DESIGN OF A SEMANTICALLY DIRECTED VISION PROCESSOR

E. M. Riseman*
A. R. Hanson†

COINS Technical Report 74C-1
January 1974

This research was supported by the Office of Naval Research
under Grant ONR 049-332

*Department of Computer and Information Science
University of Massachusetts
Amherst, Massachusetts 01002

†Division of Language and Communication
Hampshire College
Amherst, Massachusetts 01002

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified

1. ORIGINATING ACTIVITY (Corporate author)

University of Massachusetts, Amherst, Mass. 01002

2a. REPORT SECURITY CLASSIFICATION

UNCLASSIFIED

2b. GROUP

None

3. REPORT TITLE

DESIGN OF A SEMANTICALLY DIRECTED VISION PROCESSOR

4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)

Technical Report

5. AUTHOR(S) (First name, middle initial, last name)

Edward M. Riseman
Allen R. Hanson

6. REPORT DATE

January 1974

7a. TOTAL NO. OF PAGES

37

7b. NO. OF REFS

42

8a. CONTRACT OR GRANT NO.

N00014-67-A-0230-0007

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

COINS Technical Report 74C-1

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES

None

12. SPONSORING MILITARY ACTIVITY

Office of Naval Research, Code 437
Washington, D.C.

13. ABSTRACT

Most vision research to date has been applied in highly constrained environments consisting primarily of objects with straight lines, simple shape, little texture, and no color. Many of the techniques developed are going to be of little help in machine perception of natural outdoor scenes. This paper discusses the preliminary design of a semantically directed vision processor. This design allows low level visual features (e.g., motion and edge detectors, texture and color descriptors, etc.) to be interfaced with high level conceptual knowledge (e.g., trees stem from the ground, are below the sky, and are not very mobile) in the perception of complex images.

The proposed system will employ some of the standard techniques of pattern recognition coupled with semantic information structured for scene analysis. This process will involve investigating several of the most plausible models of hypotheses of what is in the scene being viewed. The system will operate by reducing information in the scene with local operators in a parallel structure. The detection of prominent features in this reduced data signal the likely presence of particular objects. Each general type of object that might be viewed will have a vision procedure which will selectively analyze the lower level mass of data; a rough value of the likelihood of the object, based upon a set of features, will be returned. When any object is detected, the semantic knowledge will interact with the visual information and guide the search for objects that are related to the given object. As more of the scene is perceived, the model and semantic net will speed up further analysis. After initial processing the bulk of the analysis will be directed in a top-down fashion. Information flow and conceptual design are sketched in Figure 5.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0102-014-6700

UNCLASSIFIED
Security Classification

TABLE OF CONTENTS

LIST OF FIGURES

I. INTRODUCTION

Research in the field of computer vision has developed in a somewhat narrow, though understandable, fashion. Naive notions of AI researchers concerning the complexity of vision processing were quickly dispelled. The initial problems were not the expected difficulty of determining what objects were depicted by the lines and vertices that appeared in the image. Rather detection of the lines and vertices themselves was the initial non-trivial problem. Since then the domain of the "block's world" of polyhedra has been extensively studied [1 - 9]. Reasonable segmentation of objects, even if some are partially occluded, can usually be carried out. Object reconstruction is aided by the use of various heuristic techniques for the grouping of regions into objects. These solutions, however, have proven to be less than trivial and seem to have left uncertainty with respect to solving more general problems in the field.

There appear to be two long-range directions for vision research that are now called for. The first involves stepping into the real world of outdoors with an enormous increase in the complexity of scenes. This will force consideration of many global properties of scenes as opposed to the micro properties employed in the block's world. Many of the current techniques may have to be discarded and new ones developed. The second involves the interface of computer vision research with powerful semantic techniques so that perception will take place as a knowledge-directed process. As Tenenbaum [10] points out:

> "We feel that the time is now ripe to confront a number of these crucial perceptual issues--information overload, segmentation of textured objects, representation of irregular objects, generality of strategies--that do not arise in the blocks world. Instead of simplifying the environment, we must learn to cope with the complex-

> ity of real-world scenes by capitalizing upon their natural redundancy of descriptive features and contextual constraints."

We will briefly discuss the need and desirability for pursuing each of the research directions mentioned, and then describe a tentative design of a semantically directed vision processing system for analysis of complex outdoor scenes. Reference to Figure 5 will facilitate the discussion.

I.1 Indoor vs. Outdoor Scenes

Almost all research in scene analysis to date has been conducted in highly constrained environments such as the blocks world or a relatively uncluttered laboratory world of corridors, rooms, doors, desks, etc. In certain aspects the laboratory world does not differ dramatically from the blocks world. In each of these environments, the image is composed of basic components such as straight lines, rectangles, and the like. It has not been necessary to dramatically alter scene analysis techniques (e.g., inclusion of color and texture information) from the block's world to handle scene analysis in the laboratory world. To date there has been only a little effort expended towards the examination of natural outdoor scenes [11-12]. These images are far more complex and bear little resemblance to the images found in the blocks or laboratory environment, particularly in terms of requirements for scene analysis. Here the world is one of non-straight lines, color, varying texture, and little control of lighting. The descriptions of objects are also more complex. There is a relatively simple and general description of pyramids and cubes; on the other hand, the description of the set of all types of trees is far more difficult to state due to the variations in shape, size, color, and texture as a function of both the season and area of the country. Consequently, many of the techniques for the blocks

world may have little applicability to the real world.

The most successful effort (possibly the only serious one completed) in the analysis of outdoor scenes is that of Yakimovsky and Feldman [12]. They utilized semantic information in a decision-theoretic approach to the analysis of several road scenes. The information includes properties of the boundaries between regions (e.g., how likely is the adjacency of two regions) and properties of the regions themselves (color, shape, etc.). After initial clustering of picture points to form regions, a decision-tree analysis is used to further join and then identify regions according to a maximum likelihood analysis based on these properties. For more complex environments, we feel that the a-priori conditional probability of a feature given a region cannot be reliably estimated (usually the number of samples is very small) and probably changes over time. Thus, it is becoming apparent that the inclusion of more complex semantic information is necessary; furthermore, the nature of this information must be such that it can be utilized in a highly flexible manner.

I.2 The Utility of Semantic Information

The blocks world is almost entirely devoid of semantic information. In fact, most of the techniques that are regularly employed in scene analysis utilize only the visible physical structure of objects in a picture. Although terms such as "line semantics" or "semantic descriptors" have been applied to this work, we feel that these approaches have been primarily syntactic. The techniques, for example, involve the difference in intensity of light and dark areas as a clue to the presence of a line, the legal ways that lines can come together and still represent the vertices of a trihedral

solid, the pattern of shadows produced by such vertices and the objects containing them, the adjacency of regions in the context of various types of vertices as heuristic clues for the combining of regions into a single object, as well as the search techniques or global strategies for parsing an entire scene into a syntactically acceptable representation using combinations of these types of analyses. The main point is that the meaning of the scene within some structured world of objects and activity is utilized little, if at all. The complexity of the real world offers a richness of knowledge in a form that has not been applied in vision research.

We can gain insight from the efforts on natural language processing. Conversation systems, such as Winograd's [13], allow semantics of linguistic information to be related to a physical structure of block objects. However, there really are no general a-priori relationships between the separate objects or types of objects in this domain. Thus, while supplying a nice setting for language research, this type of problem domain seriously handicaps the vision researcher. Lately, work in speech recognition [14-16] has been blending the syntactic and semantic approaches of natural language processing with the more classical approaches of pattern recognition applied to acoustical data. As a simple demonstration of the potential of this approach to vision, one need only look at any scene through a small window and examine what he "sees" as opposed to what he "knows" is there.

In a related piece of research, one to which we are particularly sympathetic, Tenenbaum [10] has described the preliminary structure of a knowledge-based perceptual system. Though it operates in a constrained world of walls, doors, desks, chairs and telephones, it begins to utilize higher level information. Tenenbaum's approach is to define a simple two-stage procedure for distinguishing the object sought from other objects appearing in a scene. The

system relies heavily on such data as color and range (relative size) to quickly eliminate most objects from the set of possible objects. After this reduced set is found, features which pairwise disambiguate the objects are employed. Contextual information is employed to form strategies on where to look for a particular object or what objects are nearby any objects found. Currently, the SRI work is set up to interactively determine both the features sufficient to distinguish individual objects and the strategies that can efficiently use them. A general model of each object in the scene is stored internally; when an object is found its size and orientation is correlated with the internal model which is then displayed via a graphics display.

I.3 A Perspective on the Proposed System

Tennenbaum's approach is similar to the one proposed here in that coarse processing of the data (although not of the parallel type we utilize and describe later) and application of semantic information quickly reduces the size of the set of objects to which an unknown object belongs. However, although similar in concept, our approach differs in several significant ways. First, all operations are done on the raw data as opposed to our interactive layered system that is currently being examined and is described later. Second, although the scene is more complex than the blocks world, it is still a relatively simple, straight-line world; consequently, the semantic and syntactic information can be in a reasonalby straightforward form, possibly a simple table. The range and complexity of the kind of information that is required for the analysis of an outdoor scene makes our problem considerably different. We feel that this problem domain necessitates more general and powerful utilization of context, semantic knowledge, and model-

building.

We should also mention that a number of researchers are currently examining the effectiveness and economy of multi-sensory data from touch-sensors, mechanical position sensors on wheels or camera, radar, etc. [10,17,18]. In the system to be described below only visual data is considered since we are trying to develop techniques to deal with this rich source of information. However, we are in no way precluding the use of such data; in fact, it can be incorporated in a natural way into the system we envision.

In order to get some perspective on the approaches currently employed and the variety of strategies that are still available, let us consider for a moment a plausible cognitive approach of a human. Suppose we supply a person with a large photograph and constrain his view of it by forcing him to use a magnifying glass so that at any single moment he sees the grain in the picture within a very local region. If this individual is given the problem of determining roughly what is in the picture, we feel that he will use very different procedures than those used by the current vision systems. Rather than attempting to construct an outline drawing, he will look for prominant features as clues to what the image represents. He will probably scan quickly in many different directions, but once he finds a prominent feature, his strategy might become highly context sensitive and model-directed. He might form hypotheses which encompass large numbers of assumptions and use these to direct the processing until they are verified or disproven. Our interest here is not to assert a model of human problem-solving, but rather to emphasize the need for higher-level approaches to computer vision.

As we see it, one of the crucial problems is to interface low level visual information associated with local features and high level conceptual

knowledge. We propose that this can be done by quickly filtering upward processed information from local features, finding prominent features, and possible objects in the scene, and then invoking world knowledge from a type of semantic net to direct further processing. Thus, the processing initially will be bottom-up until hypotheses are formed and then will switch to top-down. The design that we outline is an ambitious project. Consequently, we will denote some of the distinct subproblems whose solution will contribute to an effective vision processor, but which can be independently investigated.

II. LOCAL FEATURES AND PARALLEL PROCESSING

II.1. Hints from Pattern Recognition

Research in pattern recognition has made it unmistakably clear that identification of the category of a pattern of information cannot be separated from either selection of the features to be employed or the pragmatic consideration of the dimensionality of this data. Suppose we utilize an array of input points that ensures fairly good resolution, say 256 x 256. Then each snapshot of a scene is comprised of 64K points with each point consisting of 6 bits (64 distinct levels) of intensity for each of 3 colors. In terms of the classical approaches to pattern recognition, this is a staggering computational overload. From this point of view, an immediate necessity is the reduction of this data to a manageable level by retaining a subset of the most relevant data or carrying out a transformation which emphasizes that data useful for classification. However, the problem we have here is far more complex than the typical pattern recognition problem, irrespective of the dimensionality of the patterns. Nevertheless, we ask the same ques-

tions: How can the data be usefully reduced in size and which features should be employed?

II.2. Data Reduction

Many of the approaches in scene analysis to date have operated on the raw digitized data exclusively. Visual systems in the animal world, however, carry out extensive parallel preprocessing. We think this is highly desirable in our problem domain. By automatically reducing the data and preserving coarse, though significant, information, detailed analysis of critical regions of the image can be selectively carried out where it would not be practical across the whole scene. Our goal here is to specify the type of information that should be retained and a simple structure for this data reduction.

The approach we put forth is the utilization of local preprocessing functions which are applied across the entire array and produce as output an array of processed data of reduced dimensionality, but which preserve spatial relationships. Conceptually, we choose to think of this stage as a parallel feature extraction process as opposed to a sequential examination of the data. Obviously, we can simulate any parallel processing in a sequential fashion. However, we think of this stage as the initial automatic processing which is not high-level goal-directed. The feature extraction stage is of low efficiency because much computation is carried out and only a portion of this output will contribute significantly. In addition, we want it carried out quickly to enable the higher level processes to be brought into play at an early point in the sequential analysis that follows. Given the state of LSI we feel that much of the hardware that we are simulating can be economically constructed in the near future, although we are not suggesting that it be built at this time.

Let us first give an example of a simple local edge-detection function operating on an ideal noise-free image of a 2-dimensional rectangular figure. Imagine that we have a logical function f of four points:

$$f(x_1,x_2,x_3,x_4) = \begin{cases} 0 & \text{if the intensities of all points are within } \Delta \text{ of each other} \\ 1 & \text{otherwise} \end{cases}$$

This equality can be viewed as a local edge detector on a 2 x 2 subregion. It outputs a signal of 1 whenever there is a difference greater than Δ between any two points. If this function is repeated on 2 x 2 non-overlapping subareas across the whole array, then the 256 x 256 grid can be reduced to a 128 x 128 grid in which 1's denote boundaries. (Note that we are ignoring quantization noise for the moment.) Thus, the information is reduced by a factor of 4 and processed so that only detected edges will show. This can be viewed as differentiation (to detect edges) and then low pass filtering (or smoothing). If the function is then repeated on this next layer, the array is reduced to 64 x 64 and the process can be repeated, say, to the 16 x 16 level as depicted in Figure 1. Note that the function is somewhat simpler after the first layer since the input has been transformed to binary. Each layer is an outline drawing that is more smoothed (or blurred) than the previous layer.

II.3. Types of Information Extracted

Now let us examine the way in which these 6 layers of information can be dynamically analyzed in software in the usual sequential fashion. Imagine viewing a tree outlined against the sky. Since it is desirable to process an image very quickly, it is not feasible to examine all 64K points on the first level. However, it is a relatively simple task on the 5th layer of 16 x 16

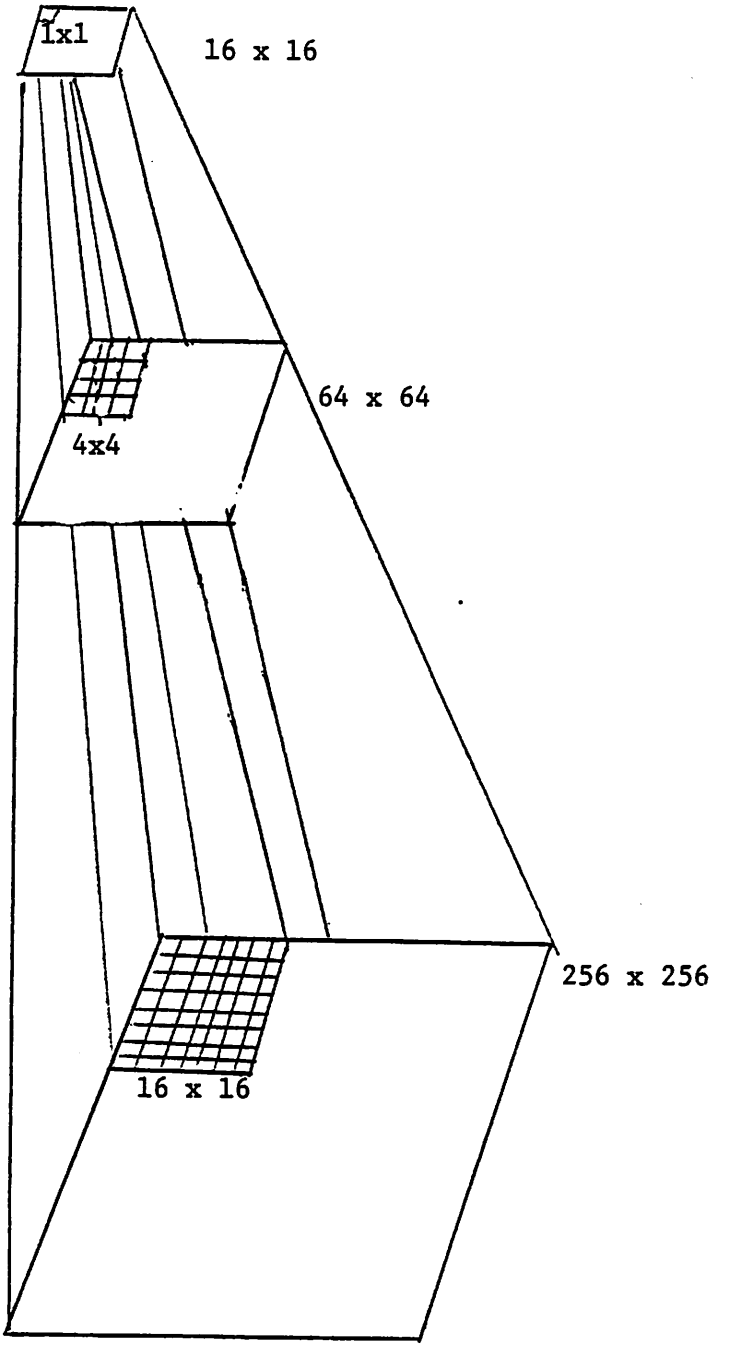

Fig. 1 - Layered Parallel Preprocessing

points to determine if there is an upright convex blob of approximately the following shape:

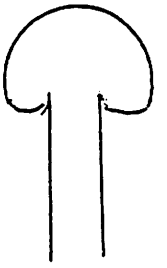

Figure 2

It is important that the functions be relatively uniform over all sub-arrays describing the scene. This will allow the information to be interpreted in a straightforward fashion. Now, it is unclear whether this is a lollipop viewed at a couple of feet or a tree at a greater distance. However, the analysis can be directed to sample portions of the upper boundary of the object in more detail, say at the 128^2 or 256^2 level. The outline should be relatively uniform for the lollipop but irregular for the tree:

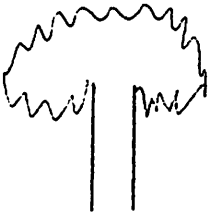

Figure 3

Thus, different knowledge of the shape of trees is available at the fine and coarse levels of processing. We will discuss later how information at the higher levels is used to direct the selective examination of the large amount of data at lower levels.

II.4. Some Alternatives for Local Operators

It should be pointed out that in the process just described, we have ignored quantization noise--edges will be overlooked if they fall on the boundary of adjacent 2 x 2 windows. There are various straightforward ways of taking care of this problem: overlapping windows on the first layer, adja-

cency operations, and/or increase in the size of the windows (say to 3 x 3 or even to 8 x 8) so that this event occurs less often. We will not discuss this aspect further. Rather we wish to address the more interesting problem of making the preprocessing work on boundaries with the presence of noise and texture.

Texture is a characteristic that is essential in the identification of boundaries. The boundary between grass and bushes might only be determined by the use of texture. However, in the case of strongly textured elements, each local operator might think that an edge has been detected and the entire region of cells could turn on. Any cell of a few points might have some dark and some light points. The effect of noise is similar. In essence, it is the spatial distribution of the intensity that is texture. In order to determine differences between regions some properties of the distribution must be determined.

The first change to accomodate these problems might be to compute average gray levels on the first layer subarrays (possibly with the expansion of size of the 2 x 2 windows). Now, a local edge would be denoted by some minimum difference in the average values of a subset of regions. On the subsequent layers, the equality function can then be reproduced with modest amounts of noise having little effect on the average values being compared. This will also solve the problem of boundaries between objects of different texture if the average gray level of the two regions is different. Rosenfeld and Thurston [19] have applied this technique in a two-pass system to detect boundaries, even when the average gray levels are the same, as long as the coarseness of the texture is different. However, the technique will break down if there is not sufficient difference in the coarseness of texture.

There are a number of variations or alternatives to the specific scheme described. A somewhat different approach involves suppression of activity of those cells when all neighboring cells have similar texture characteristics such as relatively equal mean and variance of intensity of the points in each cell. If the variances of each local region within an object are approximately the same, this information may be sufficient. A different parameterization of texture that might be useful is Haralick's spatial grey tone dependency matrix [20]. Briefly stated, this technique captures the spatial distribution of intensities by providing a count of the number of times two adjacent points have intensities i and j, and this is provided for all i and j within the allowable range of intensities. A third possibility is the use of Fourier descriptors for texture. Bajcsy [11] has shown that in some cases the directionality and quality of texture is better captured by Fourier analysis than spatial operators.

There are many possible edge detection functions that might be useful; there is a choice between separate functions of individual color intensities or a joint function of all colors. This also applies to texture descriptors. Our goal is not just detection of edges, but rather the parameterization of visual characteristics. Thus, there are other useful parallel local operators such as a color homogeneity function, a movement detection function, a straight-line angular orientation function, etc. Some of these mappings might produce non-visual information. For example, a texture function across some sub-image might output a number or vector which serves as a <u>descriptor</u> of the texture of that region. Another such mapping is a descriptor of the

speed and direction of movement of an object; note that this operation might be dependent upon comparison of previous snapshots and the determination of object boundaries. 1's or non-zeroes in these upper level arrays can denote movement within that square. A color homogeneity function mapping in a layered fashion down to say 16 x 16 will make available, in a very simple form, the information on whether a large region is of a single color since each upper point represents 256 lower level points.

To some extent, the preprocessing of the image need not be highly reliable. We do not intend to carry out detailed analyses on the upper (coarse) levels of highly processed information. Rather it will only serve as a guide for efficient detailed processing of the lower (fine) levels. Thus, in some cases, misleading processing might cause only additional computation; e.g., when an edge has been erroneously detected, this might be rectified by the use of context in the sequential processing of the lowest layers. On the other hand our best candidates for describing texture might be our local operations themselves; thus, there may be no further processing available to better characterize the texture of a region.

An additional complication but one that would provide a more flexible processing system is to incorporate feedback from the executive programs to the lower levels of the array preprocessor. Thresholds for intensity difference in edge detection, for instance, could be varied in different sections of the image under the control of the sequential analysis. This opens up a whole range of techniques for refined and tuned processing of an image.

II.5 Focus of Attention

Now we will describe a modification in the layered data structure that is under consideration. An interesting biological mechanism suggests itself

as a way to reduce the data to be processed, the foveal view of the human visual system. Only the central field of vision (several degrees) is in focus and carries detailed information of the scene; the remaining field of view seems to transmit information on a relatively gross level. Motion can be detected on the extreme periphery, but not detailed color, form, etc. As a simple experiment, one need only look at a particular object and examine what he "sees" on the periphery. However, it seems we "know" most of what is in a particular scene by using internal models to add a large amount of information to the gross view. For those interested in more detailed and biologically motivated models of human vision processing that bear resemblance to the ideas in this paper, we refer you to Arbib [21] and Didday and Arbib [22].

The suggestion for our model of a mechanical visual system is not to process all areas of the scene with equal effort. Presumably, if the region of interest were known, the major portion of computation could be carried out on the points comprising this region and points falling outside this region would receive only crude processing. This type of "selective focus" may be approximated in the proposed system by constructing the first layer in the following manner. The central field samples information in a dense fashion and areas progressively further from the center sample information with an increasingly coarser grid. The region of "focus" or central interest is resolved at the finest level and surrounding regions become less and less detailed, as shown in Figure 4.

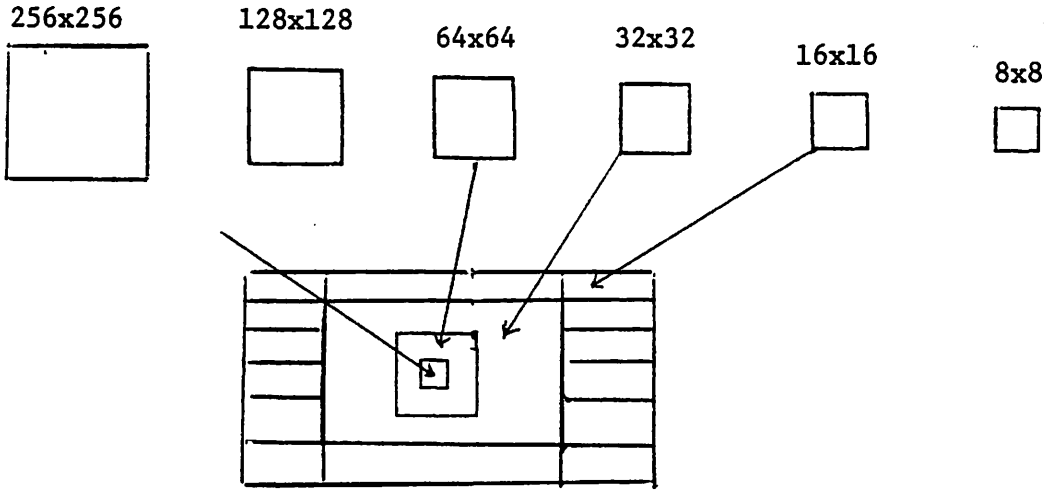

Figure 4. First Layer of a system employing a 'foveal view'

If necessary, a layered preprocessing system similar to the original layered system can now be constructed by mapping central information to the level of coarser surrounding bands and repeating this process with the new larger central area. Of course, the region of central interest may be shifted across the entire scene by altering the camera position. Although this is an interesting structure and might be useful, it will not directly affect the success or failure of our schemes for visual perception. Therefore, we will not employ this mechanism in our initial research.

II.6 Comparison with Past Parallel Processing Machines

The various layers of processed information for each of the functions employed will be assumed to be simultaneously available to the executive vision processor. Although interaction of the various processing functions might be useful, in the interest of initial simplicity we will maintain their independence. Thus, boundary detection of objects operating on several of the processed arrays will initially be a higher level process. We seek to determine how to utilize this information, not details of how the information is made available.

One could think of computational modules at each cell in each layer that can compute any of the necessary functions desired. Then we would allow the executive routine to control the switching of the functions on all cells. In the limit one can imagine a very simple general purpose minicomputer at each point in the array at each layer. In the simplest form, only a couple of these functions might be available in separate array structures. We feel that this can be intelligently discussed only when we understand the amount, the complexity, and the manner of use of the processed data. There has been extensive research on the design of general image processing machines and parallel transformations from the late 1950's to the present [23 -31]. Most of that effort was not concerned with any particular application. The design that is outlined here should not be compared with such general purpose systems. Any of these structures that can effectively perform our computations could be utilized in our scene analysis. For the time being, however, we refer the interested reader to the latest two papers which will serve as an effective introduction to this area [30 - 31]. Additionally, it should be pointed out that the limitations associated with perceptron-like parallel processing [32] are not a factor here because decisions are based upon both the processed and the raw data, upon both parallel and sequential processing.

III. THE SEMANTIC NETWORK

One of the prime goals of this research is to determine how to bring knowledge of the world (both general and specific) to bear upon the visual perception of images. The image will be processed for the purposes of identifying all large or "important" objects, and construction of a rough 3-D

model of the scene. By 'object' we mean each physical region whose boundaries should be identified (such as sky), although they are not literally objects (in the sense of manipulatable objects).

Useful knowledge might be of a general form--that trees are green and basically immobile; that people have two feet, are potentially mobile, and often appear on sidewalks, which generally have a long narrow shape; and that if the sky appears in an image, it usually appears above all other objects. Thus, we are concerned with spatial, temporal and functional relationships between the visual events of interest. The extensive base of such general information allows one to view the world in the highly structured way in which it exists. In addition, there might be available specific information about the environment under consideration. This might vary from a list of the objects that are likely to appear in the image to a complete topographical map of all objects in the environment. This information must be organized in a form which allows easy transformation into visual processes; that is, in such a way that it interfaces naturally with the visual analysis to be performed.

III.1 Representation of Semantic Information

The semantic information can be embodied in many forms. The specific physical structure does not seem to be critical at this point of the research, although two forms of representation are immediately evident, semantic networks or an axiomatized data base for theorem proving.

In the first alternative, the one towards which we are leaning, semantic information can be embedded in a directed graph structure in which the nodes are used to represent conceptual objects or their modifiers while arcs represent the relationships between them. All information which bears

directly or indirectly upon the visual image or its processing should be embodied within such a network. Thus, all physical visual attributes of an object will be associated with a node; e.g., a tree has a certain color, shape, texture, and size, and these can be stored as attribute-value pairs associated with the node for tree. In fact, the internal model of "tree" must be rich enough to embody both the general concept of a tree as well as all variations of trees that might appear in the scene being processed.

In addition to physical attributes of objects, spatial, temporal and functional relationships between objects bear useful information. One 'knows' that trees are rooted in the ground and often appear beside sidewalks. In addition, sidewalks are used by people to walk upon; therefore, one often will find people with their feet upon them. All of this information can be embedded in the network as a directed arc between the objects labelled with the name of the relationship. Quillian [33] has examined various ways of storing and retrieving information from semantic nets; e.g., the length of the path between related objects may be used as a rough estimate of the 'relatedness' between objects.

An alternative approach would be to embed the semantic information in a modified theorem-proving environment [34-37]. These theorem-provers axiomatize the semantic information at the same time that they utilize heuristic information to efficiently direct large-scale searches. Our use of a theorem-proving system would most likely be based on heuristic search methods coupled with a model approach to theorem proving [38]. As in a straightforward theorem prover, a set of rules (or clauses) is provided describing the permissible deductions, and a set of predicates and functions describe the relationships between objects. However, the use of a simple theorem proving system

applied to real world environments has the obvious disadvantages; it is necessary that information from the model and clues from prior proof procedures be blended with heuristic search procedures in order to guide the proof procedure.

Regardless of the structure chosen to embody the semantic information, the problems encountered concerning efficient use of this information remain. Questions still remain open such as how to employ heuristic information in order to drastically reduce retrieval or proof time and how to best employ models of the environment. We feel that both approaches are viable and interface with the rest of the system in a similar fashion. For the rest of this paper, we will only refer to the semantic net as the means of storing and retrieving semantic information.

III.2 Application of Semantic Information to Scene Analysis

We have outlined a relatively straightforward construction of a network for semantic retrieval of information. However, there has been little attempt in vision research to apply structures of this type. One possible reason for this is that the information in these networks is not in a functional form. If the information is stored in this network with descriptive symbolic labels (i.e., the actual words associated with the concepts), then it must be translated into a form in wiich the knowledge may be applied to the image. This is a distinct limitation of such nets. We may retrieve the information that the "trunks" of trees are generally "vertical". However, in terms of processing the image, this means that the lower portion will have a narrow boundary running up and down. Winograd [13] represented knowledge in a procedural form in his natural language processing system. This latter technique will be used to make our semantic network into a far

more flexible medium. For example, a piece of descriptive information about some object can be associated with a subprocedure; a set of programming statements. This subprocedure can describe how the image should be analyzed in order to detect that characteristic of the object. This representation of information allows it to be functional; a node has a mechanism by which it can actively operate on the image rather than remain as a passive placeholder.

Now we have a computational structure for the semantic network. We can build it in the usual descriptive form and associate with each node a procedure to apply the information. Thus, we can view the structure as a network of symbolic labels or a network of labelled subprocedures. The concept of "tree" is defined as a graph structure of concepts with words as labels ("green", "tall", "trunk", "leaves"), and a subprocedure associated with determining the presence of "tree". This approach leaves us with a network of vision procedures, one associated with each object. All information about a single object which has a direct visual component is now accessible through that object procedure.

The only other information is that relating the different objects of the environment. Thus, there will be additional arcs associated with spatial and functional knowledge. An object procedure will be connected to other object procedures by this relational knowledge. In addition to this information we will provide a set of useful entry nodes into the network. These nodes will be sets of pointers to the various objects that have particular characteristics. Thus, there will be a node associate with each measurable concept or physical attribute that can be associated with an object, and a pointer to each object that has this characteristic (or a particular value

of the attribute). For example, there will be a node for color and a list of pointers to objects which have each particular color. The object descriptions will contain the necessary modifiers operating on, in this case, the attribute color. In this way it will be relatively easy to access the subset of objects which are green, or mobile, or with straight edges. These entries will be necessary for the executive system in constructing models of the scene.

IV. VISION PROCEDURES

Up to this point we have discussed the layered preprocessing of the visual data and the network of conceptual information. Here we describe the form of the actual algorithms which operate on the visual data and interface with the semantic knowledge. The purpose of these procedures is to determine roughly the likelihood of the presence of various objects in the different regions whose characteristics have been tentatively identified.

Consider the procedure associated with "tree". There are many types of trees with varying characteristics; each of these variations must be taken care of in the semantic net as we have described. Thus, one type of tree might have a "trunk" consisting of a vertical, farily straight brown area from 1" to 5' in width and 3' to 100' high, full round "foliage" with deep green leaves, a given type of texture, ..., etc. This information must in some way be correlated with the visual data if the presence of this tree is to be determined. This portion of the process of scene analysis is very similar to the classical problem in pattern classification.

IV.1. The Pattern Recognition Approach

Pattern classification is generally viewed as at least a two-part procedure: feature extraction and classification. The selection of a set of

features upon which the decision will be based is most crucial. Reliability of decisions is directly limited by the quality of information in the feature measurements. We have already discussed the extraction of some features through extensive parallel preprocessing. Now the task is to determine the subset of information from any of the layers that signal the presence of, say, a tree. We must decide which of the concepts associated with an object have measureble and relatively invariant visual components (sometimes dependent upon such things as the season of the year, etc.). Some features describing a tree are color, shape, size, and texture of the trunk and foliage. Higher level features which are simple functions of the processed data may be quite useful. Since shape is complex and difficult to describe, one might employ a ratio of area to perimeter [6,10]. There is a great deal of flexibility in writing a procedure to examine specific portions of the information. The uppermost layers contain the coarsest information, but they are of low dimensionality so they can be examined quickly in a sequential manner. The rough shape, color, and texture might be sufficient to determine the likely absence or presence of a tree. However, since any of the more detailed layers can be accessed, as we pointed out earlier, the boundary should be uneven on a finer grid and the color might be less uniform upon a micro-examination (which may be correlated to the texture). Features which are effective in separating some pair of categories (objects) may be particularly useful in reducing critical ambiguity [10,39-40].

The second phase is the classification process. Here we view this as determining the likelihood of the presence of the goal object given the values of the features measured. The goal here is to get a <u>coarse</u> evaluation

such as "improbable", "low", "medium", "high", "almost certain". This crude decision information will be fed to an executive system (the vision monitor to be described shortly) so that it can be integrated into a global model of the scene. The coarse evaluation can be accomplished by mapping the statistical evaluation of a standard pattern classifier into the desired several confidence levels (fuzzy set theory [41] might be usefully applied here). As we mentioned earlier, one of the serious obstacles to this approach is a reliable estimate of the class conditional probability densities for each feature. This is also complicated by the occlusion of objects which may make it difficult to measure the degree of presence of a feature (such as shape). A decision tree approach [12] to this problem heuristically based on outstanding (and hopefully simple) characteristics may overcome this difficulty. Of course, statistical formulations can be employed where useful but we think they will be limited. This appears to be one of the most difficult portions of the research design to implement. This explains our motivation for easing the constraints on our classifier to one of making coarse decisions. We will leave the responsibility for removing ambiguity, making finer decisions, and correcting errors to the executive system which will bring contextual and semantic knowledge into the perceptual process.

Before we discuss the model building process, we will digress briefly to discuss the efficient application of the (possibly) many object procedures whose processing we envision to be primarily sequential. The following section outlines techniques to focus on a few of the object procedures and a few subareas on the basis of high level mappings of information on the

top layer.

IV.2 Contextual Cueing Specialists

Certain objects are relatively easy to discern in a given image. This may be due to various factors. If one sees a large expanse of blue in the top of the image with some green splotches below, a reasonable initial hypothesis is that it is blue sky with trees below; a less likely hypothesis is that it is water with some type of vegetation in the foreground. Of course, this must be coupled with information concerning the location of the observer, the direction (up or down) of the gaze, etc; the latter hypothesis is far more likely if one knows that a pond, for example, is nearby, and the observer is outdoors looking down. Similarly, a long broad straight line might denote a sidewalk or a building; a region of the image which is moving must be one of a relatively small subset of object types--person, animal, car, bicycle. If the procedures associated with these object types can be applied selectively to the particular regions in question, the construction of hypotheses as models will function more accurately and efficiently.

Specifically, the set of features which flag the likely presence of certain objects will be searched for in the top layers of the processed visual data by a set of special procedures called "contextual cueing specialists". These procedures can examine the top layer (in general, any of the layers of coarsest information) and selectively step down through the layers when more detail is required. Whenever any of these prominent features is found, one or at most a few object procedures will be activated in the

region of that feature.* Thus, the sequential scan is quick because of the small amount of processed data and the tentativeness of the decision. This initial scan very selectively invokes a few object procedures which will examine the much larger amount of data on lower levels.

V. VISION MONITOR AND MODEL CONSTRUCTION

Our goal now is to somehow amalgamate the various forms of information at a very early stage of the analysis so that a hypothesis of the scene can be formed. This model will direct further processing of the data with the goal of either completing the model or rejecting it and finding an alternative model. The desired output is an intermediate level detailed drawing of boundaries and identification of objects of any significant size.

The vision monitor is the system what will oversee the construction of the model, utilizing information from all sources. The operation of this visual system is highly modularized and basically allows a heterarchical control structure [42]; each process can call other processes and the particular sequence of program interactions is dependent upon the data in the image. Information is fed to the vision monitor which is responsible for the final decisions on the model. The entire sequential vision processing of the information is tailored for efficiency of computation and semantically driven analysis. A schematic diagram of the system is shown in Figure 5.

---

*In the next section, we discuss the construction of a model of the environment that satisfies the semantic constraints. Once the model is initially constructed the contextual specialist must agree with the model before automatically invoking the object procedure.

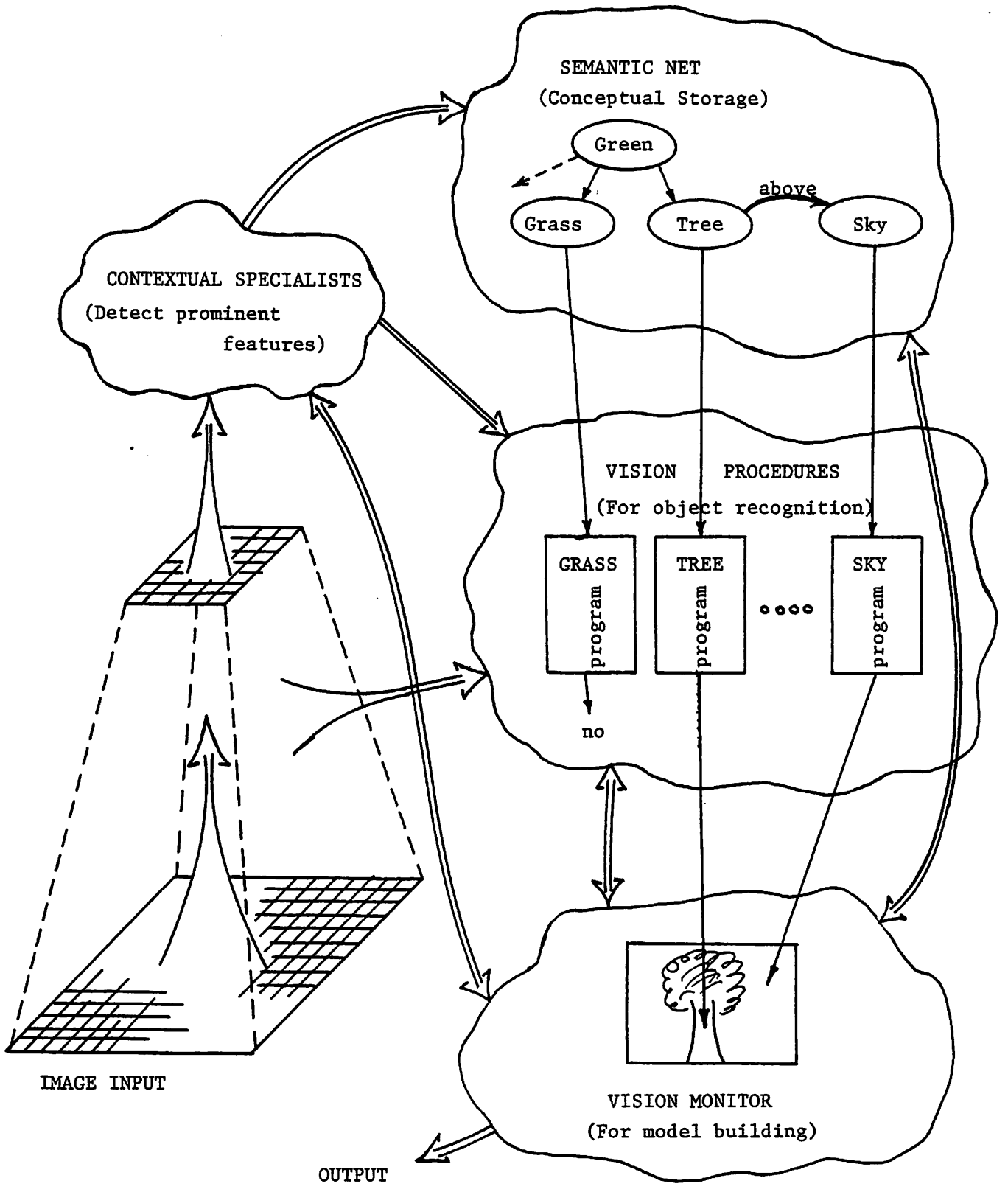

Fig. 5 The Vision Processing System

Let us summarize the operation of the system that has been developed up to this point. A digitized image is automatically preprocessed in a layered fashion. The upper levels are scanned for outstanding features by cueing specialists. Any features that are discerned with some degree of confidence call up a subset of vision procedures associated with various candidate objects. These procedures analyze the visual information selectively down to the lowest levels. If one or more of these procedures identify one or more candidates as being likely, then the semantic network can be consulted. The presence of a sidewalk should have "person" as one of its strongest candidates (and not "tree" or "building") and be investigated first. The construction of an internal model, as discussed below, would tag the hypothesis "person" as potentially mobile and subsequent processing of later scenes would tend to verify or reject this hypothesis. Once the identity of one or more objects is tentatively fixed, information from the semantic net in conjunction with high level processed data can direct further visual processing. From this point on, one can view additional processing as construction of a model of the world being viewed. This analysis will be under the control of the vision monitor although the type of processing by the systems described will still be taking place.

The semantic network will immediately be able to retrieve both general and specific information concerning possible objects related to the identified objects or prominent features. The semantic net has been set up to be applied simply and flexibly. The net can direct processing to the next subset of object choices whose presence is most likely assuming the presence of some given objects, as well as defining the probable subregions in which they will appear. Prominent characteristics such as color and size might be

accessed to determine which objects have these characteristics.

In the case of general information, the vision monitor will attempt to correlate the semantically identified objects with those already suggested by cueing specialists but yot yet evaluated because of lower priority. Any objects supported by information from both the network and the cueing specialists can be examined more carefully by their respective vision procedures. During detailed processing by an object procedure, specification could be made to examine the suitability of the context to the given object. This would lead to suspension of the present analysis and accessing the surrounding processed features and other related objects through the semantic net and/or model for gross compatibility; several bad fits of this data on the one hand, or likely matching characteristics of semantically related objects and subregions on the other, would imply very different expectations with respect to the presence of the object in the region originally examined. Of course, this analysis can be continued recursively to any depth but we do not see the need or desirability of more than a couple of layers of recursion. This illustrates the usefulness within a vision procedure of providing alternative subprocedures of varying computational complexity, depending on the desired confidence of the decision. If a topographical map is available, then the tentatively identified objects can be fitted to alternative choices in the map. Then specific nearby objects can very quickly be checked for consistency.

Usually there will be many choices open to the vision monitor in constructing the model as well as the problem of competing alternative models. In the latter case such ambiguity can be resolved by directing the visual analysis to those regions or objects which will verify or contradict one or more of the models. The coarse likelihoods of the presence of objects in

various regions (some of which may be competing) provided by the vision procedures must be used to direct the system to the globally most likely interpretation of the image. If a map of the environment is available, one might continue processing of the image in that area where the visual data is most likely to disagree with the projected model. In any case the problem of ambituity between models is open to a variety of strategies and limited search procedures probably cannot be avoided.

The initial construction of a model of the image should speed by orders of magnitude the further development of the model utilizing semantic direction. We conjecture that erroneous models will be discerned very early by inconsistencies between implied semantic data and the visual data. Thus, it is hoped that after the initial phases of analysis, the system should perceive the image even more quickly. Thus we have the start-up problem; once the system is in a steady state, the temporal relationships between objects of a given scene will become more important than the spatial relations. Although some level of detail might be glossed over due to lack of semantic importance, most key aspects of an image should be focussed upon rather quickly.

SUMMARY

This report outlines the motivation and design of a vision processing system to operate on real-world outdoor images. The system utilizes both pattern recognition procedures for analyzing visual data and higher level AI techniques for applying world knowledge. By embedding conceptual knowledge and carrying out extensive parallel preprocessing, much of the visual analysis is able to be directed in a top-down fashion. Hypotheses of what is in the scene are formed rather early in the analysis. Then the model, prom-

inent visual features, and conceptual knowledge are used for affirmation and completion of the model or rejection in favor of an alternative model.

This design represents preliminary ideas on a vision processing project that is being started. The COINS department at the University of Massachusetts has a flying spot scanner for digitizing 35 mm slides on a 256 x 256 grid in three colors. This is interfaced to a PDP-15 computer with a disk and magnetic tape. A color display system is now under construction.

The scenes that will be analyzed will probably be from the University of Massachusetts campus. The initial research will focus on the preprocessing of the image and the object recognition routines. The second phase of the work will employ a small semantic network and a rough topological map of the environment. It is envisioned that as the semantic net is enlarged and the full capabilities of model-directed analysis are obtained, more complex scenes without the use of a map will be examined.

REFERENCES

[1] A. Guzman, "Decomposition of a Visual Scene into Three-Dimensional Bodies," Proceedings of Fall Joint Computer Conference, vol. 33, 1968, pp. 291-304.

[2] D.A. Huffman, "Impossible Objects as Nonsense Sentences," Machine Intelligence 6, B. Meltzer and D. Michie (eds.), American Elsevier, 1971, pp. 295-323.

[3] M.B. Clowes, "On Seeing Things," Artificial Intelligence, vol. 2 vol. 2, Spring 1971, pp. 79-116.

[4] C.R. Brice and C.L. Fenema, "Scene Analysis Using Regions," Artificial Intelligence, vol. 1, Fall 1970, pp. 205-226.

[5] T.O. Binford and J.M. Tenenbaum, "Computer Vision," Computer, May 1973, pp. 19-24.

[6] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, 1973.

[7] P.H. Winston, "The MIT Robot," Machine Intelligence 7, B. Meltzer and D. Michie (eds.), American Elsevier, 1972, pp. 431-463.

[8] Y. Shirai, "A Heterarchical Program for Recognition of Polyhedra," Memo No. 263, Artificial Intelligence Laboratory, MIT, Cambridge, Mass., 1972.

[9] D. L. Waltz, "Generating Semantic Descriptions from Drawings of Scenes with Shadows," AI TR-271, MIT, Cambridge, Mass., 1972.

[10] J. Tenenbaum, "On Locating Objects by their Distinguishing Features in Multisensory Images," SRI Technical Note 84, AI Center, Stanford Research Institute, September 1973.

[11] R. Bajcsy, "Computer Description of Textured Surfaces," Proc. of 3rd IJCAI, August 1973, pp. 572-579.

[12] Y. Yakimovsky and J.A. Feldman, "A Semantics-Based Decision Theory Region Analyzer," Proceedings of Third Joint Conference on Artificial Intelligence, August 1973, pp. 580-588.

[13] T. Winograd, Understanding Natural Language, Academic Press, New York, 1972.

[14] D.R. Reddy, L.D. Erman, R.D. Fennell, and R.B. Neely, "The Hearsay Speech Understanding System," Proc. of 3rd Joint Conf. on AI, August 1973, pp. 185-193.

[15] D.E. Walker, "Speech Understanding Through Syntactic and Semantic Analysis," Proc. of 3rd Joint Conference on AI, August 1973, pp. 208-215.

[16] W.A. Lea, M.F. Medress, and T.E. Skinner, "Use of Syntactic Segmentation and Stressed Syllable Location in Phonemic Recognition," 84th Meeting of Acoustical Society of America, November 1972.

[17] B.M. Dobrotin and V.D. Scheinman, "Design of a Computer Controlled Manipulator for Robot Research," Proc. of 3rd Joint Conf. on AI, August 1973, pp. 291-297.

[18] M.H. Smith and L.S. Coles, "Design of a Low Cost General Purpose Robot," Proc. of 3rd Joint Conf. on AI, August 1973, pp. 324-335.

[19] A. Rosenfeld and M. Thurston, "Edge and Curve Detection for Visual Scene Analysis," IEEE Transactions on Computers, May 1971, pp. 562-569.

[20] R. Haralick, "Textured Features for Image Classification," IEEE Trans. on Systems, Man, and Cybernetics, pp. 610-621, Nov. 1973.

[21] M.A. Arbib, The Metaphorical Brain, John Wiley & Sons, 1972.

[22] R.L. Didday and M.A. Arbib, "Eye Movements and Visual Perception: A 'Two Visual System' Model," Tech. Report 73c-9, Computer and Information Science, University of Massachusetts, Amherst, December 1973.

[23] S.H. Unger, "A Computer Oriented Toward Spatial Problems," Proc. IRE, Oct., 1950, p. 1744.

[24] B.H. McCormick, "The Illinois Pattern Recognition Computer-ILLIAC III," IEEE Trans. on Electronic Computers, 1963, pp. 791-813.

[25] R. Narasimhan, "Labelling Schemata and Syntactic Descriptions of Pictures," Information and Control, 1964, pp. 151-179.

[26] M.J.E. Golay, "Hexagonal Parallel Pattern Transformations," IEEE Trans. Comp., vol. C-18, August, 1969, pp. 733-740.

[27] A. Rosenfeld, "Connectivity in Digital Pictures," Journal of ACM, vol. 17, Jan. 1970, pp. 146-160.

[28] E.G. Johnston, "The PAX II Picture Processing System," Picture Processing and Psychopictorics, Eds., B.S. Lipkin and A. Rosenfeld, Academic Press, 1970, pp. 427-512.

[29] B.S. Gray, "Local Properties of Binary Images in Two Dimensions," IEEE Trans. Comp., vol. C-20, May, 1971, pp. 5551-5561.

[30] M.J.B. Duff, D.M. Watson, T.J. Fountain, and G.K. Shaw, "A Cellular Logic Array for Image Processing," Pattern Recognition, vol. 5, September 1973, pp. 229-247.

[31] B. Kruse, "A Parallel Picture Processing Machine," IEEE Trans. on Comp. Vol. C-22, December 1973, pp. 1075-1087.

[32] M. Minsky and S. Papert, Perceptrons: An Introduction to Computational Geometry, MIT Press, Cambridge, 1969.

[33] M.R. Quillian, "The Teachable Language Comprehender: A Simulation Program and Theory of Language," Comm. of ACM, vol. 12, August 1969, pp. 459-476.

[34] R.E. Fikes and N.J. Nilsson, "STRIPS: A new approach to the application of theorem proving to problem solving," Artificial Intelligence 2, pp. 189-208, 1971.

[35] C. Hewitt, "PLANNER: A Language for Manipulating Models and Proving Theorems in a Robot," AI Memo 168, AI Project MAC, Cambridge, Mass: MIT, 1968.

[36] G.J. Sussman and D.V. McDermott, "From PLANNER to CONNIVER - A Genetic Approach," Proc. of 1972 FJCC, pp. 1171-1179, 1972.

[37] D. Bobrow and B. Raphael, "New Programming Languages for AI Research," Tutorial presented at 3rd Joint Conf. on AI, August, 1973.

[38] B. Meltzer, "A New Look at Mathematics and its Mechanization," Machine Intelligence 6, (B. Meltzer and D. Michie, Eds.), Edinburgh: Edinburgh University Press, pp. 63-70, 1968.

[39] A. Hanson and E. Riseman, "System Design of an Integrated Pattern Recognition System," (Abstract), Proc. of First Int. Joint Conf. on Pattern Recognition, p. 342, October 1973; available as Tech. Report 73C-5, Comp. & Information Science, University of Massachusetts, Amherst, June, 1973.

[40] A. Hanson and E. Riseman, in Preparation.

[41] L.A. Zadeh, "Fuzzy Sets," Information and Control, vol. 8, pp. 338-353, 1965.

[42] M. Minsky and S. Papert, AI Progress Report, MIT Technical Report, January 1972.