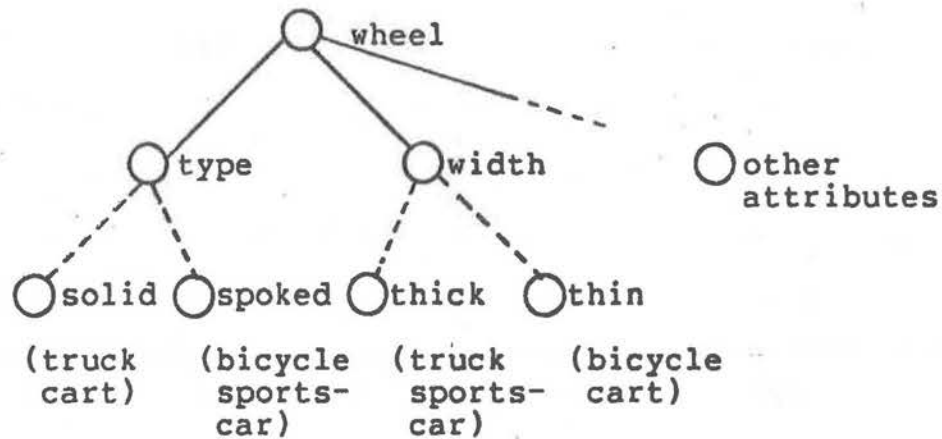


Figure 4.3.1. A simple set labelling structure.

When an instance of a wheel is detected, it acts a cue for any of the four vehicles, but as attribute values of the wheel are obtained, the set of possible models becomes automatically constrained to the set corresponding to the attribute value. If both attributes become available, simple set intersection of the models for each attribute value will further constrain the possibilities. Thus partial information, as might be available in the view of the wheel from the

side (type only) will be useful, and additional information will always be well used. The order of appearance of the attributes is of no consequence, as would be the case if the encoding were either procedural or in the form of a discrimination tree.

A preliminary analysis of the grammatical representation of the problem domain can easily produce such set labelling structures for any feature's attribute which can be appropriately quantized.

For the body knowledge grammar, lines, blobs, and connections are treated in this way. Consider what happens in the case of lines: The description for "line-hand-5" includes

(\$1 line (curve 17))

This means that any line with curvature of 17 can act as a cue for the "\$1" component of the model "line-hand-5". Since these image construction descriptions are intended as prototypes, we would also expect lines with similar curvature values to cue this role in the model. Thus the range over which the attribute values may vary, and still fulfill the description, is expanded out to an interval by an arbitrary extent. In the case of line curvature, the range is expanded 8 degrees from the prototype, so any line with curvature between 9 and 25 degrees will cue the model. Each use of "line" in a description can be similarly analyzed, until a

large list of curvature ranges with cued models has been produced. Over this list, overlapping portions of adjacent ranges are merged, until a list of mutually exclusive ranges with their attached set of cued models has resulted. Some adjacent ranges may differ only slightly in the list of models they cue, so another merging operation collapses across such ranges. The result is a partitioning of the range of values that the attribute "curvature" can take on, together with a list of possible models, that is, a list of all description-label pairs which specify a line with the attribute within the range. A partial example is shown in figure 4.3.2.

```
(line curve
      . . . . .
      (9 18
        (line-lower-leg-1 component $3)
        (line-lower-leg-2 component $2)
        (line-hand-3 component $4)
        (line-hand-3 component $2)
        (line-hand-4 component $5)
        (line-hand-4 component $3)
        (line-trunk-3 component $2)
        (line-trunk-4 component $3)
        (line-head-1 component $4)
      (19 23
        (line-head-1 component $4)
        (line-head-1 component $3)
        (line-lower-leg-1 component $3)
        (line-lower-leg-2 component $2)
        (line-hand-3 component $2)
        (line-hand-4 component $5)
        (line-trunk-3 component $2)
        (line-trunk-4 component $3)
        (line-head-2 component $3)
        (line-head-2 component $2)
      . . . . .
```

Figure 4.3.2. A partial example of the set labelling data structure for the curvature of lines.



With this process we have accomplished both the generalization from the prototypical descriptions and the development of the set labelling structure. Whenever an instance of a line is found in the image, and its curvature is known, it will store the value of that attribute as a pointer into the set labelling structure, because for the purpose of the interpretation it need not be known more precisely than the range into which it falls.

The same procedure is carried out for the "ratio" attribute of "blobs", and for the "angle" and "ratio" attributes of "connections". In the case of "connections", the set labelling is used to its full advantage because it is often the case that the angle between two lines can be determined (because it is a local property) but that the ratios of the connecting lines is not known.

A model possibility list is developed for every element used in a description, including terminals and non-terminals. In a sense this is a complete inversion of the grammar. The resulting structure can be thought of as a cue table (Mackworth, 1977a) for the entire body knowledge.

It is important to realize that it is the declarative and uniform structure of the knowledge grammar which permits the automatic development of these useful structures.[24]

---

[24] This part of the implementation, as well as all the following parts, was accomplished in Franzlisp, with the UNIX operating system on a VAX-11/780.

#### 4.4. Feature-Based Operations

Before the steps in the use of the body grammar in image interpretation can be explained we must first consider the availability of features. Features are made available in limited areas of the image, defined as concentric circles about a central fixation location. An inner circle defines the foveal area, an area in which line information is available, and a larger circle is the peripheral area, the area of available blob data. The center point of these circles, called the fixation center can be moved to any location in the image. These circles, whose radii are arbitrarily set, represent the availability of information at different resolution levels because of the structure of the human retina. Figure 4.4.1 shows an image with a typical fixation.

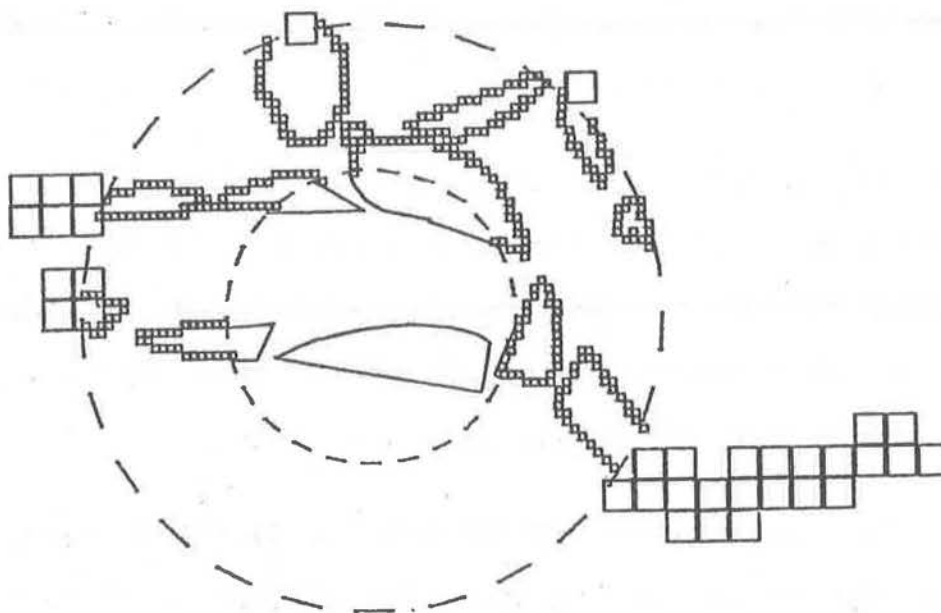


Figure 4.4.1. A typical fixation of an image. The area in 128x128 resolution indicates the periphery, and the 1024x1024 area is the fovea. The rest of the image is shown in 32x32 resolution.

Features within the fixation may not be complete in the sense that some of their attributes may not be determined. For example, the curvature of a line may be found, but one of the endpoints might lie outside of the fovea and so not be detected. Features will, however, always specify lists of model possibilities - lists of roles that they may play in object descriptions.

One approach to interpretation is to begin model invocation. This operation involves the examination of the descriptions specified in the model possibilities, followed by attempts to establish the existence of the remaining required features, and then testing the required relations among them. This approach can provide a dynamic determination of whether processing should proceed top-down or bottom-up (see Havens, 1978). As well, it can provide a means of iterative refinement of segmentation and interpretation (see Mackworth, 1978). Model invocation can be costly because of the search involved and the possibility of redundant operations.

In order to validate a model in terms of the image, the relations of the description must be verified with appropriate bindings of features. Model invocation methods are eventually used to accomplish these tests, and will be described in the context of the body grammar in the following section. The purpose of this section is to examine the idea that some of the model possibilities (or roles) might be eliminated before the complete model invocation procedures are used.

As we have seen in the review of the work of Waltz (1972) and Mackworth (1977b), the operations of network consistency may be used to reduce sets of possible bindings whenever an appropriate constraining relation can be identified.

As a step towards discovering such relations, we note that the coarse layer features can be expected to be larger than the fine layer features. Furthermore, several fine layer features can be expected within the same image extent as a single coarse layer feature. For the body drawing problem domain, each blob feature coincides with a number of line features. An image hierarchy retains this information, as described in section 4.1. In some cases, a line feature may be related to several blob features, particularly if the body parts in the original drawing are close together, but in many cases line features will only have this image hierarchy connection with a single blob feature.

In such cases of confidence about the coincidence of features from different layers, it seems a reasonable assumption that a group of line features which are all related to the same blob, will also exhibit the specific line-based relations which would be necessary for their confirmation of support for some more complex object. In a natural vision situation there would be many more sources of the formation of groups of fine layer features. Motion, colour, and texture could all provide the criteria for the development of rough groupings within which we might expect continuity of





specializations of the coarse layer feature's interpretations.

In the example, of the possible interpretations for "blob-17", only "extremity" has a counterpart among the remaining line interpretations (as either "line-hand-1" or "line-foot-3"), and so the others are eliminated. Similarly, the "line-lower-leg-1" possibility has no counterpart among the blob's interpretations, and so it is eliminated. The result is shown in figure 4.4.4.

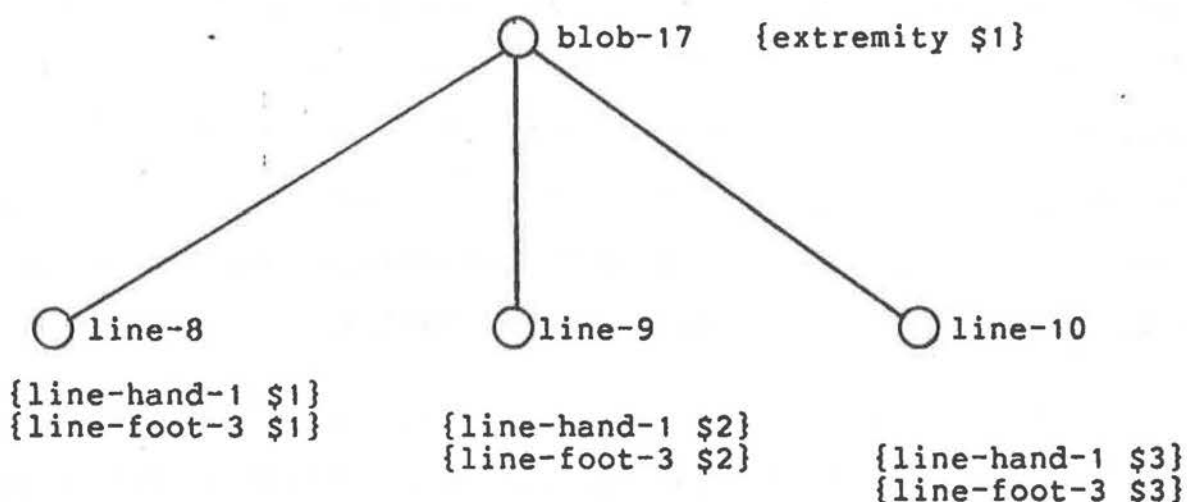


Figure 4.4.4. The final situation after the inter-level consistency has been applied.

Once these two consistency requirements are met, the number of remaining model possibilities is significantly reduced. One further constraining relation is available on the basis of the junctions between lines. Recall from section 4.3 that model possibilities are also assigned to the points of connection between lines. These roles are based on the



attributes (1) the angle at the junction, and (2) the ratio of the lengths of the lines forming the junction. These junctions must have interpretations which are compatible with the interpretations of the lines which meet at the junction. Again, this consistency requirement is required only for the generic models. For example, if one of the model possibilities for a junction is "{line-hand-1 \$4}", and one of the possibilities for a line involved in the junction is "{line-hand-1 \$3}", then they will be considered compatible. A closer examination might reveal that the required connection "\$4" is not intended to involve the line bound as "\$3". This more detailed examination based on the contents of the schemata descriptions is reserved for a point after the feature-based operations are complete. It is the intention that these feature-based operations remain simple enough that set intersection is adequate for their implementation.

The consistency requirement across junctions has the appropriate format for the application of network consistency methods. Full arc consistency, as reviewed in section 2.2, requires complete relaxation, with several iterations. Each of the feature based consistency relations is only applied in a single pass over the features, or junctions. The condition of consistency is not required by any of the subsequent processes, and the single pass makes a significant reduction in the number of model possibilities.



#### 4.5. Model-Based Operations

To review to this point, features such as lines, blobs and connections have been extracted within a fixation area of the image. On the basis of the properties of these features, lists of possible interpretations have been assigned. Each possible interpretation is really a role that the feature may play in one of the declarative schemata descriptions of more complex objects. The feature-based operations have made a major reduction in these lists of roles.

The purpose of the processes described in this section is to confirm the precise conditions as laid out by the schemata, and thereby assert the existence of more complex structures. These more complex structures will, in turn, be assigned model possibilities in still more complex objects.

In section 4.2 it was noted that the set of declarative schemata descriptions which encodes the body knowledge may be viewed as a grammar of the problem domain. The model-based operations may be seen as an attempt to parse the image, and develop a parse tree result. The leaf nodes of the parse tree are the collected features, the middle nodes are the simple body parts, the higher nodes are the more complex body parts, and the root node represents the entire body form.

Each node in this developing parse tree will be called a description instantiation, meaning that whenever one of the schemata descriptions is verified, one such node is generated.

Each node will be composed of three types of information: (1) its generic name (the schema of which it is an instance), (2) the arcs pointing to the nodes beneath it which act as its supporting evidence, and (3) attribute value pairs for the attributes which are specified in the schema.

Due to the non-uniform availability of features over the retina, it will often be the case that schemata instances will be partly developed, but not complete. For example, the schema description of "line-foot-1" might be satisfied by two lines and a junction, but the third line might not be available, either because it falls outside of fixation, or because it has not yet been considered. The approach which has been taken is to record these partial instantiations for any given schema as a network whose nodes are the possible bindings for the required objects and whose arcs are the relations required among objects. This section will present an algorithm which can be used to extract any newly completed instance of the schema which might result from the addition of a new binding possibility into the network.

The system's interpretation method is entirely bottom-up. This choice of strategy is not reflected at all in the body model representation, but is local to the control programs. Other planned versions of the system will be able to implement a variety of types of top-down control.

This section will outline and discuss the model-based operations in the context of a single fixation. The problem of intelligently selecting these locations, and of combining information across fixations will be deferred to the following section. This partition is natural because the size of the foveal and peripheral radii of feature availability is arbitrarily chosen, and it is possible to set these values to cover the entire image and thereby extract all features in a single fixation - as if the image subtended a very small visual angle.

The task is one of parsing to as high a level as can be supported by the available features. Each of the two layers of the grammar is complete, and can be used independently, so only the fine layer will be discussed. This layer is more complex because of the "elaboration structure" in the attribute specification portions of the descriptions.

There are a number of issues relating to this phase of interpretation. In this section we shall concentrate on two particular issues and show how they motivate the mechanism developed for interpretation. The first issue, termed the "locally legal interpretations issue" is a result of the uncertainty of the order in which elements should be considered, combined with the commitment to initiate the development of the understanding of the scene before having extracted features over the entire image. The second, the issue of "representing relation instances", is a result of the local

uncertainty of the underlying three dimensional structure of a self deforming scene object such as the human body.

#### 4.5.1. Locally Legal Interpretations Issue

Each prototypical schemata description of image constructions in the grammar is unique. However, once the generalization over attribute specifications takes place (as shown in section 4.3), a collection of lines in the image may satisfy the criteria for a number of descriptions. This is particularly true for constructions intended to be at different scales. For example "line-hand-1" and "line-hips-2" are similar. As a result, locally legal interpretations will be found which turn out to be incorrect in a larger context, so it is important to not make too great a commitment to a completed description, by, for example, allowing it to control the parse, searching for its other required elements.

This problem is also encountered at a higher level (toward the root node) in the parse of a body form image. The problem is more vividly illustrated at this level. Suppose that the image shown in figure 4.5.1 is to be interpreted. It is apparent that the parts labelled "1" and "2" belong to the same leg (crossed in front of the body) and that parts "3" and "4" belong to the other. It might be the case that the first leg to be recognized in the image is the one made up of parts "2" and "3", which is completely legal in a local sense.



Figure 4.5.1. Locally legal, but globally illegal structures in body form problem domain.

One solution to the problem requires a mechanism whereby a single feature may support a number of hypothesized models, not only different models, but also several versions of the same model. The particular solution used here involves a variation of network consistency methods (Mackworth, 1977b), so first we shall examine the difficulties in applying those methods directly, a line of reasoning which will reintroduce the "locally legal interpretations" issue within a stricter formalism.

#### 4.5.2. Applying Network Consistency

A network may be established for each schemata description. The purpose of the network will be to retain a record of the features, found to any point, which express a possibility of supporting the description. Each object label in the description will be represented by a node, and the required relations between the objects will be the edges. The nodes will be viewed as variables, with possible bindings from the set of features which have specified the corresponding label in that description. For example, consider the schema description for the image construction "line-hand-1"[25]:

```
(line-hand-1 nil
  (component
    (($1 line (curve 53))
     ($2 line (curve 0))
     ($3 line (curve 0))
     (($4 connect ($1 $2) (angle 134) (ratio 108))
      ($5 connect ($2 $3) (angle 90) (ratio 225))
      ($6 connect ($3 $1) (angle 92) (ratio 41))))
```

We may form the network as shown in figure 4.5.2, for which the three required objects are nodes, and the relations are arcs.

---

[25] The attribute development portion has been deleted.

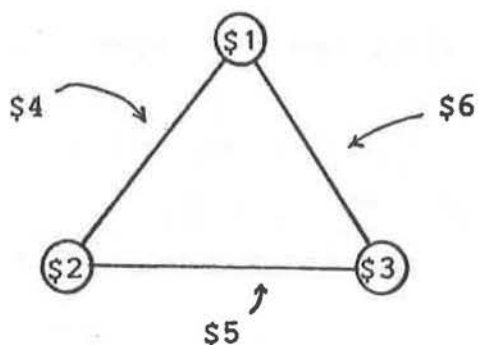


Figure 4.5.2. A network constructed from a schema description.

Within the fixation, roles of the features are considered, one at a time. Each role results in an entry to the network for the schema specified in the role. For example, if "line-1" is found to have the role "{line-hand-1 component \$1}", then the domain for the "\$1" node will be updated to reflect the possibility as shown in figure 4.5.3.

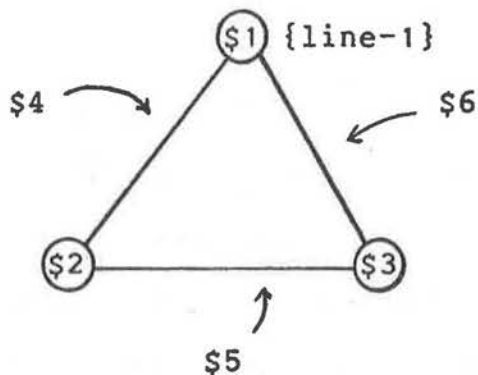


Figure 4.5.3. A network constructed from a schema description with an entry made.

As more roles, and other lines are considered, several entries will be made to the network. With each entry, the required relations among elements are examined, and if established, they are entered as part of the extension of the arcs. A

more advanced state of development is shown in figure 4.5.4.

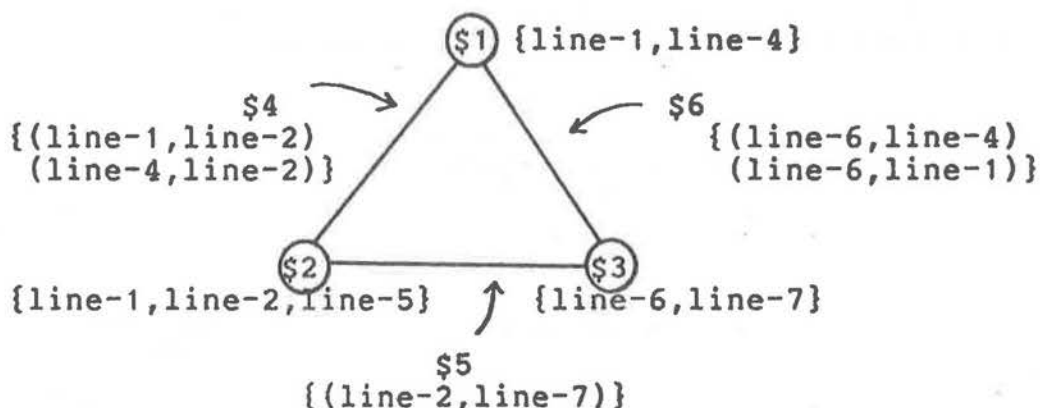


Figure 4.5.4. A network constructed from a schema description after several entries.

This is the classical format for the application of network consistency methods towards the reduction of the sets of possible bindings, and ultimately to determine instances of the description.

Due to the design goals of the system, there are reasons why these methods may not be applied directly. The central issue is a difference in approach. Network consistency methods rely on the availability of all information: the complete sets of possible variable bindings, and all relations among them. The spirit of this system is to reach some understanding after a minimum amount of feature extraction, in an incomplete knowledge situation, and in particular, to avoid the assumption of availability of relations among features in parallel over an image, an availability which has been demonstrated contrary to the operations of human vision (Treisman and Gelade, 1980).



For the example in figure 4.5.4, arc consistency would empty the domain for node "\$3" on the first pass, which would propagate to empty all the domains. The next fixation, or even the next feature in the same fixation might provide "line-3" as a possibility, with relations as shown in figure 4.5.5.

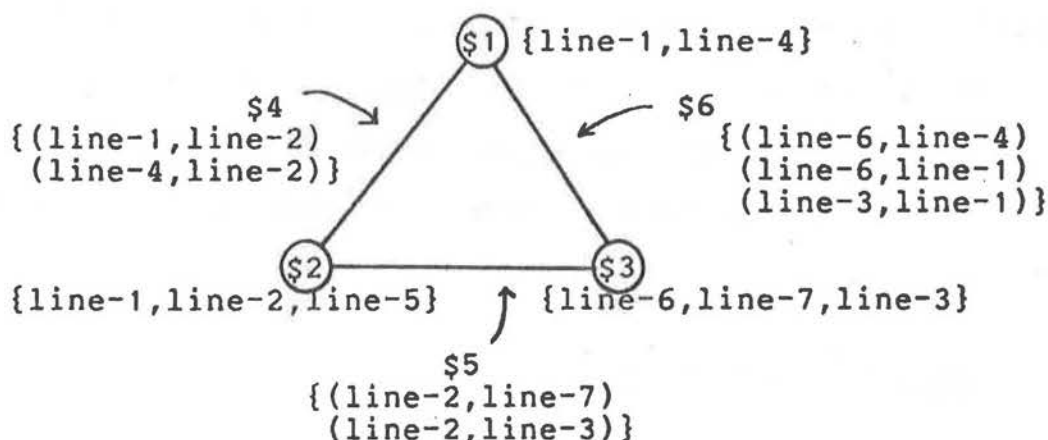


Figure 4.5.5. A network constructed from a schema description after several entries.

The conclusion is that the networks will have to be examined as each new piece of information becomes available. One straightforward way to do this is to apply arc consistency over the network each time a new variable is entered. This will, unfortunately, result in the rediscovery of solutions returned at previous points in the interpretation.

One possible remedy would be to remove variable bindings once they take part in some solution over the network. This would introduce the unfavourable condition of not being able to deal with the "locally legal interpretations" issue as

described earlier: removing a binding possibility excludes its involvement in other solutions.

A better solution is to restrict the application of the consistency methods. A temporary arc consistency is applied starting at the node for which a new entry is made. The set of domain variables for that node is confined to the single entry. The result is a list of all potential solutions which have not been previously returned. These solutions may, however, include domain variables which have been used in previous solutions for the same schema.

#### 4.5.3. Incremental Consistency

The following is a formalization of this variation of arc consistency, which will be called incremental consistency. The formulation follows closely after that of the AC-2 algorithm for arc consistency as provided by Mackworth (1977b).

For each node  $i$  of the network, assume  $F_i$  to be the set of all features which express the potential to fulfill the schema role represented by the node  $i$ . We would like to know, at all times the value of :

$(D_1, \dots, D_n)$  where  $D_i$  is a subset of  $F_i$  such that the elements of  $D_1, \dots, D_n$  are arc consistent.

Define the neighbourhood of a node in the network:

$Q_i = \{j | P_{ij} \text{ is required}\}$

where  $P_{ij}$  represents a required relation between nodes  $i$  and  $j$ . In the ongoing example from figure 4.5.5,

$$Q_i = \{\$4 \$6\}$$

For each  $x$  in  $D_i$  define for each  $j$  in  $Q_i$

$$R_{ijx} = \{y | P_{ij}(x,y)\}$$

This means that each relation is described as an extension, distributed over the elements which enter into the relation. For the example,

$$R(\$3, \$1, \text{line-6}) = \{\text{line-4}, \text{line-1}\}$$

Whenever a feature  $x$  specifies a role in  $i$  of the description, and should therefore be added to the network, we establish  $R_{ijx}$  for each  $j$  in  $Q_i$ . Then we apply  $NEW(x,i)$  which returns the subset of all arc consistent bindings which have not been previously returned.

The newly entered feature will be called the originating value. The algorithm first sets up a temporary domain for the node at which the originating value is entered. This domain  $D_i$  consists of that single value. The algorithm propagates outward from this node. As the propagation proceeds from node  $i$  to node  $j$ , if node  $j$  has not yet been visited then the working subset of variables  $D_j$  for that node is set to those values which meet the  $P_{ij}$  and  $P_{ji}$  relations with the values in

$D_i$ . If the node has been visited before,  $D_j$  will be intersected with the set of those which meet the relations. If the node  $j$  is updated in either way, it is put onto the list (REM) of nodes from which the propagation must yet take place.

```

procedure NEW(x,i)
   $D_i$   $\leftarrow$  {x}
   $D_k$   $\leftarrow$   $\emptyset$  for all  $k \neq i$ 
  REM  $\leftarrow$  {i}
  while REM not empty do
    begin
      select and delete any  $i$  from REM
      for each  $j$  in  $Q_i$  do
        begin
           $X_j$   $\leftarrow$   $\bigcup_{a \text{ in } D_i} R_{ija}$ 

          if  $D_j = \emptyset$  then
            begin
               $D_j$   $\leftarrow$   $X_j$ 
              REM  $\leftarrow$  REM  $\cup$  {j}
            end
          else if  $D_j$  not a subset of  $X_j$  then
            begin
               $D_j$   $\leftarrow$   $X_j \cap D_j$ 
              REM  $\leftarrow$  REM  $\cup$  {j}
            end
          if  $D_j = \emptyset$  then return nil
        end
      end
    end
  return { $D_1, \dots, D_n$ }
end NEW

```

Figure 4.5.6. Incremental Consistency Algorithm.

Note that the procedure will terminate and return "nil" if any of the originating value's required relations is not met for at least one value. Processing will only continue to the second iteration in the event that the originating value has each of its required relations fulfilled. At that point it is likely that there will be a new solution over the

network and the instantiation sets are being reduced. If a single value results in each of the  $D_i$ , then it is certain to be a solution, and if some  $D_i$  results with more than one entry, a search is required to find the actual solution.

This constitutes the first phase of the model-based operation: simply run through the list of model possibilities for each feature, and enter the possibilities into the network for the appropriate schema description. Then run the incremental consistency algorithm, and all newly formed sets of binding candidates will be returned. As we shall see, there are important steps that must be taken upon finding such satisfied descriptions, but the basic idea is that the object supported by the description will itself be introduced into a network for the description of a higher level object, and so forth, until the process can no longer develop more complex objects.

#### 4.5.4. Representing Relation Instances

We wish to avoid examining the conditions for the existence of a relation more than once for each pair of objects or features. If, for example, a "connect" relationship is found between two lines during consideration of their involvement in the description "line-hand-2", then we would like to retain information about their connection for examination in the event that these same two lines become candidates in some other schema description. For this reason, relation

instances have identities of their own, and carry attribute values in exactly the same way that objects do. Each object has associated with it a list of relation instances in which it is known to take part.

In some cases attempts to establish the relation instances are made "on demand" during the course of the interpretation, and in other cases, such as for line connections, relation instances are collected within the fovea exhaustively with each fixation, as if they were themselves features.

As seen in section 2.2, each description of objects has associated methods of developing attributes. Similarly, there are specific methods, of the same form, for the development of attributes of relations. Figure 4.5.7 shows the methods for the development of attributes for the "near" relation. The binding labels "\$1" and "\$2" indicate the two (scene) objects which have been judged to be "near".

```
(near
  ((ok <- (same (side $1) (side $2)))
   (ratio <- (times 100. (quotient (size $1) (size $2))))
   (angle-x <- (diff (car (a3d $1)) (car (a3d $2))))
   (angle-y <- (diff (cadr (a3d $1)) (cadr (a3d $2))))
   (angle-z <- (diff (caddr (a3d $1)) (caddr (a3d $2))))))
```

Figure 4.5.7. Example of specifications for the evaluation of attributes for a relation.

We have seen in section 4.2 that objects may have special attribute structures called "elaborations" which contain a

number of alternative sets of values which could not be definitely determined at the time of completing its description. In this case, the attribute values for the relation will also be stored as elaborations, representing the possible combinations of the elaborations of the two objects entering into the relation. This gives the appearance of a combinatorial explosion in the number of elaborations, but the bulk of the elaborations never contribute to any more complex structure. This is because there are constraints on the attributes of the relation which are required by the descriptions that specify the more complex objects. For example if a certain "hand" with two possible orientations is found to be "near" a lower arm with four possible orientations, then each of the attributes "angle-x", "angle-y", and "angle-z" will have eight possible orientations. But, the description for the object "arm" specifies a tight range of possible values for these attributes, and hence any instance of "arm" will only retain a few possible elaborations for that relation.

#### 4.6. Selecting Processing Locations

The previous sections have demonstrated how a parse tree might be developed out of available image features, such that the body structure and its attributes are represented. Within a single fixation, however, there may not be enough information to develop a root node, "body". The nature of the knowledge representation, together with the status of the interpretation provide an ideal means of intelligent selection of processing locations such that the entire interpretation can be effectively accomplished.

Before addressing this issue, we must consider what is meant by the term "interpretation". Interpretation might require that every line in the image be used in support of a complete parse tree for an object at the fine layer, with every attribute of each body part computed. In this case, the issue of selecting processing locations is not important. Since every location must be fixated foveally, a raster scan would be appropriate. On the other hand, interpretation might require only a complete body to be determined on the basis of coarse layer features, but the degree of uncertainty associated with the representation makes this alternative unattractive.

There is a compromise position which can be motivated by the phenomenology of vision. During the normal viewing of an object such as a bookcase, only a small portion of the field



of view is available in fine detail, and so one might be absolutely sure that a few of the books were actually books. The books which are not seen foveally will likely conform to some coarser representation for books. The coexistence of these two representations is adequate to permit the subjective experience of having seen all of the books in detail. It is quite unlikely that one would fixate on each book in a bookcase unless searching.

This idea can be expressed computationally within the body drawing interpretation system. For any coarse layer instance of an object, define a correspondence to be a fine layer object instance which meets the following criteria:

- (1) The instances are related by specialization-generalization links.
- (2) Attribute values of the two instances are similar, particularly the "size" attribute.
- (3) The two instances have roughly the same location.
- (4) The image construction which forms the basis of the fine layer object instance does not support any other (and different) object instances.

Of greater interest are the coarse layer object instances which do not have correspondences. We cannot be sure of the validity of these interpretations, yet there are still investigations that can be made. Suppose, for example, that the

coarse layer object has a component description, and that one of its components has a correspondence. In this case, we say that the composed object instance has an inferred correspondence. We could expand this definition to include objects which have a component with an inferred correspondence also.

With these concepts we can define the default objectives of interpretation to be the development of an instance of the body based on the coarse layer grammar, which has an inferred correspondence in the fine layer objects. Other, more specific demands of a task could produce the requirement for fine layer information about some specific body part and thereby extend the objectives, but in the absence of such requests the default objectives are adequate to confirm the existence of a body.

It follows that processing may be directed to areas of the image which can permit the most rapid arrival at the objectives. This can be formulated as

- (1) Foveal processing requirement: The locations of coarse layer objects are of interest for fixation if they compose some more complex object which has no correspondence at the fine layer.

This requirement pinpoints locations which have the greatest opportunity to propagate the certainty associated with foveal fixation out to the peripheral objects through correspondence. The body drawing interpretation system uses this rule as a

means of selecting fixation locations. The next chapter includes examples of the operations of the selection process.

There may not always be the appropriate configuration for the application of the foveal requirement, so another rule must be formulated.

- (2) Peripheral processing requirement: In the absence of foveal requirements, fixation locations should be selected to expand the area which is interpreted at the coarse layer.

In the body drawing system, this rule is implemented by making available an 32x32 grid over the image which indicates the amount of detail[26] in the grid square. Depending on the size of the peripheral radius, locations are selected which contain high detail such that the peripheral area will merge with the area already processed. This maximizes the chance of developing a foveal requirement, and at the same time works towards a complete coarse layer interpretation.

---

[26] This is simply a measure of the amount of line in the grid square.

## 5. Working Examples

This chapter demonstrates the operations of the computer implementation with the help of two examples. Outputs from computer runs are included, and are all prefixed with a vertical line to distinguish them from the annotations. The first example shows the processing taking place at a single fixation, with quite a wide field of view both peripherally and foveally. This example will be used to demonstrate the feature collection, feature-based model reduction, and the model invocation phases of interpretation. The second example shows a series of six fixation locations being selected by the system, and demonstrates the results at each step. This sequence is sufficient to demonstrate the location selection criteria and the integration across fixations.

### 5.1. A Single Fixation

The body form drawing that will be used in these examples is shown in figure 5.1.1. For this example, the radius of peripheral vision has been set at 375 units across the entire image area of 1024x1024, while the fovea was chosen as 325. This large fovea is used in order that the interpretation processes can be shown to develop to the point of recognizing complex structures, without requiring refixation. Figure 5.1.2 and 5.1.3 indicates the areas that are included in the example.

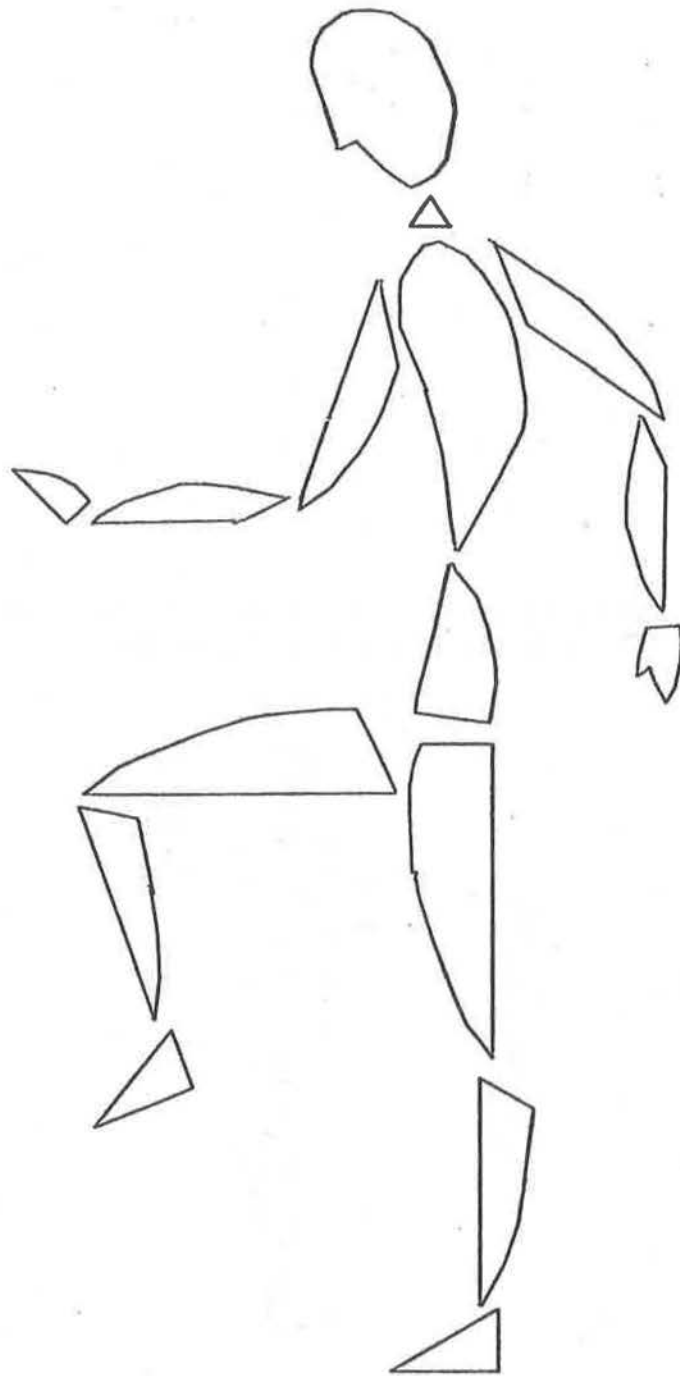


Figure 5.1.1 The body form line drawing to be used as the example.

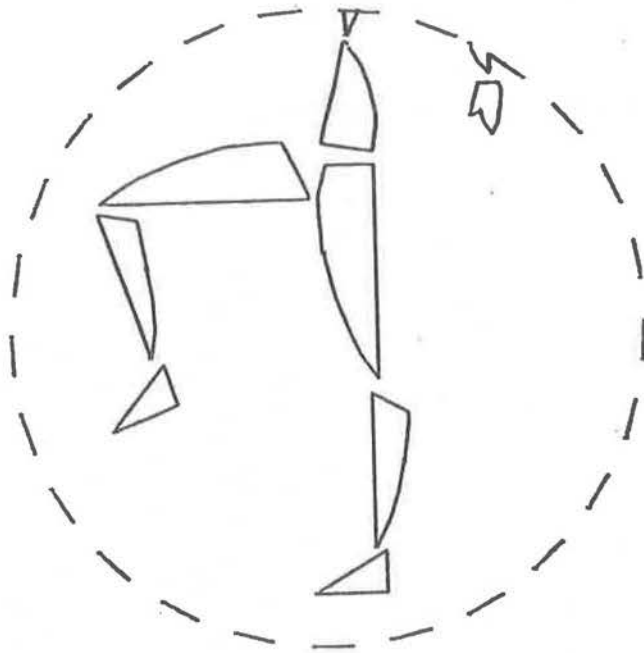


Figure 5.1.2. Area of available fine layer features in the single fixation at point (350,325).



Figure 5.1.3. Area of available coarse layer features in the single fixation at point (350,325).

The system is instructed to fixate at location (350,325). Next a list is produced of all the features which were collected within the fixation. Their properties are included in the list where known, and are otherwise nil. The number of possible roles that the feature may take in the interpretation is listed at the far right hand side under the heading "models".

-> (see1 350 325)

after feature collection at 350 325 (325/375)

node	point 1	point 2	curve	models
line-30	512 541	503 536	-8 1	82
line-29	524 517	512 542	9 18	12
line-28	535 571	524 517	36 44	12
line-27	512 570	535 571	-8 1	82
line-26	503 536	512 570	9 18	12
line-25	524 579	nil nil	-8 1	82
line-24	524 579	nil nil	54 61	10
line-23	199 251	129 225	-8 1	82
line-22	184 293	199 251	-8 1	82
line-21	129 225	184 293	-8 1	82
line-20	160 438	171 299	19 23	14
line-19	117 447	160 438	-8 1	82
line-18	171 299	117 447	-8 1	82
line-17	121 456	309 514	36 44	12
line-16	339 456	309 514	-8 1	82
line-15	121 456	339 456	-8 1	82
line-14	409 54	333 54	-8 1	82
line-13	409 99	409 54	-8 1	82
line-12	333 54	409 99	-8 1	82
line-11	434 236	396 101	19 23	14
line-10	396 259	434 236	-8 1	82
line-9	396 101	396 259	-8 1	82
line-8	405 272	353 490	46 53	12
line-7	405 490	353 490	-8 1	82
line-6	405 272	405 490	-8 1	82
line-5	349 512	376 614	-8 1	82
line-4	403 503	349 512	-8 1	82
line-3	376 614	403 503	45 45	16
line-2	nil nil	379 623	93 104	6
line-1	nil nil	379 623	9 18	12

node	center	longaxis	shortaxis	ratio	models
blob-12	525 551	525 576 528 520 512 552 536 552	213 237	4	
blob-11	517 651	nil nil 528 592 504 648 528 648	nil nil	0	

blob-10	414	192	405	127	400	260	432	188	400	196	363	512	2
blob-9	383	557	382	611	400	512	368	560	408	560	238	297	3
blob-8	385	398	401	298	379	496	408	408	360	392	363	512	2
blob-7	385	70	350	59	416	96	392	56	384	88	213	237	4
blob-6	nil	nil	391	653	nil	nil	nil	nil	nil	nil	nil	nil	0
blob-5	nil	nil	288	672	nil	nil	nil	nil	nil	nil	nil	nil	0
blob-4	175	262	152	245	192	288	184	248	168	272	203	212	3
blob-3	199	659	nil	nil	256	664	200	648	nil	nil	nil	nil	0
blob-2	248	484	152	469	340	464	252	464	244	512	363	512	2
blob-1	156	390	171	328	124	448	176	396	144	388	363	512	2

Next the system goes through the steps to reduce the models using the feature-based operations as described in section 4.4. The three steps shown below correspond to the three consistency relations which are exploited. The step marked "2-level" indicates the interlevel consistency and the "C-filter" step is the consistency at junctions of lines. The printout shows the reduction in terms of the total number of models for the line features at each step in the process. Note that the image hierarchy information is not adequate for "line-16" to enter into the groupings of features, and hence its model possibilities are not reduced.

```

..... before grouping 1690 models at line level
..... after  grouping 824 models at line level
..... after  2-level  435 models at line level
..... after  C-filter 170 models at line level

```

after feature-based model reduction

node	point 1	point 2	curve models
line-30	512	541 503 536	-8 1 2
line-29	524	517 512 542	9 18 2
line-28	535	571 524 517	36 44 1
line-27	512	570 535 571	-8 1 2
line-26	503	536 512 570	9 18 2
line-25	524	579 nil nil	-8 1 16
line-24	524	579 nil nil	54 61 8
line-23	199	251 129 225	-8 1 3
line-22	184	293 199 251	-8 1 3
line-21	129	225 184 293	-8 1 3
line-20	160	438 171 299	19 23 1



line-19	117	447	160	438	-8	1	2
line-18	171	299	117	447	-8	1	2
line-17	121	456	309	514	36	44	12
line-16	339	456	309	514	-8	1	82
line-15	121	456	339	456	-8	1	2
line-14	409	54	333	54	-8	1	3
line-13	409	99	409	54	-8	1	3
line-12	333	54	409	99	-8	1	3
line-11	434	236	396	101	19	23	1
line-10	396	259	434	236	-8	1	2
line-9	396	101	396	259	-8	1	2
line-8	405	272	353	490	46	53	1
line-7	405	490	353	490	-8	1	2
line-6	405	272	405	490	-8	1	2
line-5	349	512	376	614	-8	1	2
line-4	403	503	349	512	-8	1	2
line-3	376	614	403	503	45	45	1
line-2	nil	nil	379	623	93	104	2
line-1	nil	nil	379	623	9	18	1

node	center		longaxis			shortaxis			ratio models				
blob-12	525	551	525	576	528	520	512	552	536	552	213	237	1
blob-11	517	651	nil	nil	528	592	504	648	528	648	nil	nil	0
blob-10	414	192	405	127	400	260	432	188	400	196	363	512	1
blob-9	383	557	382	611	400	512	368	560	408	560	238	297	2
blob-8	385	398	401	298	379	496	408	408	360	392	363	512	2
blob-7	385	70	350	59	416	96	392	56	384	88	213	237	4
blob-6	nil	nil	391	653	nil	nil	nil	nil	nil	nil	nil	nil	0
blob-5	nil	nil	288	672	nil	nil	nil	nil	nil	nil	nil	nil	0
blob-4	175	262	152	245	192	288	184	248	168	272	203	212	3
blob-3	199	659	nil	nil	256	664	200	648	nil	nil	nil	nil	0
blob-2	248	484	152	469	340	464	252	464	244	512	363	512	1
blob-1	156	390	171	328	124	448	176	396	144	388	363	512	1

The system now goes through all the known coarse layer features and attempts the model-based operations as described in section 4.5. Not all of the blob features have enough known about their attributes to suggest possible models. First, the feature "blob-12" is considered as a possible binding for the "\$1" role in the "image1" description of "extremity". The full description is as follows:

```
(extremity nil
  (image1
    (($1 blob (ratio (200 300)))) nil
```

```
((ends <- (list (pt11 $1) (pt21 $1)))
 (location <- (cofg $1))
 (size <- (lengthb $1))
 (delta <- (times .25 (lengthb $1))) )
```

There is only one required object for this description, and as is shown below, the requirement on its "ratio" property is met, so the description's requirements are fulfilled, and a new node "extremity-1" is constructed, and its property values generated as specified in the description out of the properties of the line feature.

```
attempting solutions for blob-12 as (extremity image1 $1)
master-role list for blob-12 in (extremity image1 $1)
(---- nil $1 (nil blob-12 (nil)))
attempt to elaborate extremity from (($1 . blob-12))
verifying ($1 blob-12)
with ((ratio (200 300))) found= (213 237)
node:extremity-1
  type      extremity
  description image1
  bindings  (($1 blob-12))
  ends      ((525 . 576) (528 . 520))
  location  (525 . 551)
  size      56.08
  delta     14.02
```

Similarly, possible model roles are considered for the other blob features, resulting in the generation of the following nodes:

```
node:lower-limb-1
  type      lower-limb
  description image
  bindings  (($1 blob-10))
  ends      ((405 . 127) (400 . 260))
```

```

location      (414 . 192)
size          133.09
delta         26.61

node:central-body-1
type          central-body
description   image1
bindings     (($1 blob-9))
ends         ((382 . 611) (400 . 512))
location     (383 . 557)
size         100.62
delta        35.21

node:extremity-2
type          extremity
description   image1
bindings     (($1 blob-9))
ends         ((382 . 611) (400 . 512))
location     (383 . 557)
size         100.62
delta        25.15

node: upper-limb-1
description   image
bindings     (($1 blob-8))
ends         ((401 . 298) (379 . 496))
location     (385 . 398)
size         199.21
delta        39.84

```

Up to this point, an attempt has been made to enter each of the newly established nodes into a model role for some more complex construction, such as "limb". All of these attempts so far have failed to inspire any close examination because the relations required to satisfy such models have not been established. The node "lower-limb-1" does, however, meet a relation with another node, and so the following description for "limb" is considered.

```

(limb nil
  (component
    (($1 extremity)
    ($2 lower-limb)

```

```

($3 upper-limb))
(($4 b-connect ($1 $2 nil nil)
  (ratio (25 60)))
 ($5 b-connect ($2 $3 nil nil)
  (ratio (60 80))))
((proximal-end <- (free2 $5))
 (distal-end <- (free1 $4))
 (location <- (location $5))
 (delta <- (delta $3))
 (size <- (times 2.3 (size $2))))

(specialization (($1 right-arm)
  ($2 right-leg)
  ($3 left-arm)
  ($4 left-leg)) nil nil) )

```

In the examination of the possible fulfilled models, it is noted that the node being considered, "upper-limb-1", has all of the relations which are required of it in the description, and so the Incremental Consistency algorithm described in section 4.5 attempts to return a list of potential bindings, and fails because no appropriate "extremity" has been encountered yet.

```

**** relation b-connect-1 established between lower-limb-1
                                           upper-limb-1

attempting solutions for upper-limb-1 as (limb component $3)

    rem= ($2)
    dj= (lower-limb-1)
    xj= (lower-limb-1)
    dlist= (nil $3 (upper-limb-1)
           $2 (lower-limb-1))

    rem= ($1)
    dj= nil
    xj= nil
    dlist= (nil $3 (upper-limb-1)
           $2 (lower-limb-1)
           $1 nil)

solutions not found

```

As more blob features are considered, more nodes are generated. When the next "lower-limb" is encountered, another attempt is made to establish a node for "limb". The Incremental Consistency algorithm actually returns a binding list as a potential solution, but a subsequent examination discovers that the upper and lower parts of the potential limb are supported by the same image construction ("blob-8") and so the node is not generated.

```

node:lower-limb-2
  type      lower-limb
  description image
  bindings  (($1 blob-8))
  ends      ((401 . 298) (379 . 496))
  location  (385 . 398)
  size      199.21
  delta     39.84

**** relation b-connect-2 established between extremity-2
                                           lower-limb-2
**** relation b-connect-3 established between lower-limb-2
                                           upper-limb-1

attempting solutions for lower-limb-2 as (limb component $2)

      . . . . .

      rem= ($1)
      dj= (lower-limb-2)
      xj= (lower-limb-1 lower-limb-2)
      dlist= (nil $2 (lower-limb-2)
              $1 (extremity-2)
              $3 (upper-limb-1))

      rem= nil
      dj= (lower-limb-2)
      xj= (lower-limb-2)
      dlist= (nil $2 (lower-limb-2)
              $1 (extremity-2)
              $3 (upper-limb-1))

      solutions returned

      (((($2 . lower-limb-2) ($1 . extremity-2) ($3 . upper-limb-1)))

```

master-role list for lower-limb-2 in (limb component \$2)

```
(----- nil
  $1
  (nil extremity-1
    (nil)
    extremity-2
    (nil $4 (lower-limb-2)))
  $2
  (nil lower-limb-1
    (nil $5 (upper-limb-1))
    lower-limb-2
    (nil $4 (extremity-2) $5 (upper-limb-1)))
  $3
  (nil upper-limb-1 (nil $5 (lower-limb-2 lower-limb-1))))
```

```
attempt to elaborate limb from (($2 . lower-limb-2)
                                ($1 . extremity-2)
                                ($3 . upper-limb-1))
```

not-unique

The system proceeded considering the blob features, and generates more nodes:

```
node:central-body-2
  type          central-body
  description    image2
  bindings       (($1 blob-7))
  ends           ((385 . 70))
  location       (385 . 70)
  size           75.66
  delta          37.83

node:central-body-3
  type          central-body
  description    image1
  bindings       (($1 blob-7))
  ends           ((350 . 59) (416 . 96))
  location       (385 . 70)
  size           75.66
  delta          26.48

node:lower-limb-3
  type          lower-limb
  description    image
  bindings       (($1 blob-7))
  ends           ((350 . 59) (416 . 96))
```

```

location      (385 . 70)
size          75.66
delta         15.13

node:extremity-3
type          extremity
description   image1
bindings      (($1 blob-7))
ends          ((350 . 59) (416 . 96))
location      (385 . 70)
size          75.66
delta         18.91

**** relation b-connect-4 established between extremity-3
                                                lower-limb-1
**** relation b-connect-5 established between extremity-3
                                                lower-limb-3

```

As new relations are found for the node "extremity-3", another attempt is made to find a solution from among the possible bindings for the components of "limb". This time all the requirements are met.

```

attempt to elaborate limb from (($1 . extremity-3)
                                ($2 . lower-limb-1)
                                ($3 . upper-limb-1))

verifying ($1 extremity-3) with nil
verifying ($2 lower-limb-1) with nil
verifying ($3 upper-limb-1) with nil

**** relation b-connect-6 established between extremity-3
                                                lower-limb-1

verifying ($4 b-connect-6)
with ((ratio (25 60))) found= 56

**** relation b-connect-7 established between lower-limb-1
                                                upper-limb-1

verifying ($5 b-connect-7)
with ((ratio (60 80))) found= 66

node:limb-1
type          limb

```

```

description  component
bindings    ($5 b-connect-7)
            ($4 b-connect-6)
            ($1 extremity-3)
            ($2 lower-limb-1)
            ($3 upper-limb-1)
proximal-end (379 . 496)
distal-end   (350 . 59)
location     (400 . 279)
delta        39.84
size         306.11

```

```

**** relation b-connect-8 established between limb-1
                                           central-body-1

```

The current status of the interpretation is summarized in the partial parse tree shown below:

```

| limb-1
|   extremity-3 (blob-7)
|   lower-limb-1 (blob-10)
|   upper-limb-1 (blob-8)

```

The process continues, finding more basic coarse level body parts until a second limb is detected.

```

node:central-body-4
  type      central-body
  description image2
  bindings  (($1 blob-4))
  ends      ((175 . 262))
  location  (175 . 262)
  size      58.72
  delta     29.36

node:central-body-5
  type      central-body
  description image1
  bindings  (($1 blob-4))
  ends      ((152 . 245) (192 . 288))
  location  (175 . 262)
  size      58.72
  delta     20.55

node:extremity-4
  type      extremity

```



```

description image1
bindings (( $1 blob-4))
ends ((152 . 245) (192 . 288))
location (175 . 262)
size 58.72
delta 14.68

node:upper-limb-2
type upper-limb
description image
bindings (( $1 blob-2))
ends ((152 . 469) (340 . 464))
location (248 . 484)
size 188.06
delta 37.61

node:lower-limb-4
type lower-limb
description image
bindings (( $1 blob-1))
ends ((171 . 328) (124 . 448))
location (156 . 390)
size 128.87
delta 25.77

node:limb-2
type limb
description component
bindings ($5 b-connect-16)
($4 b-connect-15)
($2 lower-limb-4)
($1 extremity-4)
($3 upper-limb-2)
proximal-end (340 . 464)
distal-end (152 . 245)
location (138 . 458)
delta 37.61
size 296.41

```

Now that a second limb has been established, the description for the "body-half" is satisfied. The actual schema description is provided below:

```

(body-half nil
  (component
    (( $1 limb)
     ($2 limb)
     ($3 central-body))
    (($4 b-connect ($1 $3 proximal-end nil)
      (ratio (150 450)))

```

```

($5 b-connect ($2 $3 proximal-end nil)
  (ratio (150 450)))
((head-end <- (midpoint (proximal-end $1)
  (proximal-end $2)))
  (center-end <- (free2 $4))
  (location <- (location $3))
  (delta <- (delta $3))
  (size <- (plus (size $1) (size $3))))
(specialization (($1 upper-body)
  ($2 lower-body)) nil nil) )

```

```

node:body-half-1
  type          body-half
  description   component
  bindings      ($5 b-connect-19)
                ($4 b-connect-18)
                ($2 limb-2)
                ($3 central-body-1)
                ($1 limb-1)
  head-end      (359 . 480)
  center-end    (382 . 611)
  location      (383 . 557)
  delta         35.21
  size          406.73

```

after coarse models invoked

The "body-half" was the largest structure which could be supported in the context of the limited diameter of peripheral features. To this point, each of the model possibilities for the blob features has been entered into the interpretation process, and so now the fovea is processed. The parse tree for the body-half is shown below:

```

body-half-1
  limb-1
    extremity-3 (blob-7)
    lower-limb-1 (blob-10)
    upper-limb-1 (blob-8)
  limb-2
    extremity-4 (blob-4)
    lower-limb-4 (blob-1)
    upper-limb-2 (blob-2)
  central-body-1 (blob-9)

```

The operations for the fovea are identical to those for the periphery, using the same routines. At some point a node is generated for "line-foot-1".

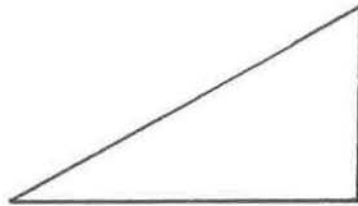
```
node:line-foot-1-1
  type          line-foot-1
  description    component
  bindings       ($6 connect-7)
                 ($5 connect-8)
                 ($4 connect-9)
                 ($1 line-21)
                 ($2 line-22)
                 ($3 line-23)
  a2d            20
  proximal-end  (184 . 293)
  location       (160 . 248)
  size           87
```

Working in a strict bottom-up fashion, the system recognizes that the scene structure "foot" can be supported by the image construction "line-foot-1". At this point the system cannot know whether it will be a right or left foot, and so both possibilities are retained as the elaborations of the node for the foot.

```
attempt to elaborate foot from
(($1 . line-foot-1-1))
verifying ($1 line-foot-1-1) with nil
(($1 line-foot-1-1))
node:foot-1
  type          foot
  description    image1
  bindings       (($1 line-foot-1-1))
  proximal-end  (184 . 293)
  size           87
  location       (160 . 248)
  a3d            (-20 90 0)
  extra          (E00007 E00008)
                  elaboration:E00007
                  side           right
                  elaboration:E00008
```

side            left

The "line-foot-1" construction is shown below.



If it is the left foot, then the outside is facing the viewer, and if it is the right, then the inside faces the viewer. In either case, the three dimensional rotation from the rest position is the same, so the attribute "a3d" (three-dimensional orientation) does not appear in the elaboration, but rather in the main node.

An image construction for "line-lower-leg-1" is developed next, which in turn prompts the generation of a node for "lower-leg".

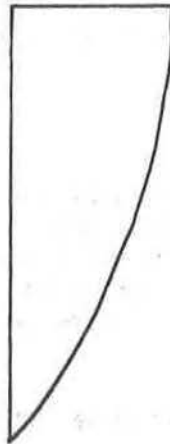
```
node:line-lower-leg-1-1
  type          line-lower-leg-1
  description   component
  bindings      ($6 connect-10)
                ($5 connect-11)
                ($4 connect-12)
                ($1 line-18)
                ($2 line-19)
                ($3 line-20)
  size         157
  a2d          20
  proximal-end (117 . 447)
  location     (154 . 370)
  distal-end   (171 . 299)
```

```

attempt to elaborate lower-leg from
(($1 . line-lower-leg-1-1))
verifying ($1 line-lower-leg-1-1) with nil
(($1 line-lower-leg-1-1))
node:lower-leg-1
  type          lower-leg
  description    image1
  bindings       (($1 line-lower-leg-1-1))
  size           157
  location       (154 . 370)
  proximal-end   (117 . 447)
  distal-end     (171 . 299)
  extra          (E00009 E00010 E00011 E00012)
                 elaboration:E00009
                 side           left
                 a3d            (0 0 -20)
                 elaboration:E00010
                 side           left
                 a3d            (-20 90 0)
                 elaboration:E00011
                 side           right
                 a3d            (-20 90 0)
                 elaboration:E00012
                 side           right
                 a3d            (0 180 20)

```

Again the sides are kept in the elaborations, but in this case, different three-dimensional orientations are also possible. The image construction is shown below:



The bulged side may either be the outside or the back of the lower leg, producing different orientations relative to the rest position. Similarly, "upper-leg-1-1" is eventually generated:

```
node:upper-leg-1
  type      upper-leg
  description image4
  bindings  (($1 line-upper-leg-4-1))
  size      218
  location  (222 . 470)
  proximal-end (339 . 456)
  distal-end (121 . 456)
  extra     (E00017 E00018 E00019 E00020)
            elaboration:E00017
              side      left
              a3d       (0 180 -90)
            elaboration:E00018
              side      left
              a3d       (90 90 0)
            elaboration:E00019
              side      right
              a3d       (0 0 90)
            elaboration:E00020
              side      right
              a3d       (90 90 0)
```

Connections are made between the body parts in the form of "near" relations. As discussed in section 4.5, these relations takes on attribute value pairs in much the same way as do the interpretation nodes for objects. The method for the development of attribute values for the "near" relation is shown below:

```
((ok <- (same (side $1) (side $2)))
 (ratio <- (getratiox (size $1) (size $2)))
 (angle-x <- (diff (car (a3d $1)) (car (a3d $2))))
 (angle-y <- (diff (cadr (a3d $1)) (cadr (a3d $2))))
 (angle-z <- (diff (caddr (a3d $1)) (caddr (a3d $2))))))
```

The result of establishing that the upper and lower leg are "near" is shown in the following printout. Because the body

parts have several values for their orientations ("a3d") the attribute values of the "near-2" relation have elaborations (prefaced with the letter "X") to store the possibilities.

```

near-2
  (type near
   location (119 . 451)
   args (lower-leg-1 upper-leg-1 (117 . 447) (121 . 456))
   ratio 72
   extra (X00021 X00022 X00023 X00024 ...))

X00021
  (xargs (($2 . E00017) ($1 . E00009))
   angle-z 70
   angle-y -180
   angle-x 0
   ok left)

X00022
  (xargs (($2 . E00017) ($1 . E00010))
   angle-z 90
   angle-y -90
   angle-x -20
   ok left)

X00023
  (xargs (($2 . E00018) ($1 . E00009))
   angle-z -20
   angle-y -90
   angle-x -90
   ok left)

X00024
  (xargs (($2 . E00018) ($1 . E00010))
   angle-z 0
   angle-y 0
   angle-x -110
   ok left)

  . . . . .

X00027
  (xargs (($2 . E00020) ($1 . E00011))
   angle-z 0
   angle-y 0
   angle-x -110
   ok right)

```

The system attempts to establish a "leg" on the basis of these body parts. During the process of testing the schema for "left-leg", the requirements for some of the component objects are found to not hold, and so some of the possible elaborations for the body parts are eliminated from further consideration in this context.

```
attempt to elaborate left-leg from (($3 . upper-leg-1)
                                   ($2 . lower-leg-1)
                                   ($1 . foot-1))

verifying ($1 foot-1 E00007 E00008)
with ((side left))
      found= nil
      req= (side left)
           trying elaboration: E00007
           found=right
           elaboration deleted
           trying elaboration: E00008
           found=left

. . . . .
```

Similarly, the required properties of the "near" relations are examined, and candidate elaborations for the relation nodes are eliminated.

```
verifying ($5 near-2 X00021 X00022 X00023 X00024 ...)
with ((angle-x (-145 10))
      (angle-y (0 0))
      (angle-z (0 0))
      (ratio 72))

. . . . .

      req= (angle-y (0 0))
           trying elaboration: X00021
           found=-180
           elaboration deleted
           trying elaboration: X00022
           found=-90
           elaboration deleted
           trying elaboration: X00023
           found=-90
```



```

elaboration deleted
trying elaboration: X00024
found=0

```

.....

The result is that not many of the elaboration possibilities remain valid in the context of a "left-leg". The bindings show all of the local possibilities, which are then examined for compatibility. The result is a single possible value for the orientation of the leg. The constraints of the allowable angles at the connections of the components of the leg has pruned the elaborations.

```

node:left-leg-1
  type          left-leg
  description    component
  bindings       ($5 near-2 X00024 X00027)
                 ($4 near-1 X00014 X00015)
                 ($3 upper-leg-1 E00017 E00018)
                 ($2 lower-leg-1 E00009 E00010)
                 ($1 foot-1 E00008)
  proximal-end  (339 . 456)
  size          392.5
  knee-location(121 . 456)
  location      (119 . 451)
  foot-base     (-20 0)
  extra         (E00029)
                elaboration:E00029
                xargs      ($5 . X00024)
                           ($4 . X00014)
                           ($3 . E00018)
                           ($2 . E00010)
                           ($1 . E00008)
                foot-posture 0
                knee-posture 110
                a3d         (90 90 0)

left-leg-1
  foot-2
    line-foot-1-1 (line-21 line-22 line-23)
  lower-leg-1
    line-lower-leg-1-1 (line-18 line-19 line-20)
  upper-leg-1

```

line-upper-leg-4-1 (line-15 line-17 line-16)

This same group of body parts is then developed into a "right-leg" node also. These two possibilities are valid, but with different values for the elaborations of the component parts.

```
node:right-leg-1
  type          right-leg
  description    component
  bindings       ($5 near-2 X00024 X00027)
                ($4 near-1 X00014 X00015)
                ($3 upper-leg-1 E00019 E00020)
                ($2 lower-leg-1 E00011 E00012)
                ($1 foot-1 E00007)
  proximal-end  (339 . 456)
  size          392.5
  knee-location(121 . 456)
  location      (119 . 451)
  foot-base     (-20 90)
  extra         (E00030)
                elaboration:E00030
                xargs      ($5 . X00027)
                          ($4 . X00015)
                          ($3 . E00020)
                          ($2 . E00011)
                          ($1 . E00007)
                foot-posture 0
                knee-posture 110
                a-3d      (90 90 0)
```

This process continues, until the other leg in the image is established, along with the "hips", and then a node is created for the entire "lower-body". Actually two such "lower-body" nodes are supported in the image, with different interpretations for the sides of the legs.

```
node:lower-body-1
  type          lower-body
  description    component
  bindings       ($5 near-8)
                ($4 near-5)
                ($3 hips-1)
                ($1 right-leg-1 E00030)
```

```

top          ($2 left-leg-2 E00053)
a3d         (376 . 614)
location    (10 90 0)
size        (369 . 535)

```

```

node:lower-body-2
type        lower-body
description component
bindings   ($5 near-7)
           ($4 near-6)
           ($3 hips-1)
           ($1 right-leg-2 E00054)
           ($2 left-leg-1 E00029)
top         (376 . 614)
a3d        (10 90 0)
location   (369 . 535)
size       509.0

```

At this point, the interpretation has gone as far as it can within the limited diameter of available features as defined by the foveal and peripheral radii. The complete parse tree to this point is:

```

lower-body-2
  right-leg-2
    foot-4
      line-foot-1-2 (line-12 line-13 line-14)
    lower-leg-2
      line-lower-leg-1-2 (line-9 line-10 line-11)
    upper-leg-2
      line-upper-leg-6-1 (line-8 line-7 line-6)
  left-leg-1
    foot-2
      line-foot-1-1 (line-21 line-22 line-23)
    lower-leg-1
      line-lower-leg-1-1 (line-18 line-19 line-20)
    upper-leg-1
      line-upper-leg-4-1 (line-15 line-17 line-16)
  hips-1
    line-hips-2-1 (line-3 line-4 line-5)

lower-body-1
  right-leg-1
    foot-2
      line-foot-1-1 (line-21 line-22 line-23)
    lower-leg-1
      line-lower-leg-1-1 (line-18 line-19 line-20)

```

```

    upper-leg-1
      line-upper-leg-4-1 (line-15 line-17 line-16)
left-leg-2
  foot-4
    line-foot-1-2 (line-12 line-13 line-14)
  lower-leg-2
    line-lower-leg-1-2 (line-9 line-10 line-11)
  upper-leg-2
    line-upper-leg-6-1 (line-8 line-7 line-6)
hips-1
  line-hips-2-1 (line-3 line-4 line-5)

```

At the line level, the interpretation is correct for 50% of the hypothesised constructions. Those which were incorrect could not take part in some larger structure and so do not appear in the parse tree.

In order to provide some idea of the amount of time taken by the interpretation processes, this same example was processed with peripheral and foveal diameters which covered the entire image. The CPU seconds taken on a VAX-11/780 are shown beside each step.

fixation (feature collection)	16 seconds
feature based model possibility reduction	125 seconds
blob based interpretation	42 seconds
line based interpretation	82 seconds
total	<u>265 seconds</u>

The entire system is written in Franzlisp, including mathematical operations and it runs interpretively.

## 5.2. Example with Multiple Fixations

This second example is different from the previous one in that the diameters of available features are smaller, and several fixation locations are selected by the system in order to accomplish the interpretation through the propagation of the fine layer results into the periphery. The system attempts to select locations which will facilitate this propagation.

The foveal radius has been reduced to only 125 units, and the periphery reduced to 250. The location of the first fixation is arbitrarily chosen to be (449 192). Figure 5.2.1 shows the areas processed. After the usual collection of features, and the feature based operations, a number of coarse body parts are supported, as shown below.

```
-> (see 449 192)
after feature collection at 449 192 (125/250)

node point 1 point 2 curve models
line-7 409 99 nil nil -8 1 82
line-6 nil nil 409 99 -8 1 82
line-5 434 236 396 101 19 23 14
line-4 396 259 434 236 -8 1 82
line-3 396 101 396 259 -8 1 82
line-2 405 272 nil nil 46 53 12
line-1 405 272 nil nil -8 1 82

node center longaxis shortaxis ratio models
blob-3 414 192 405 127 400 260 432 188 400 196 363 512 2
blob-2 385 398 401 298 nil nil 408 408 360 392 nil nil 0
blob-1 385 70 350 59 416 96 392 56 384 88 213 237 4

..... before grouping 436 models at line level
..... after grouping 242 models at line level
..... after 2-level 146 models at line level
..... after C-filter 89 models at line level
```

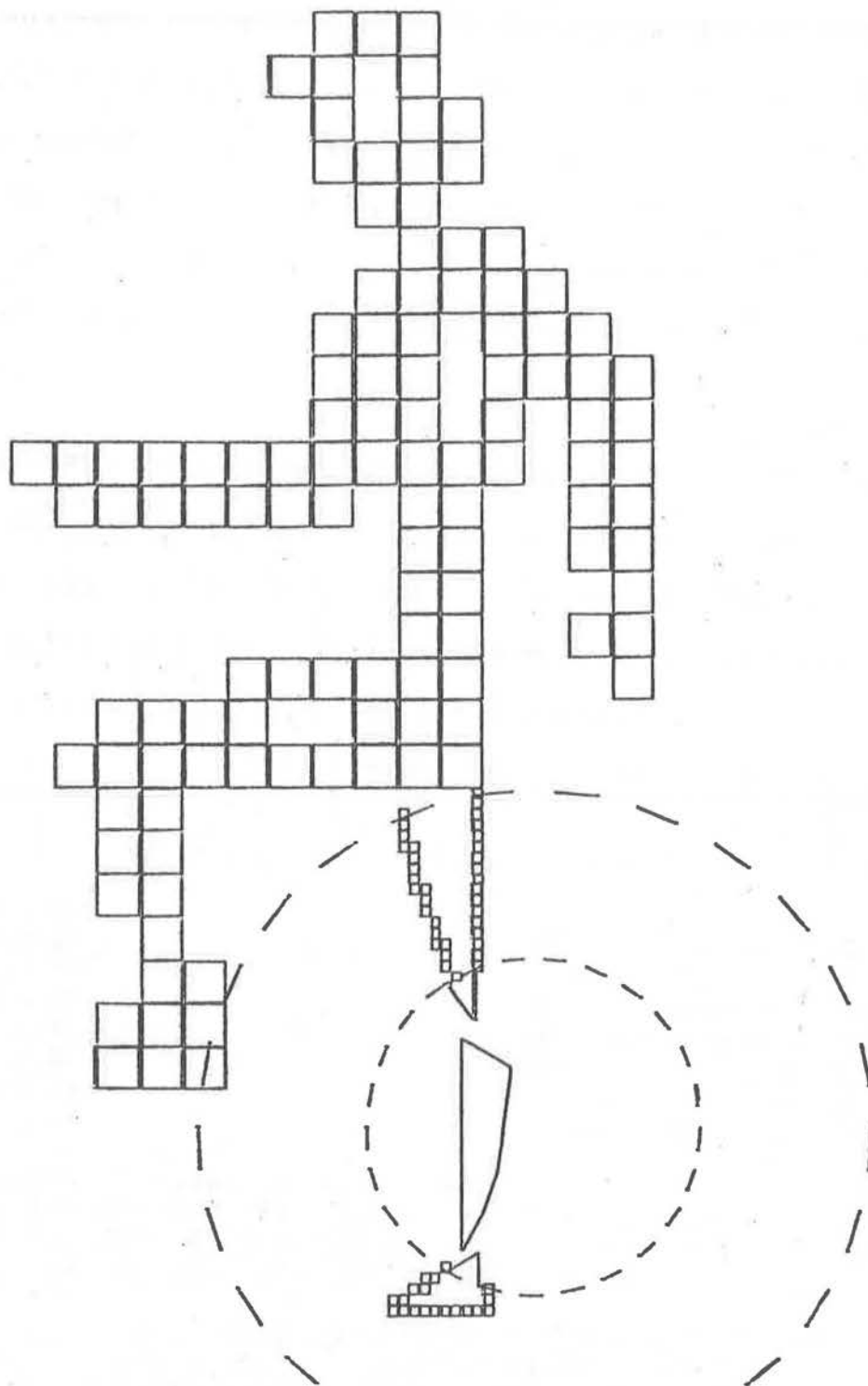


Figure 5.2.1. The first fixation (at location 449 192). The small squares indicate periphery, and the large squares show unprocessed areas.

```
node:lower-limb-1
  type          lower-limb
  description   image
  bindings      (($1 blob-3))
  ends          ((405 . 127) (400 . 260))
  location      (414 . 192)
  size          133.09
  delta         26.61
```

```
node:central-body-1
  type          central-body
  description   image2
  bindings      (($1 blob-1))
  ends          ((385 . 70))
  location      (385 . 70)
  size          75.66
  delta         37.83
```

```
node:central-body-2
  type          central-body
  description   image1
  bindings      (($1 blob-1))
  ends          ((350 . 59) (416 . 96))
  location      (385 . 70)
  size          75.66
  delta         26.48
```

```
node:lower-limb-2
  type          lower-limb
  description   image
  bindings      (($1 blob-1))
  ends          ((350 . 59) (416 . 96))
  location      (385 . 70)
  size          75.66
  delta         15.13
```

```
node:extremity-1
  type          extremity
  description   image1
  bindings      (($1 blob-1))
  ends          ((350 . 59) (416 . 96))
  location      (385 . 70)
  size          75.66
  delta         18.91
```

At the line level, there is one hypothesized object.

```
node:line-lower-leg-1-1
  type          line-lower-leg-1
```

```

description  component
bindings    ($6 connect-2)
            ($5 connect-3)
            ($4 connect-4)
            ($1 line-3)
            ($2 line-4)
            ($3 line-5)
size        158
a2d         0
proximal-end (396 . 259)
location     (405 . 174)
distal-end  (396 . 101)

node:lower-leg-1
type        lower-leg
description image1
bindings    (($1 line-lower-leg-1-1))
size        158
location    (405 . 174)
proximal-end (396 . 259)
distal-end  (396 . 101)
extra       (E00005 E00006 E00007 E00008)
            elaboration:E00005
            side          left
            a3d           (0 0 0)
            elaboration:E00006
            side          left
            a3d           (0 90 0)
            elaboration:E00007
            side          right
            a3d           (0 90 0)
            elaboration:E00008
            side          right
            a3d           (0 180 0)

```

To this point, there is no construction known at the coarse level which has components and yet does not have a corresponding model at the line level. Thus it is not possible to use the foveal requirement for processing location (as described in section 4.6). As a result, the peripheral requirement is used. An area of the image which has a lot of detail, as measured by the density of lines, is selected such that peripheral processing of the new location will merge with the existing periphery.



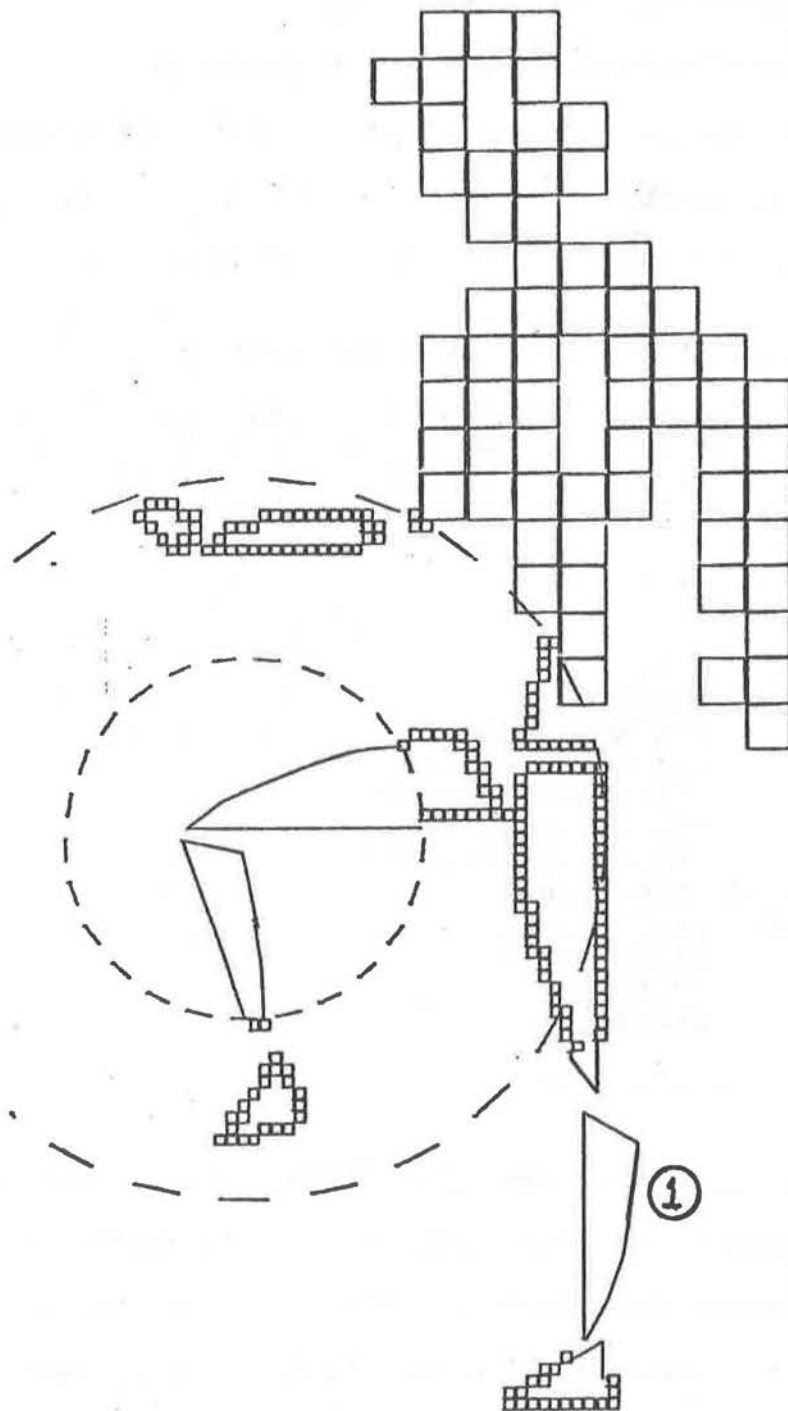


Figure 5.2.2. The second fixation at 162 448. The previously processed area is also shown.

```
|located: lower-limb-2 nil at ((385 . 70) (350 . 59) (416 . 96))
|located: lower-limb-1 as (lower-leg)
```

```
|next location selected as 162 448 peripheral
```

Processing at the second location (shown in figure 5.2.2) results in a number of models at the coarse level, the most interesting of which is the "limb-1" construction.

```
|after feature collection at 162 448 (125/250)
```

```
..... before grouping 361 models at line level
..... after  grouping 167 models at line level
..... after   2-level 117 models at line level
..... after   C-filter 117 models at line level
```

```
.....
|node:limb-1
|  type           limb
|  description    component
|  bindings       ($5 b-connect-7)
|                 ($4 b-connect-6)
|                 ($2 lower-limb-4)
|                 ($1 extremity-2)
|                 ($3 upper-limb-2)
|  proximal-end   (340 . 464)
|  distal-end     (152 . 245)
|  location       (138 . 458)
|  delta          37.61
|  size           296.41
```

.....

No line level object is developed which can correspond to any of the components of this limb, and so the basic criteria for the foveal requirement is met. The location of one of the components is chosen as the next fixation, as shown in figure 5.2.3.

```
|located: limb-1 uninstantiated
```

```
|next location selected as 160 288 foveal
```

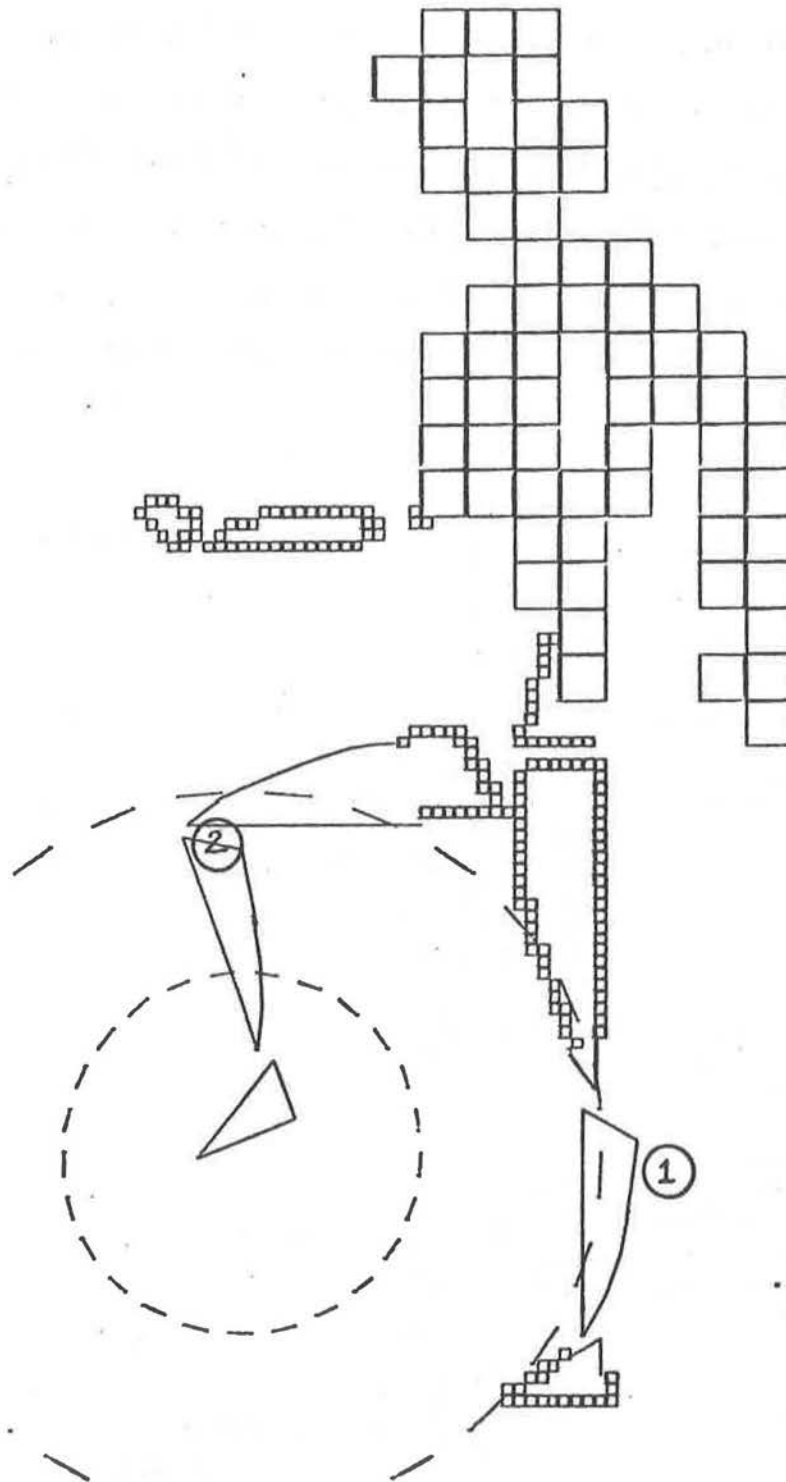


Figure 5.2.3. The third fixation at 160 228. The previously processed areas are also shown.

The result of processing at this new location is that "foot-1" is detected. This new line level object is recognized as corresponding to the "extremity-2" involved in "limb-1", so an inferred correspondence is made between "limb-1" and "leg". It is now no longer necessary to process the remaining portions of this leg with the high resolution fovea, because the more detailed interpretation has been propagated to the entire object.

after feature collection at 160 288 (125/250)

. . . . .

```
node:line-foot-1-1
  type      line-foot-1
  description  component
  bindings   ($6 connect-9)
             ($5 connect-10)
             ($4 connect-11)
             ($1 line-13)
             ($2 line-14)
             ($3 line-15)
  a2d       20
  proximal-end (184 . 293)
  location    (160 . 248)
  size       87

node:foot-1
  type      foot
  description  image1
  bindings   (($1 line-foot-1-1))
  proximal-end (184 . 293)
  size       87
  location    (160 . 248)
  a3d        (-20 90 0)
  extra      (E00011 E00012)
             elaboration:E00011
             side          right
             elaboration:E00012
             side          left

located: limb-1 as leg
```

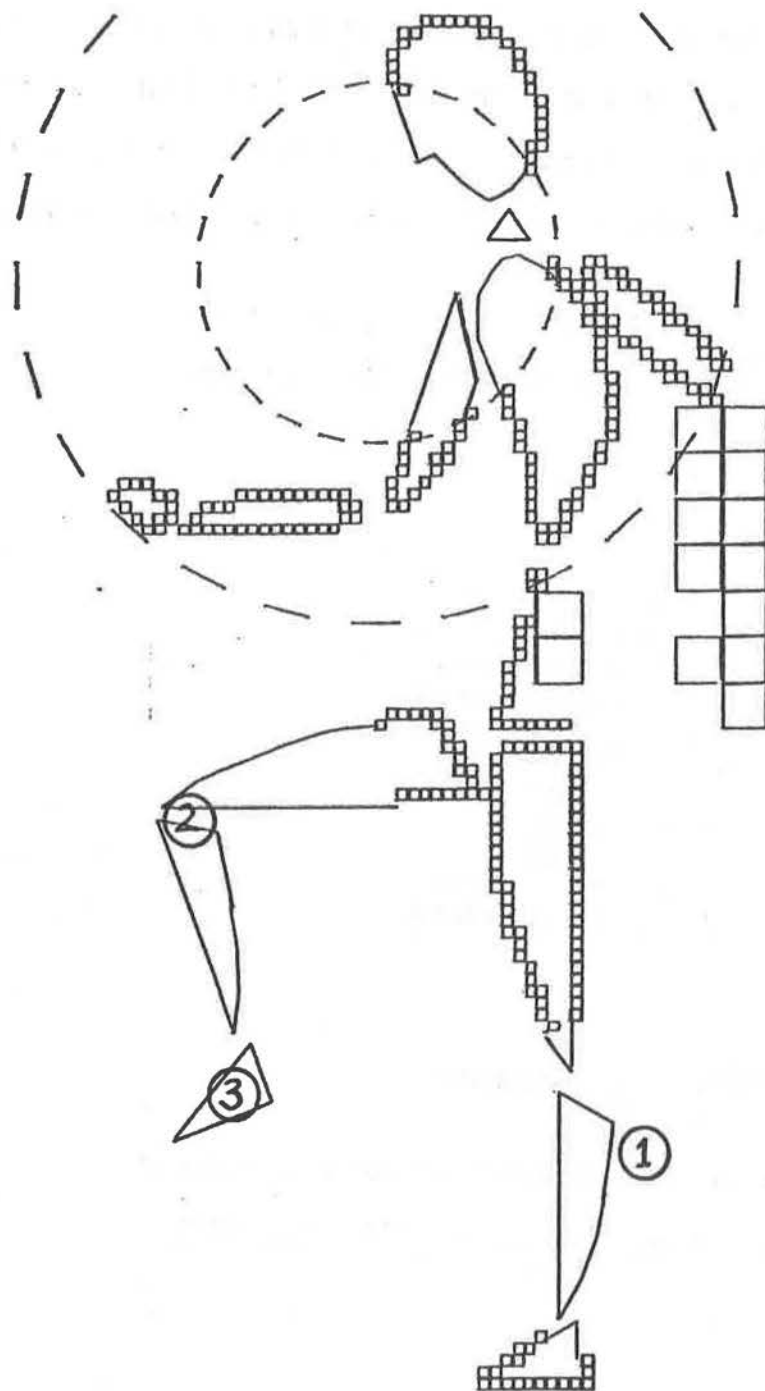


Figure 5.2.4. The fourth fixation at 270 832.

The next location, shown in figure 5.2.4, is chosen on the basis of peripheral requirement because no new coarse level objects have been detected. At the new location, another limb is detected, which results in another foveal requirement resulting in the inferred correspondence of that limb.

```
|next location selected as 270 832 peripheral
|after feature collection at 270 832 (125/250)
```

. . . . .

```
|node:limb-2
|  type           limb
|  description    component
|  bindings       ($5 b-connect-11)
|                 ($4 b-connect-10)
|                 ($3 upper-limb-3)
|                 ($2 lower-limb-3)
|                 ($1 extremity-3)
|  proximal-end   (336 . 804)
|  distal-end     (112 . 665)
|  location       (272 . 668)
|  delta          28.0912797857271
|  size           254.2618532143585
```

. . . . .

```
|located: limb-2 uninstantiated
|located: limb-1 as leg
|
|after next location selected as 96 672 foveal
|after feature collection at 96 672 (125/250)
```

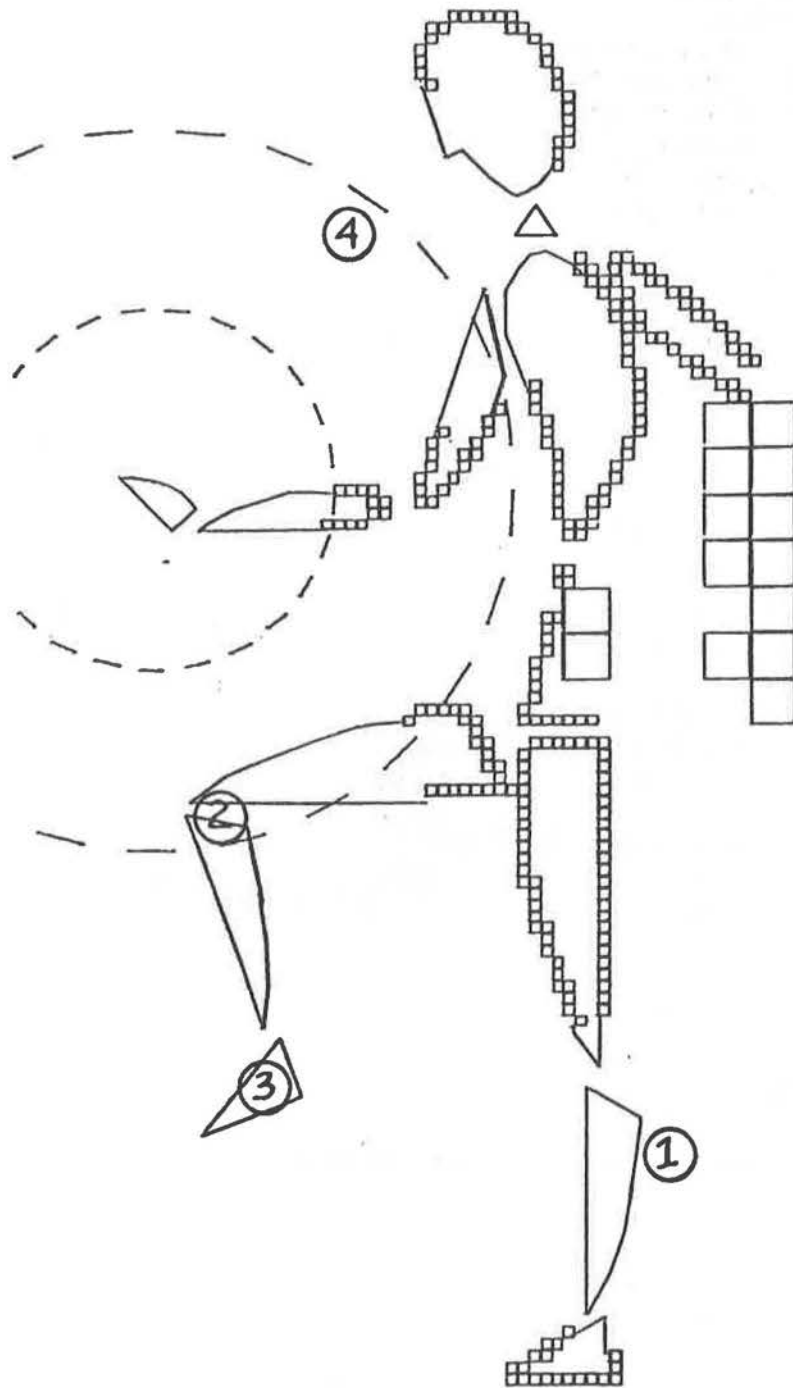


Figure 5.2.5. The fifth fixation at 96 672.

```

node:line-hand-2-1
  type      line-hand-2
  description component
  bindings  ($6 connect-26)
            ($5 connect-27)
            ($4 connect-25)
            ($3 line-33)
            ($2 line-34)
            ($1 line-35)
  size      58
  a2d       -136
  proximal-end (118 . 651)
  location    (95 . 666)
  distal-end  (72 . 681)

node:hand-1
  type      hand
  description image2
  bindings  (($1 line-hand-2-1))
  posture   open
  location  (95 . 666)
  proximal-end (118 . 651)
  distal-end  (72 . 681)
  size      58
  extra     (E00023 E00024)
            elaboration:E00023
            side      right
            a3d       (0 0 136)
            elaboration:E00024
            side      left
            a3d       (0 180 -136)

```

.....

```

located: limb-2 as arm
located: limb-1 as leg

```

```

next location selected as 448 832 peripheral

```



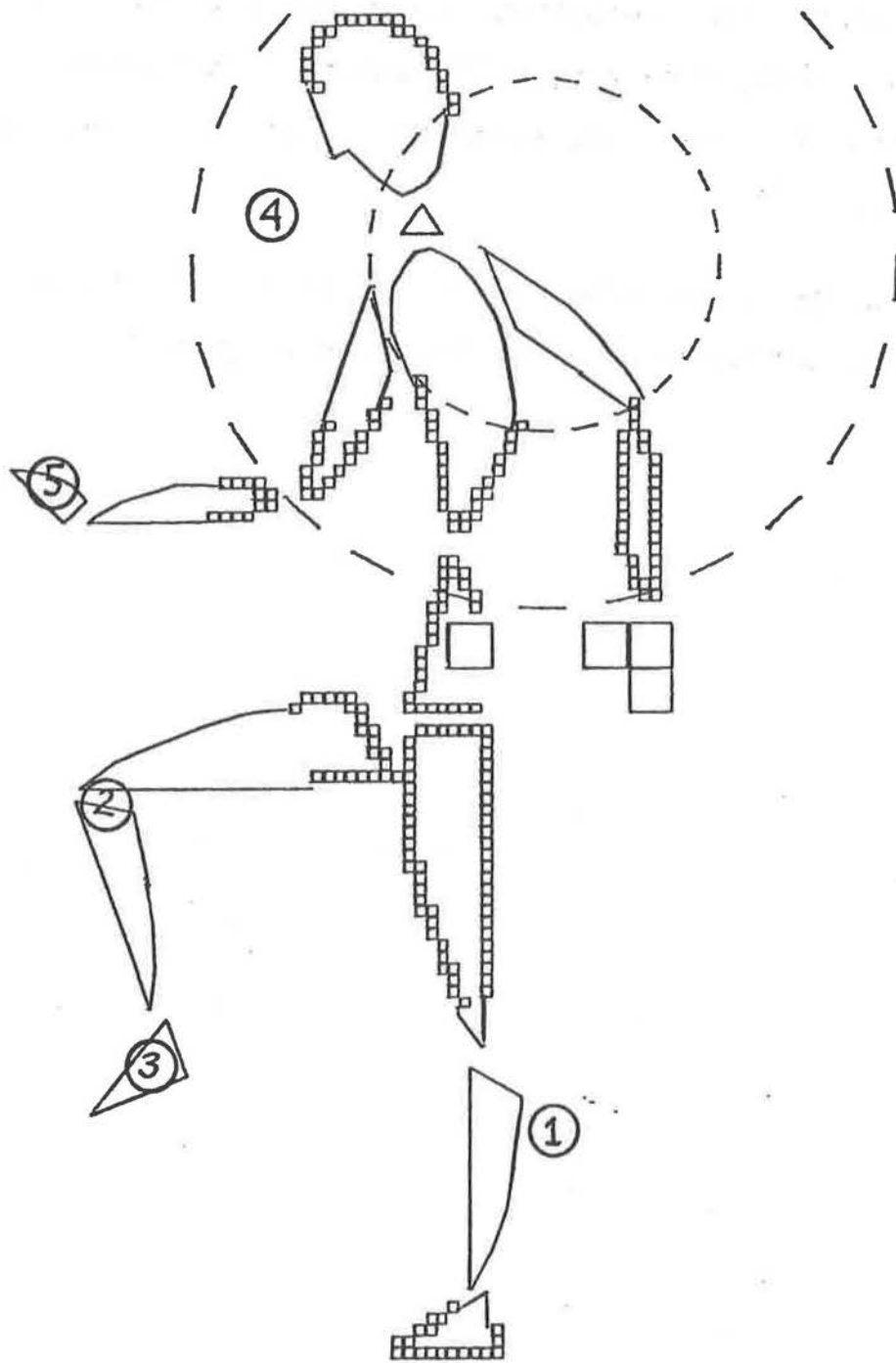


Figure 5.2.6. The sixth fixation at 448 832.

Of course, the preset sizes of the foveal and periperal diameters influence the resulting locations selected. The same image with different setting will result in different selections. Two of the four examples in Appendix D demonstrate this feature.

There are suggestions provided for extensions to these basic selection requirements in the concluding chapter 7.

## 6. Related Issues

### 6.1. Grouping and Feature Integration

Often human visual tasks may be couched in terms of the recognition of specific models expressed as the composition of more basic features. There are two steps involved in this recognition:

- (1) The identification of the necessary features.
- (2) The localization in space of the particular combinations of features comprising the models, and the examination of the way features interact to determine the validity of the models.

There is compelling evidence that the operations at these two steps are quite different, even though responses indicative of recognition may be based on either step. The Feature Integration Theory of attention (Treisman and Gelade, 1980) proposes that individual image features are detected rapidly and in parallel, but, in order that an object be identified as consisting of two or more separate features, locations must be processed serially with foveal attention. If focal attention is prevented, illusory perceptions will be formed through combining features incorrectly (Treisman and Schmidt, 1981).

There is an expense which accompanies the application of foveal attention. Treisman, Sykes, and Gelade (1977) have demonstrated this expense in the context of experiments

requiring subjects to detect a target in a display of distractor objects. The amount of time required to detect a target made up of a conjunction of features increases linearly with the number of distractors, but the detection of targets which are identifiable on the basis of features alone depends little on display size.

Consider an example taken from experiment IV of Treisman and Gelade (1980). In separate blocks, subjects were required to detect the letter "R" in a field of "P"s and "Q"s or in a field of "P"s and "B"s. The prediction was found to be correct: that detection time increased linearly with display size for the "PQ" distractors, and less for the "PB" distractors, even though it took longer for subjects to detect "R"s in a field of "B"s than in a field of "Q"s alone (see Figure 6.1.1).

P Q Q P Q	P B B P P
Q P P Q P	B P P B P
Q R P Q P	P P B R B
P Q P P Q	P B B P B
P Q Q Q P	B P B B P
(a)	(b)

Figure 6.1.1. Two displays of the type used in Feature Integration Theory experiments. (a) conjunction target R, (b) feature target R.

Figure 6.1.2 shows an analysis of these sets of letters in terms of the model possibilities associated with each

feature of the letters in the two different contexts.[27] The differences in reaction time can be explained as operations taking place on groupings of different spatial extent. In the displays with the "PB" distractors, the occurrence of the target may be determined within a grouping consisting of the entire display, simply by the presence of all the features necessary to make up the "R". If this criteria is met, the location of the single critical feature (the one with the single model possibility) can be used as a location to form a smaller grouping based on the letter alone to confirm the target's presence. In the case of the distractors "PQ" all the features that make up an "R" are always present, and there is no critical feature, so the tighter grouping must be applied to each letter in the display sequentially.

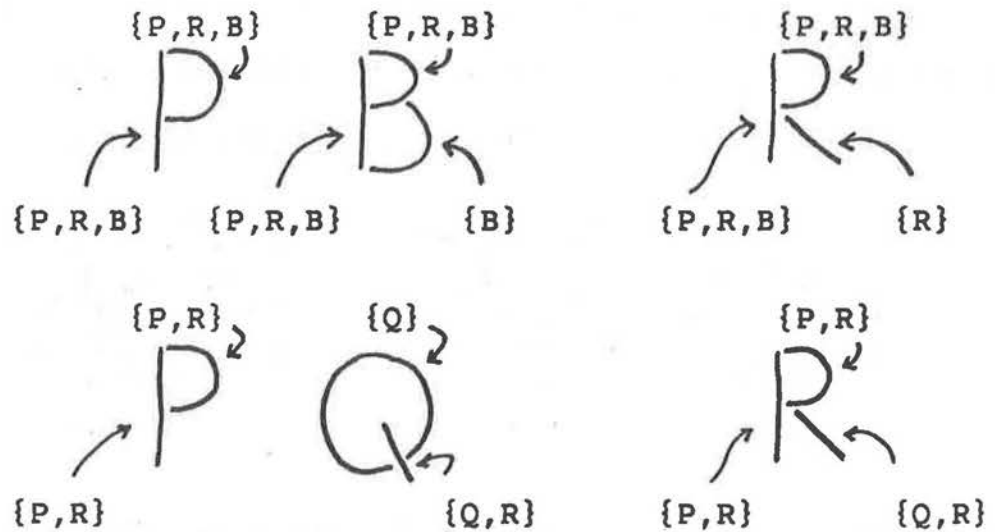


Figure 6.1.2. Analysis of display configurations shown in terms of model possibilities.

[27] A rather arbitrary set of features has been chosen for the letters in order to pursue the example.

These operations are similar in nature to those proposed as "grouping consistency" for the body drawing interpretation system. Each involves the formation of groups and subsequent use of the presence of features, and their model possibilities towards interpretation. For the feature integration experiments, recognition either succeeds or fails at each grouping operation. The proposed method used in the body drawing system considers many more possible models, and uses the grouping operation to move closer to interpretation by eliminating local model possibilities of the features.

Kahneman and Henik (1977) have formulated a "group-processing" model of the application of attention which is also similar to the application of grouping consistency. Their model proposes a pre-attentive grouping operation which selects large scale objects for subsequent analysis. The experiments which demonstrate the validity of this model employ displays such as that of figure 6.1.3.

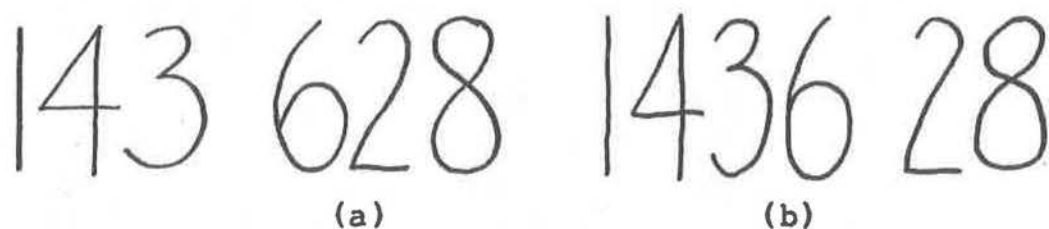


Figure 6.1.3. Group processing digit detection display.

One of the two displays such as shown in figure 6.1.3 is presented briefly and the task is to detect a specified target digit. The results show that groups are processed

separately, but that processing is almost uniform within groups.

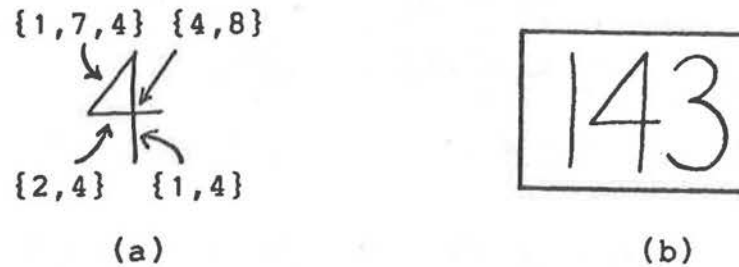


Figure 6.1.4. Features available at two resolutions.

Assume that high resolution feature information is available, and that for each such feature, a set of model possibilities is established (as shown in figure 6.1.4a). Also assume the availability of coarse level information which gives the identification of larger objects (figure 6.1.4b).

Consider the following interpretation of these results: In the first stage, the global objects are detected, as are the high resolution features specifying their model possibility sets. These sets can only be assigned, however, to the established objects, as depicted in figure 6.1.5.

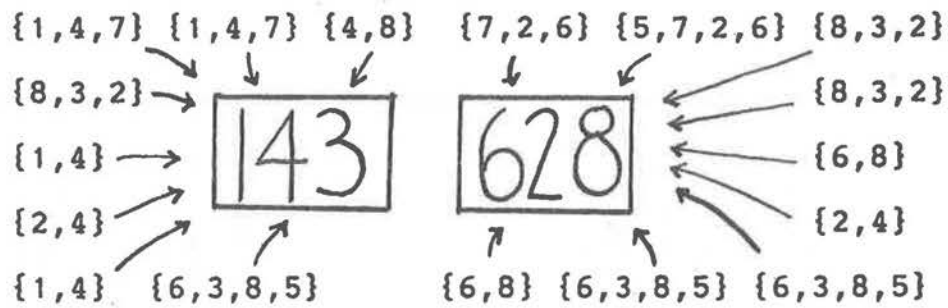


Figure 6.1.5. Low resolution objects detected and model possibilities assigned to high resolution features, which are roughly located.

At this point, there are obviously too many features associated with the object for it to be a single digit, so a subsequent breakdown of objects takes place. In that this second phase requires a higher resolution, it can only take place over a smaller area, so one of the two main objects is selected for more detailed examination (see figure 6.1.6).

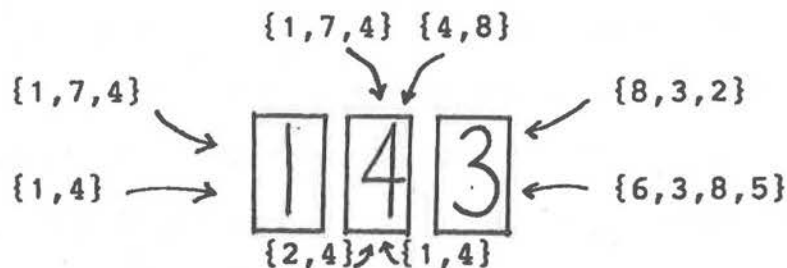


Figure 6.1.6. Features assigned to objects detected at a finer level of resolution, for one of the low level objects.

A second examination of the possibility sets reveals that the required elements are available for only one digit in each of the defined positions, and hence their identities can be established in parallel, without serial application of attention



to each of the specific locations.

Feature integration theory proposes that object identification may take place in parallel, based on features alone, or serially based on conjunctions of features when necessary. The group processing results indicate intermediate steps at which features are assigned to objects detected at low resolution, and once this assignment is complete, some model possibilities may be discarded by using grouping consistency. The location of objects to which these features are attached will become more refined if necessary, to the point of either allowing object identification through confirmation of the presence of the required elements alone, or if necessary by considering the relations among features.

The identity of features may be determined over a wide visual field, but without specific location. Location may become more specific through attachment to low resolution image elements, but only over a more restricted visual field. Finally, the actual location may be determined to permit feature integration. This final locating action operates over a small area of the visual field, and therefore requires serial application if more than one location is to be searched.[28]

---

[28] The relation between these aspects of Cognitive Psychology and the feature based interpretation methods was first expressed in Browse (1981).

## 6.2. Picture Grammars

The declarative schemata representation used for the body form problem domain has roots in some early computer vision research. In the 1960's, there emerged the requirement for structural descriptions rather than categorization of pictorial data. Grammatical structures became the object of investigation toward this end. If a class of pictorial objects could be represented as a grammar, and images could be parsed using that grammar, then the resultant parse tree or trace of the application of productions could provide a structural description of the image.

The fundamental problems in the application of grammatical methods to images are:

- (1) It is necessary to identify a set of image primitives which can act as terminal symbols of the grammar.
- (2) It is necessary to develop a means of specifying and using the complex relations which exist among image elements. Grammars for languages utilize the implicit and uniform relation between symbols of the grammar, which is simply their ordered sequence of appearance.

Both of these issues were addressed in early picture grammar systems. Ledley's (1964) system traced around the perimeter of objects detected in an image. Local characteristics of the line segments found during the trace were used as

the basis for the development of a set of primitives to be used as the terminal symbols of a grammar (straight line, clockwise curve, etc.). This tracing also provided a strict ordering of grammatical structures, which allowed the direct use of the implicit concatenation relation as used in language based grammatical applications.

Shaw (1969) developed a picture description language (PDL) which also used line segments as terminal symbols. An inventory of connecting relations was given which were used to construct descriptions of objects. These connecting relations were allowed, not only among the terminal symbols, but also among larger scale picture objects. These relations were all connectivity relations, requiring every picture part, or terminal symbol to designate a "head" and "tail" part, in order that the relations could be specified.

Evans (1969) recognized the advantages of using more sophisticated relationships in the specification of objects. For example:

```
(TRI (X Y Z) ((PT X) (PT Y) (PT Z)
              (ELS X Y) (ELS X Z) (ELS Y Z)
              (NONCOLL X Y Z)))
```

This describes a triangle as three points. The "ELS" predicate requires that there exists a line segment between the points specified, and "NONCOLL" requires that the points specified as its arguments are not collinear.

Stanton (1972) identified this progression towards more relational information in picture grammars. The purpose of such representations was to make explicit a class of objects which might be depicted in the image. Stanton recognized that no clear specification of the relations and predicates was being provided, and that the modularity and generality of the resulting systems was diminishing.

Stanton (1970) devised a system, RAMOS which combines the descriptive structures of picture grammars with a set of primitive operations over a data base. Relations among image objects were expressed as combinations of the primitive operations in a way which anticipated the use of logic programming systems such as PROLOG in the analysis of visual information.

Consider the example from Stanton (1972):

```
F( ... A:SQUARE B:TRI JOIN(A,B) ...)
```

This is a description for a situation in which "a square is joined to a triangle". The "F" operator indicates the instruction to find the situation, and the requirements for the situation are given within the parentheses. The predicate "JOIN" is given in terms of the primitive operations:

```
JOIN G(A,B) F(C.SIDE(A) D.SIDE(B) SAME(C,D))
```

This means that given A and B, find a side of A (call it C), and find a side of B (call it D) such that C and D are the

same.

The equivalent PROLOG description would be:

```
square-joined-to-triangle(*A,*B) <-
    square(*A) & triangle(*B) & join(*A,*B).

join(*A,*B) <-
    side-of(*A,*C) & side-of(*B,*C).
```

The use of the "same" predicate is not necessary in the PROLOG version because of the required unification on the variable "\*C".

Stanton's requirement for the clear depiction of the structure of relations was taken as an indication of the need for a procedural component, though within the context of declarative systems such as PROLOG, this is not the only option. Stanton also identified the need for an ability to invert the predicates required among image structures to enable not only testing, but obtaining relations from an image. This remains an open problem which is independent of whether declarative or procedural methods are employed.

The declarative schemata system used for the body form information as described in this document may be viewed as an extension to the concept of a picture grammar. These extensions are identified and described below:

Logic programming connection: The declarative schemata system represents knowledge in a way which can easily be translated

to logic programming systems (see section 4.2). This gives a direction for the further exploration of the issues of incorporating structural descriptions of relations and predicates as proposed by Stanton.

Multi-layer grammars: Parsing processes are inherent in vision systems which use component hierarchies as the basis for the structural descriptions of objects. Havens (1978) makes an argument for the similarity of such systems. The declarative schemata system provides a grammatical counterpart for the specialization hierarchy through the incorporation of connections between the layers of the grammar whose basis is at different resolution levels.

Cue/model approach: Most of the research relating to picture grammars preceded the development of the idea of labelling image elements with interpretation possibilities (Huffman, 1971; Clowes, 1971). The declarative schemata system makes an explicit connection between these approaches through the provision of the ability to analyze the grammatical structure and automatically generate the cue/model structure necessary for the application of interpretation labelling methods.

Attribute structure: The declarative schemata provide a means of specifying attributes of the non-terminals of the grammar. These specification methods are associated with each schema in much the same way that attribute evaluation methods are associated with productions in attribute grammars (Knuth, 1968).

The main difference is, that in the use of attribute grammars, a context-free parse tree is first generated, and then afterwards, the attribute values are developed. In the declarative schemata system, the attributes are evaluated immediately as the schema is applied[29]. The results (the attribute values) then enter into the decision as to the applicability of subsequent schemata by their involvement in required relations. This attribute structure thus provides a context-sensitivity mechanism as well as a uniform means of developing the semantics associated with parsing an image.

Access to three dimensions: One problem with picture grammars was that they were only applicable to two dimensional problem domains. Picture grammars provided no means of structuring the depictions such that coherent three-dimensional objects could be represented (see Stanton, 1972). The declarative schemata system enables the mapping from view-oriented descriptions of objects to underlying representations of the three-dimensional aspects of objects.

---

[29] This sometimes results in multiple value possibilities.

### 6.3. View Based Representation

The body form interpretation system represents problem domain knowledge in a way which is a hybrid of two and three-dimensional structures. The underlying three-dimensional representation of the human body gives possible ranges of orientation and relative length of the body segments. This formulation does not express the shape of the objects, and so it is not appropriate for matching operations such as those used by Marr and Nishihara (1976). As an alternative, specific prototypical views are given with mappings between these image domain structures and the underlying model. Constraints from both representations are used in the development of an interpretation for the image.

A related proposal is found in Minsky's (1975) frame system for the representation of knowledge. In its application to visual scene analysis, different viewpoints are represented separately, with transformations provided which take one such frame to the next, thereby encoding the effect of perspective change. Minsky also argues that the idea of dimensionality in a representation is not completely appropriate in the discussion of propositional systems. This notion is borne out by the body form representation system. The same basic propositional format is used to encode the three-dimensional requirements for the body parts as is used to describe the two-dimensional relations required among elements in the image constructions.



Pinker (1980; Pinker and Finke, 1980) has proposed a model for the representation of physical space which uses both underlying three-dimensional information and specific perspectives. Subjects were shown to be able to develop mental images of objects from angles which they had not experienced. Scanning and line-of-sight tasks indicated that subjects were utilizing emergent two-dimensional aspects of their images.

The notion of a three-dimensional model for object knowledge which can be arbitrarily rotated, and from which perspectives may be generated is an appealing idea. Caution must be taken, however, not to usurp an understanding of the visual processes being explained by assuming the ability to "look at" this internal model (see Pylyshyn, 1973; Minsky, 1975).

Pinker's model emphasizes the primacy of the three-dimensional structures. Other studies indicate that, for known objects, there exist special privileged perspectives, suggesting that particular views are not necessarily constructed from the three-dimensional model, but may have an "existence" of their own (Palmer, Rosch, and Chase, 1981). As an informal indication of this idea, consider that it would be more difficult to recognize an elephant from a photograph taken from above than from a photograph taken from the side. Yet the case would be reversed for an ant. It would be more difficult to recognize the ant from the side. This is because the familiarity of particular perspectives has a role in the

structure of visual knowledge. Such differences are difficult to explain in a system which only projects perspectives from a three-dimensional structure. In their experiments, Palmer Rosch and Chase established canonical views using a number of converging measures: goodness of view, selected angle to take a photograph, and imagined viewpoint. Subjects were shown to be faster in identifying photographs of these canonical views. The results were the same in a condition in which the subjects were told ahead of time what the viewpoint was to be.

The notion of canonical concepts in semantic memory has been demonstrated (Rosch, 1975; Mervis and Rosch, 1981). Particularly familiar concepts such as "dog" appear to have a special status, forming a base for both generalizations ("animal"), and specializations ("collie"). This idea that "the familiar" forms the basis of knowledge structures carries into the realm of visual knowledge in the notion of canonical views.

The structures which were devised for the knowledge about the body form line drawing problem domain are intended as a step towards solving the problems involved in maintaining both two and three-dimensional representations explicitly.

## 7. Conclusions

This document describes a computational vision system which interprets a set of line drawings of human body forms. The system was devised in response to two research goals:

- (1) To develop declarative structures for the representation of the knowledge about specific objects, as required for visual interpretation, and to separate the process of interpretation from this knowledge.
- (2) To incorporate fundamental aspects of human vision into a computational system. Specifically, to enable interpretation based interaction among levels of resolution, and to provide a means of intelligent selection of processing locations.

The first step in the development of the computer system was to devise a declarative schemata format for the representation of visual knowledge. This format is similar to, and extends picture grammars in a number of ways. The basic structure of the body form knowledge follows the component hierarchy for the domain. Two types of features are known to the system, each of which forms the basis for a separate representation layer. These layers are interconnected by links indicative of the specialization/generalization hierarchy. The declarative schemata system provides prototypical perspective information, with explicit mappings into an underlying three-dimensional model.

The second step accomplishes an analysis of the declarative schemata contents. This analysis generates an extensive cue table associating each generic object with a list of interpretation possibilities, or roles that they might play in some larger structure. These roles are conditional on the values of attributes of the objects. This cue table is devised to permit a set labelling mechanism for the maintenance of model possibilities, which permits the effective use of partial information about image features. This phase also relaxes the conditions on the prototypical view representations so that they cover a wider class of depictions.

During the interpretation process, features are available in concentric areas of limited diameter fovea and periphery. At each fixation, feature based grouping and consistency provide a means of pruning the lists of interpretation possibilities associated with each feature. This constitutes the first phase of interpretation based interaction between levels of detail.

A strictly bottom-up method for invoking the examination of specific schemata is employed, which allows systematic control of partially fulfilled schemata instances through the application of incremental consistency. The systems maintains provision for multiple contexts by carrying several possible values only for those attributes which are affected by context.

An analysis of the results of the invocation of the schemata yields criteria for the intelligent selection of processing location. The notion of a correspondence between interpretations based on different levels of resolution is introduced. This is extended to the idea of inferred correspondence. The use of inferred correspondence of coarse layer interpretation as a goal of the system enables the propagation of detailed foveal based results into the peripheral area, removing the requirement for fine layer processing of the entire image.

Throughout the description of the system, attempts have been made to furnish the details of relations that operations might have to Cognitive Psychology research.

One of the main advantages of the separation of object knowledge and process is that it facilitates transferral to other problem domains, and permits experimentation with other interpretation techniques. Some of the directions that these extensions might take are listed below:

(1) An interesting extension would be to devise a representation for the digits and the letters of the alphabet. The operation of the resulting system might then be aligned with the experimental results of Psychology research in visual attention, which often utilize letters and digits as display items. Having a representation for letters might also allow computational based studies of the selectional processes in

reading, and the examination of the role of peripheral information (see Rayner, 1978).

(2) Other problem domains may require enhancements of the declarative schemata format. For the level of description used for the body form knowledge, there is always a fixed number of components for each object (ie., two arms, not three). In other domains objects may be specified with unknown numbers of components. For example, in the sketch map domain (Mackworth, 1977b), a mountain range is composed of an arbitrary number of mountain symbols. Extensions to other domains will be necessary to develop a robust and generally useful representational tool.

(3) The body drawing interpretation system is entirely "bottom-up" in its operation. "Top-down" components might be instituted in several ways. "Top-down" expectations could be formulated using prior knowledge of the expected position of the body in the image. Currently, the model invocation at the coarse and fine layer do not interact. It would be possible to use a global to local "top-down" component by ordering the considerations at the fine layer on the basis of the results from the coarse layer. Preliminary interpretation, based on an even more coarse level of resolution could be used to develop a preliminary context for the processing at any location (see Palmer, 1977). A third type of "top-down" control could be introduced as hypothesis testing. Once a schema is partially fulfilled, it could direct processing towards the

discovery of its remaining requirements.

Another interesting direction would be the further examination of the use of logic programming in visual interpretation systems. The declarative schemata system maps well into such representations, but easier access to control structure would be required in order to allow the experimentation in the processes of interpretation.

The body drawing interpretation system uses very simple criteria for selecting processing locations. These criteria were intended only as an example of what might be accomplished. The schemata themselves could be examined in terms of expected areas of related objects, even without the support of coarse layer results. The nature of the task involved in vision is an obvious candidate as a determining factor in location selection.

## References

- Alpern, M.  
1972, "Eye Movements," in Handbook of Sensory Physiology Vol VII/4: Visual Psychophysics, D. Jameson and L.M. Hurvich (eds.), Springer, Berlin.
- Antes, J.R.  
1974, "The Time Course of Picture Viewing," Journal of Experimental Psychology 103, 62-70.
- American Academy of Orthopedic Surgeons  
1965, Joint Motion: Method of Measuring and Recording, Chicago.
- Bajcsy, R. and Rosenthal, D.A.  
1980, "Visual and Conceptual Focus of Attention," in Structured Computer Vision, S. Tanimoto and A. Klinger (eds.), Academic Press, New York, 133-149.
- Barrow, H.G. and Tenenbaum, J.M.  
1978, "Recovering Intrinsic Scene Characteristics From Images," in Computer Vision Systems, A.R. Hanson and E.M. Riseman (eds.), Academic Press, New York, 3-26.
- Barrow, H.G. and Tenenbaum, J.M.  
1981, "Computational Vision," Proc. IEEE 69, 572-595.
- Bartlett, F.C.  
1932, Remembering, A Study in Experimental and Social Psychology, Cambridge University Press, Cambridge.
- Biederman, I.  
1981, "On the Semantics of a Glance at a Scene," in Perceptual Organization, M. Kubovy and J.R. Pomerantz (eds.), Erlbaum, Hillsdale, New Jersey, 213-253.
- Blakemore, C. and Campbell, F.W.  
1969, "On the Existence of Neurons in the Human Visual System Selectively Sensitive to the Orientation and Size of Retinal Images," Journal of Physiology 203, 237-260.
- Brady, M.  
1982, "Computational Approaches to Image Understanding," ACM Computing Surveys 14(1), 3-72.
- Breitmeyer, B. and Ganz, L.  
1976, "Implications of Sustained and Transient Channels for Theories of Visual Pattern Masking, Saccadic Suppression, and Information Processing," Psychological Review 83, 1-36.



- Brooks, R.A.  
1981, "Model-Based Three Dimensional Interpretations of Two Dimensional Images," Proceedings of the Seventh International Joint Conference on Artificial Intelligence, Vancouver, 619-624.
- Browse, R.A.  
1980, "Mediation Between Central and Peripheral Processes: Useful Knowledge Structures," Proc. Third Conf. of the Canadian Society for the Computational Studies of Intelligence, Victoria, Canada, 166-171.
- Browse, R.A.  
1981, "Relations Between Schemata-Based Computational Vision and Aspects of Visual Attention," Proc. Fourth Annual Conference of the Cognitive Science Society, Berkeley.
- Browse, R.A.  
1982, "Interpretation-Based Interaction Between Levels of Detail," Proc. Fourth Conf. of the Canadian Society for the Computational Studies of Intelligence, Saskatoon, Canada, 27-32.
- Bruner, J.S. and Potter, M.C.  
1964, "Interference in Visual Recognition," Science 144, 424-425.
- Buswell, G.T.  
1935, How People Look At Pictures, University of Chicago Press.
- Catanzariti, E. and Mackworth, A.K.  
1978, "Forests and Pyramids: Using Image Hierarchies to Understand Landsat Images," Proc. 5th Canadian Symposium on Remote Sensing.
- Chomsky, N.  
1957, Syntactic Structures, Mouton and Co., The Hague.
- Chomsky, N.  
1965, Aspects of the Theory of Syntax, The M.I.T. Press, Cambridge.
- Clowes, M.B.  
1971, "On Seeing Things," Artificial Intelligence 2(1), 79-112.
- Didday, R.C. and Arbib, M.A.  
1975, "Eye Movements and Visual Perception: A Two Visual System Model," International Journal of Man-Machine Studies 7, 547-569.

- Eriksen, C.W. and Hoffman, J.E.  
1972, "Temporal and Spatial Characteristics of Selective Encoding from Visual Displays," Perception and Psychophysics 12, 201-204.
- Eshkol, N. and Wachmann, A.  
1958, Movement Notation, Arrowsmith, Bristol.
- Evans, T.G.  
1969, "Descriptive Pattern Analysis Techniques," in Automatic Interpretation and Classification of Images, A. Grasselli (ed.), NATO Advanced Study Institute, Pisa, 79-96.
- Fahlman, S.E.  
1979, NETL: A System for Representing and Using Real-World Knowledge, M.I.T. Press, Cambridge.
- Farley, A.M.  
1976, "A Computer Implementation of Constructive Visual Imagery and Perception," in Eye Movements and Psychological Processes, R.A. Monty and J.W. Senders (eds.), Halstead Press, New York, 473-490.
- Fisher, D.F., Monty, R.A., and Senders, J.W.  
1981, Eye Movements: Cognition and Visual Perception, Erlbaum, Hillsdale, New Jersey.
- Friedman, A.  
1979, "Framing Pictures: The Role of Knowledge in Automated Encoding and Memory for Gist," Journal of Experimental Psychology: General 108(3), 316-355.
- Funt, B.V.  
1976, "Whisper: a Computer Implementation Using Analogues in Reasoning," TR-76-09, Computer Science Dept., University of British Columbia, Vancouver, Canada.
- Garner, W.R.  
1974, The Processing of Information and Structure, Erlbaum, Pontomac, Maryland.
- Gilchrist, A.L.  
1977, "Perceived Lightness Depends on Perceived Spatial Arrangement," Science 195, 186-187.
- Gilchrist, A.L.  
1980, "When Does Perceived Lightness Depend on Perceived Spatial Arrangement?," Perception and Psychophysics 28, 527-538.

- Gould, J.D.  
1976, "Looking At Pictures," in Eye Movements and Psychological Processes, R.A. Monty and J.W. Senders (eds.), Halstead Press, New York, 323-345.
- Gould, J.D. and Schaffer, A.  
1965, "Eye Movement Patterns During Visual Information Processing," Psychonomic Science 3, 317-318.
- Graham, N.  
1981, "Psychophysics of Spatial-Frequency Channels," in Perceptual Organization, M. Kubovy and J.R. Pomerantz (eds.), Erlbaum, Hillsdale, New Jersey, 1-25.
- Gregory, R.L.  
1966, Eye and Brain, McGraw-Hill, New York.
- Guzman, A.  
1968, "Decomposition of a Visual Scene Into Three-Dimensional Bodies," Proc. AFIPS 1968 Fall Joint Computer Conference, 291-304.
- Haber, R.N.  
1978, "Visual Perception," in Annual Review of Psychology, M.R. Rosenzweig and L.W. Porter (eds.), Erlbaum, Hillsdale, New Jersey.
- Hanson, A.R. and Riseman, E.M.  
1975, "The Design of a Semantically Directed Vision Processor," Technical Report No. 75C-1, University of Massachusetts, Amherst, Massachusetts.
- Hanson, A.R. and Riseman, E.M.  
1978, "Segmentation of Natural Scenes," in Computer Vision Systems, A.R. Hanson and E.M. Riseman (eds.), Academic Press, New York, 129-164.
- Hanson, A.R. and Riseman, E.M.  
1980, "Processing Cones: A Computational Structure for Image Analysis," in Structured Computer Vision, S. Tanimoto and A. Klinger (eds.), Academic Press, New York, 101-131.
- Havens, W.S.  
1976, "Can Frames Solve the Chicken and Egg Problem?," Proc. First Conf. of the Canadian Society for the Computational Studies of Intelligence, Vancouver, Canada, 232-242.
- Havens, W.S.  
1978, "A Procedural Model of Recognition for Machine Perception," TR-78-3, Computer Science Dept., University of British Columbia, Vancouver, Canada.

- Hinton, G.F.  
1981, "Shape Representation in Parallel Systems," Proceedings of the Seventh International Joint Conference on Artificial Intelligence, Vancouver, 1088-1096.
- Hochberg, J.  
1968, "In the Mind's Eye," in Contemporary Theory and Research in Visual Perception, R.N. Haber (ed.), Holt, Rinehart, and Winston, New York, 309-321.
- Hochberg, J.E. and Brooks, V.  
1978, "Film Cutting and Visual Momentum," in Eye Movements and the Higher Psychological Functions, J.W. Senders, D.F. Fisher, and R.A. Monty (eds.), Erlbaum, Hillsdale, New Jersey.
- Hoffman, J.E.  
1980, "Interaction Between Global and Local Levels of a Form," Journal of Experimental Psychology: Human Perception and Performance 6(2), 222-234.
- Horn, B.K.P.  
1975, "Obtaining Shape From Shading Information," in The Psychology of Computer Vision, P.H. Winston (ed.), McGraw-Hill, New York, 115-155.
- Hubel, D.H. and Weisel, T.N.  
1979, "Brain Mechanisms of Vision," Scientific American 241(3), 150-162.
- Huffman, D.A.  
1971, "Impossible Objects as Nonsense Sentences," in Machine Intelligence 6, B. Meltzer and D. Michie (eds.), American Elsevier, New York, 295-323.
- Just, M.A. and Carpenter, P.A.  
1978, "Inference Processes During Reading: Reflections from Eye Fixations," in Eye Movements and the Higher Psychological Functions, J.W. Senders, D.F. Fisher, and R.A. Monty (eds.), Erlbaum, Hillsdale, New Jersey.
- Kahneman, D.  
1973, Attention and Effort, Prentice Hall, Englewood Cliffs, New Jersey.
- Kahneman, D. and Henick, A.  
1977, "Effects of Visual Grouping on Immediate Recall and Selective Attention," in Attention and Performance VI, S. Dornic (eds.), Erlbaum, Hillsdale, New Jersey, 307-332.

- Kahneman, D. and Treisman, A.  
1983, "Changing Views of Attention and Automaticity," in Varieties of Attention, R. Parasuraman, R. Davies, and J. Beatty (eds.), Academic Press, New York.
- Kelly, M.D.  
1971, "Edge Detection in Pictures by Computer Using Planning," in Machine Intelligence 6, B. Meltzer and D. Michie (eds.), American Elsevier, New York, 397-409.
- Kinchla, R.  
1974, "Detecting Target Elements in Multielement Arrays: A Confusability Model," Perception and Psychophysics 15, 149-158.
- Kinchla, R. and Wolfe, J.  
1979, "The Order of Visual Processing: "Top-Down", "bottom-Up", or "Middle-Out", " Perception and Psychophysics 25, 225-231.
- Klein, G.A. and Kurkowski, F.  
1974, "Effect of Task Demand on the Relationship Between Eye Movement and Sentence Complexity," Perceptual and Motor Skills 39, 463-466.
- Knuth, D.E.  
1968, "Semantics of Context-Free Languages," Mathematical Systems Theory 2(2), 127-146.
- Laberge, D.  
1976, "Perceptual Learning and Attention," in Handbook of Learning and Cognitive Processes, Vol 4, W.K. Estes (ed.), Erlbaum, Hillsdale, New Jersey.
- Latour, P.L.  
1962, "Visual Threshold During Eye Movements," Vision Research 2, 261-262.
- Ledley, R.S.  
1964, "High-Speed Automatic Analysis of Biomedical Pictures," Science 146(9), 216-223.
- Levine, M.D.  
1980, "Region Analysis Using a Pyramid Data Structure," in Structured Computer Vision, S. Tanimoto and A. Klinger (eds.), Academic Press, New York, 57-100.
- Lockhead, G.R.  
1972, "Processing Dimensional Stimuli: A Note," Psychological Review 79, 410-419.

- Loftus, G.R.  
1972, "Eye Fixations and Recognition Memory for Pictures," Cognitive Psychology 3, 525-551.
- Loftus, G.R. and Mackworth, N.H.  
1978, "Cognitive Determinants of Fixation Location During Picture Viewing," Journal of Experimental Psychology: Human Perception and Performance 4, 565-572.
- Lowe, D.  
1975, "Processing Information About Location in Brief Visual Displays," Perception and Psychophysics 18, 309-316.
- Leushina, L.I.  
1965, "On Estimation of Position of Photostimulus and Eye Movement," Biofizika 10, 130-136.
- Mackworth, A.K.  
1973, "Interpreting Pictures of Polyhedral Scenes," Artificial Intelligence 4(2), 121-137.
- Mackworth, A.K.  
1975, "Model-Driven Interpretation in Intelligent Vision Systems," Perception 5, 349-370.
- Mackworth, A.K.  
1977a, "How to See a Simple World: an Exegesis of Some Computer Programs for Scene Analysis," in Machine Intelligence 8, E.W. Elcock and D. Michie (eds.), John Wiley, New York, 510-540.
- Mackworth, A.K.  
1977b, "On Reading Sketch Maps," Proceeding of the Fifth International Joint Conference on Artificial Intelligence, Cambridge, 598-606.
- Mackworth, A.K.  
1977c, "Consistency in Networks of Relations," Artificial Intelligence 8(1), 99-118.
- Mackworth, A.K.  
1978, "Vision Research Strategy: Black Magic, Metaphors, Mechanisms, Miniworlds and Maps," in Computer Vision Systems, A.R. Hanson and E.M. Riseman (eds.), Academic Press, New York, 53-61.
- Mackworth, A.K. and Havens, W.S.  
1981, "Structuring Domain Knowledge for Visual Perception," Proceedings of the Seventh International Joint Conference on Artificial Intelligence, Vancouver, 625-627.



- Mackworth, N.H. and Morandi, A.J.  
1967, "The Gaze Selects Informative Details Within Pictures," Perception and Psychophysics 2, 547-552.
- Marcotty, M., Ledgard, H.F., and Bochmann, G.V.  
1976, "A sampler of Formal Definitions," ACM Computer Surveys 8(2), 191-276.
- Marr, D.  
1976, "Early Processing of Visual Information," Phil. Trans. Royal Society of London 275B(942), 483-524.
- Marr, D.  
1982, Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, W.H. Freeman, San Francisco.
- Marr, D. and Hildreth, E.  
1980, "Theory of Edge Detection," Proc. Royal Soc. London B(207), 187-217.
- Marr, D. and Nishihara, H.K.  
1976, "Representation of the Spatial Organization of Three Dimensional Shapes," Report 377, A.I. Lab, M.I.T.
- Martin, M.  
1979, "Local and Global Processing: The Role of Sparsity," Memory and Cognition 7, 476-484.
- Mervis, C.B. and Rosch, E.  
1981, "Categorization of Natural Objects," in Annual Review of Psychology, M.R. Rosenzweig and L.W. Porter (eds.).
- Miller, J.  
1981, "Global Precedence in Attention and Decision," Journal of Experimental Psychology: Human Perception and Performance 7, 1161-1174.
- Minsky, M.  
1975, "A Framework for Representing Knowledge," in The Psychology of Computer Vision, P.H. Winston (ed.), McGraw-Hill, New York.
- Monty, R.A. and Senders, J.W.  
1976, Eye Movements and Psychological Processes, Halstead Press, New York.
- Navon, D.  
1977, "Forest Before Trees: The Precedence of Global Features in Visual Perception," Cognitive Psychology 9, 353-383.

- Neisser, U.  
1976, Cognition and Reality, Freeman, San Francisco.
- Neisser, U.  
1967, Cognitive Psychology, Appleton, Century, Crofts, New York.
- Newell, A.  
1973, "Production Systems: Models of Control Structures," in Visual Information Processing, W.G. Chase (ed.), Academic Press, New York.
- Nishihara, H.K.  
1981, "Intensity, Visible-Surface, and Volumetric Representations," Artificial Intelligence 17, 265-284.
- Noton, D. and L. Stark,  
1971a, "Eye Movements and Visual Perception," Scientific American 224, 34-43.
- Noton, D. and Stark, L.  
1971b, "Scanpaths in Eye Movements During Pattern Perception," Science 171, 308-311.
- Noton, D. and Stark, L.  
1971c, "Scanpaths in Saccadic Eye Movements Whilst Viewing and Recognizing Patterns," Vision Research 11, 929-942.
- Oshima, M. and Shirai, Y.  
1981, "Object Recognition Using Three-Dimensional Information," Proceedings of the Seventh International Joint Conference on Artificial Intelligence, Vancouver, 601-606.
- Palmer, S.E.  
1975, "Visual Perception and World Knowledge: Notes on a Model of Sensory-Cognitive Interaction," in Explorations in Cognition, D.A. Norman and D.E. Rumelhart (eds.), Freeman, San Francisco, 279-307.
- Palmer, S.E.  
1977, "Hierarchical Structure in Perceptual Representation," Cognitive Psychology 9, 441-474.
- Palmer, S.E., Rosch, E., and Chase, P.  
1981, "Canonical Perspective and the Perception of Objects," in Attention and Performance IX, J. Long and A. Baddeley (eds.), Erlbaum, Hillsdale, New Jersey.
- Pantle, A. and Sekuler, R.  
1968, "Velocity Sensitive Mechanisms in Human Vision," Vision Research 8, 445-450.



- Parker, R.E.  
1978, "Picture Processing During Recognition," Journal of Experimental Psychology: Human Perception and Performance 4, 284-293.
- Parks, T.  
1965, "Post-Retinal Visual Storage," American Journal of Psychology 78, 145-147.
- Piaget, J.  
1967, Biology and Knowledge, Gallimard Press, Paris.
- Pinker, S.  
1980, "Mental Imagery and the Third Dimension," Journal of Experimental Psychology: General 109(3), 354-371.
- Pinker, S. and Finke, R.A.  
1980, "Emergent Two-Dimensional Patterns in Images Rotated in Depth," Journal of Experimental Psychology: Human Perception and Performance 6(2), 244-64.
- Posner, M.I.  
1978, Chronometric Exploration of Mind, Erlbaum, Hillsdale, New Jersey.
- Pylyshyn, Z.W.  
1973, "What the Mind's Eye Tells the Mind's Brain: a Critique of Mental Images," Psychology Bulletin 80(1), 1-24.
- Pylyshyn, Z., Elcock, E.W., Marmor, M., and Sander, P.  
1978a, "Exploration in Visual-Motor Space," Proc. Second Conf. of the Canadian Society for the Computational Studies of Intelligence, Toronto, Canada
- Pylyshyn, Z., Elcock, E.W., Marmor, M.M., and Sander, P.T.  
1978b, A System for Perceptual-Motor Based Reasoning, Report 42, Department of Computer Science, U. of Western Ontario, London, Canada.
- Rayner, K.  
1975, "The Perceptual Span and Peripheral Cues in Reading," Cognitive Psychology 7, 65-81.
- Rayner, K.  
1978, "Eye Movements in Reading and Information Processing," Psychological Bulletin 85(3), 618-660.
- Rayner, K., McConkie, G.W., and Erlich, S.  
1978, "Eye Movements and Integrating Information Across Fixations," Journal of Experimental Psychology: Human Perception and Performance 4(4), 529-544.

- Riggs, L.A.  
1965, "Visual Acuity," in Vision and Visual Perception, C.H. Graham (ed.), Wiley, New York.
- Riggs, L.A.  
1973, "Curvature as a Feature of Pattern Vision," Science 181, 1070-1072.
- Roberts, L.G.  
1965, "Machine Perception of Three Dimensional Solids," in Optical and Electro-optical Information Processing, J.T. Tippett, D. Berkowitz, L. Clapp, C. Koester, and A. Vanderburgh (eds.), M.I.T. Press, Cambridge, Massachusetts, 159-197.
- Rosch, E.  
1975, "Cognitive Representations of Semantic Categories," Journal of Experimental Psychology: General 104, 193-233.
- Rosenthal, D. and Bajcsy, R.  
1978, "Conceptual and Visual Focussing in the Recognition Process as Induced by Queries," Proc. Fourth International Joint Conference on Pattern Recognition, Kyoto, 417-420.
- Roy, R. and Sutro, L.L.  
1982, "Simulation of Two Forms of Eye Motion and Its Possible Implications for the Automatic Recognition of Three-Dimensional Objects," IEEE Transactions on Systems, Man, and Cybernetics SMC-12(3), 276-288.
- Senders, J.W., Fisher, D.F., and Monty, R.A.  
1978, Eye Movements and the Higher Psychological Functions, Erlbaum, Hillsdale, New Jersey.
- Shaw, A.C.  
1969, "A Formal Picture Description Scheme as a Basis for Picture Processing Systems," Information and Control 14(1), 9-52.
- Shiffrin, R.M. and Schneider, W.  
1977, "Controlled and Automatic Human Information Processing: II. Perceptual Learning, Automatic Attending and General Theory," Psychological Review 84, 127-190.
- Shirai, Y.  
1975, "Analyzing Intensity Arrays Using Knowledge About Scenes," in The Psychology of Computer Vision, P.H. Winston (ed.), McGraw-Hill, New York, 93-113.

- Stanton, R.B.  
1970, "Computer Graphics - the Recovery of Descriptions in Graphical Communication," Phd Thesis, Electronic Computation, University of New South Wales.
- Stanton, R.B.  
1972, "The Interpretation of Graphics and Graphic Languages," in Graphic Languages, F. Nack and A. Rosenfeld (eds.), North-Holland Publishing Co., Amsterdam, 144-159.
- Stevens, K. A.  
1981, "The Visual Interpretation of Surface Contours," Artificial Intelligence 17, 47-75.
- Tanimoto, S.L.  
1976, "Pictorial Feature Distortion in a Pyramid," Computer Graphics and Image Processing 5(3), 333-352.
- Tanimoto, S.L.  
1980, "Image Data Structures," in Structured Computer Vision, S. Tanimoto and A. Klinger (eds.), Academic Press, New York, 31-55.
- Tanimoto, S.L. and Pavlidis, T.  
1975, "A Hierarchical Data Structure for Picture Processing," Computer Graphics and Image Processing 4(2), 104-119.
- Treisman, A.M.  
1982, "Perceptual Grouping and Attention in Visual Search for Features and for Objects," Journal of Experimental Psychology: Human Perception and Performance 8, 194-214.
- Treisman, A.M. and Gelade, G.  
1980, "A Feature Integration Theory of Attention," Cognitive Psychology 12, 97-136.
- Treisman, A.M. and Schmidt, H.  
1981, "Illusory Conjunctions in the Perception of Objects," Cognitive Psychology 14, 107-141.
- Treisman, A.M., Sykes, M., and Gelade, G.  
1977, "Selective Attention and Stimulus Integration," in Attention and Performance VI, S. Dornic (ed.), Erlbaum, Hillsdale, New Jersey.
- Uhr, L.  
1972, "Recognition Cone Networks that Preprocess, Classify, and Describe," IEEE Transactions on Computers 21, 758-768.

- Uhr, L.  
1980, "Psychological Motivation and Underlying Concepts," in Structured Computer Vision, S. Tanimoto and A. Klinger (eds.), Academic Press, N.Y., 1-30.
- van Wijngaarden, A., Mailloux, B.J., Peck, J.E., and Koster, C.H.A.  
1969, "Report on the Algorithmic Language ALGOL 68," Report MR 101, Mathematisch Centrum, Amsterdam.
- Walker-Smith, G.J. and Gale, A.G.  
1977, "Eye Movement Strategies Involved in Face Perceptions," Perception 6, 313-326.
- Waltz, D.L.  
1972, "Generating Semantic Descriptions From Drawings of Scenes with Shadows," Technical Note TR-271, AI Lab, M.I.T.
- Warrington, E.K. and Taylor, A.M.  
1973, "The Contribution of the Right Parietal Lobe to Object Recognition," Cortex 9, 152-164.
- Warrington, E.K. and Taylor, A.M.  
1975, "The selective Impairment of Semantic Memory," Quarterly Journal of Experimental Psychology 27, 635-657.
- Watson, A.B.  
1982, "Summation of Grating Patches Indicates Many Types of Detector at One Retinal Location," Vision Research 22, 17-25.
- Weisstein, N. and Harris, C.S.  
1974, "Visual Detection of Line Segments: an Object Superiority Effect," Science 186, 752-755.
- Westheimer, G.H.  
1954, "Eye Movement Responses to a Horizontally Moving Visual Stimulus," Archives of Ophthalmology 52, 932-934.
- Westheimer, G.H.  
1982, "The Spatial Grain of the Perifoveal Visual Field," Vision Research 22, 157-162.
- Wilson, H.R. and Bergen, J.R.  
1979, "A Four Mechanism Model for Threshold Spatial Vision," Vision Research 19, 19-32.
- Winston, P.H.  
1972, "The MIT Robot," in Machine Intelligence 7, B. Meltzer and D. Michie (eds.), American Elsevier, New York, 431-463.

- Witkin, A.P.  
1981, "Recovering Surface Shape and Orientation from Texture," Artificial Intelligence 17(1-3), 17-45 .
- Woodham, R.J.  
1978, "Reflectance Map Techniques for Analysing Surface Defects in Metal Castings," Technical Report 457, AI Lab, M.I.T.
- Woodham, R.J.  
1981, "Analyzing Images of Curved Surfaces," Artificial Intelligence 17(1-3), 117-140.
- Yarbus, A.L.  
1967, Eye Movements and Vision, Plenum Press, New York.
- Zeki, S.M.  
1978, "Uniformity and Diversity of Structure in Rhesus Monkey Prestriate Visual Cortex," Journal of Physiology 277, 273-290.

#### Reference Notes

- (1) Schmidt, H., Ph.D. Thesis (in preparation), Dept. of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania.
- (2) Mulder, J. Ph.D. Thesis (in preparation), Dept. of Computer Science, University of British Columbia, Vancouver, Canada.