SRI International

July 1975

REPRESENTATION AND USE OF KNOWLEDGE IN VISION

by

H. G. Barrow
J. M. Tenenbaum
Artificial Intelligence Center
Stanford Research Institute
Menlo Park, California 94025

Technical Note 108

This Technical Note is an expanded version of a paper that originally appeared in SIGART, No. 52 (June 1975).

SRI International®

333 Ravenswood Ave. • Menlo Park, CA 94025
(415) 859-6200 • TWX: 910-373-2046 • Telex: 334 486

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUL 1975		**00-07-1975 to 00-07-1975**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Representation and Use of Knowledge in Vision	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **SRI International,333 Ravenswood Avenue,Menlo Park,CA,94025**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**20**	

ABSTRACT

This paper summarizes the present state of research in scene analysis. It identifies fundamental information-processing principles relevant to representation and use of knowledge in vision and traces limitations of existing programs to compromises of these principles necessitated by extant processors. Some specific and general recommendations are offered regarding a productive course of research for the next decade.

I INTRODUCTION

Visual perception, whether by a human or a machine, is a process of explaining sensory data in terms of a priori models of the world. In its simplest form, the perceptual process might entail partitioning a picture into "coherent" regions, based on, say, models of texture or color homogeneity. Higher levels of perception might entail partitioning the picture into "meaningful" regions, based on models of particular objects, classes of objects, likely events in the world, likely configurations, and even nonvisual events. Vision might be viewed as a vast, multilevel optimization problem, involving a search for the best interpretation simultaneously over all levels of knowledge.

There is an important point to note here. The perceptual process, as applied to a given scene, does not have a single "correct" solution; the desired optimum must depend on the goals, interests, background, and even sensory acuity of the perceiver.

This paper addresses the following key artificial intelligence (AI) issues within the vision paradigm formulated above:

(1) What knowledge is necessary to segment and interpret various classes of scene?

(2) How should knowledge be represented and used?

(3) Is there knowledge general enough to encompass a wide class of scene types?

(4) How should knowledge be acquired?

We begin with a brief summary of how knowledge has been represented and used in some existing systems.

II BACKGROUND

Research to date has resulted in a few scene-analysis systems that perform tolerably well in segmenting and interpreting images in some limited domains, notably polyhedra, human faces, indoor scenes, and simple landscape scenes, and in a few uniform procedures for partitioning arbitrary pictures into coherent regions. Table 1 presents just a few representative systems that indicate the current level of competence in the field.

The systems described in Table 1 depend strongly on characteristics of their particular domains. The toy worlds are usually characterized by sensorily deprived images (e.g., no shadows or texture), a small number of object prototypes, precise structural object descriptions, few constraints of a purely structural nature among objects or even pieces of objects, and, perhaps most significantly, the existence of unambiguous, goal-independent interpretations. The worlds of faces and indoor scenes are characterized by strong contextual constraints, unambiguous interpretations, classification by function, and slightly richer sensory data, notably color. From these worlds it is a significant step to the outdoor world of landscape scenes, where overwhelming amounts of image detail, ambiguity, weak contextual constraints, lack of functional names, variability of members of a given conceptual class, and many possible varieties of context occur.

It became apparent rather early that arbitrarily complex reasoning involving many types and levels of knowledge may be necessary to produce a good description of a scene. Considering the richness and complexity of the world, it was felt necessary to limit the range of difficulties to be faced by limiting the problem domain. This led to research into narrow systems with deep knowledge about a very restricted class of scenes, for example, programs in the world of polyhedral objects know about gravity and support. By relying heavily on particular characteristics of the domain, such programs could afford to ignore knowledge and broad principles necessary for scene analysis in general. The selected characteristics are often caricatures of some limited aspect of reality, and are commonly implicitly embedded in the computer program in a way that makes it difficult to appreciate not only their significance for performance, but sometimes even that they are represented at all. Thus scene-analysis systems for the world of polyhedra are based on the straight line as an element of structure, whereas

Table 1

REPRESENTATIVE SCENE-ANALYSIS SYSTEMS

Domain	Author	Description of Program
Toy worlds		
Polyhedra	Roberts	Interpretation of extracted line drawings as three-dimensional objects, based on generic prototypes of wedge, rectangular prism, and hexagonal prism (1).
	Falk	Interpretation of extracted, noisy, line drawings based on nine fixed-size object prototypes (2).
	Shirai	Extraction of line drawings for arbitrary configurations of bricks using heuristics that predict additional lines in a partially completed drawing (3).
	Waltz	Interpretation of ideal line drawings as arbitrary trihedral solids based on a catalog of legal edge and vertex types (4).
Curved objects	Agin	Description of several curved objects, such as Barbie dolls and toy horses, by fitting generalized cylinders to range data (5).
	Horn	Reconstruction of arbitrary curved surfaces, such as noses, from brightness profiles (6).
	Turner	Interpretation of ideal line drawings as objects with second-order curved surfaces. Interpretation of extracted line drawings as cylinders, cones, and polyhedra (7).
Human faces	Kelly	Sequential location of prominent features, e.g., eyes, nose, and mouth, top-down in a predetermined, contextually dependent order within crude face outlines extracted from a coarse image (8).
	Fischler	Simultaneous location of prominent facial features by finding optimal embedding of a flexible template (9).
Indoor scenes	Garvey and Tenenbaum	Location of objects, e.g., door, tabletop, and chair, in multisensory images using interactively and automatically generated, distinguishing-feature strategies (10).
Outdoor scenes		
Landscapes	Lieberman	Location of prominent landscape features, e.g., sky, trees, and water, top-down in a predetermined, contextually dependent sequence based on images crudely segmented into homogeneously colored regions (11).
Road scenes	Yakimovsky	Segmentation of simple road scenes into regions yielding the optimal joint interpretation according to semantics based on Bayesian statistics (12).
Cityscapes	Ohlander	Segmentation of outdoor scenes by recursively partitioning color and brightness distributions (13).

most landscape programs totally ignore all structure other than adjacency. Hindsight enables us to realize that, in many cases, while simplification of the domain revealed some of the central issues, it actually made scene analysis more difficult.

The performance and generality of programs within any domain depend on the amount of available knowledge about the domain, the amount of flexibility available in using that knowledge, and the degree of modularity with which new knowledge is incorporated. Winston[14] has documented this phenomenon in successive generations of blocks-world programs written at MIT. Robert's line-extraction program, for example, used implicit knowledge of the importance of straight edges in concatenating gradient points into line fragments, merging collinear fragments into lines, and segmenting lines at corners. This knowledge and its rigid application in a sequence of independent local operations proved inadequate because some lines were (inevitably) missed, shadows, reflections, and surface markings caused extra lines and errors made at one stage were compounded at later stages. These deficiencies motivated Shirai to invoke additional domain-dependent knowledge in the line-finding process.

Shirai's program found edges, one at a time, at each step using heuristic knowledge about line drawings of polyhedral bodies implicit in the context of previous results to predict where additional lines would most likely be found. The program knew to look first for the outer boundaries of an object or cluster of objects, presuming high contrast between white blocks and black table, and then to guess lines that were extensions of, or were parallel or perpendicular to, lines already found. The search for predicted lines was concentrated around vertices where they were most likely to be found. The priority of search depended on the specificity of the prediction. The additional knowledge helped Shirai's program obtain significantly better performance and reliability.

Another interesting contrast is found in a comparison of the techniques for line-drawing interpretation employed by Roberts and by Waltz. Interpretation to Roberts meant choosing a stored, three-dimensional model which, after translation, rotation, and scaling, could be projected to match part of the image. The recognized piece was then cut away and the remainder considered. Unfortunately many distinct decompositions of cluttered scenes and even of single objects could usually be obtained with a given set of models. Waltz's program avoided this difficulty by abandoning object models and relying instead on a large number of local descriptions of edges and vertices and rules for combining them to form legal scenes of trihedral solids. (Comparisons of this type are of course limited because programs seldom share exactly the same objectives.)

The work of both Kelly and Fischler on faces also provides some illuminating contrasts on how the use of knowledge affects program performance and generality. Both programs model faces with caricatures of prominent features, such as head outline, eyes, nose, and mouth, whose relative positions are allowed to vary within prescribed limits. Each feature is represented procedurally in terms of its gross attributes (e.g., nose might be represented by a routine that tests for two dark spots corresponding to nostrils), whose meaning derives primarily from the global context. However, these programs use their common knowledge of faces in dramatically different ways. Kelly's strategy was to search the image top-down for a collection of edges that could correspond to a plausible head outline. This outline was used to predict approximate location in which to search for facial features using the crude procedural templates. The program looked for the usually most reliable features and used them to refine further the predicted locations for subsequent features. Global confidence accrued as additional features were verified. The eyes were sought in an area defined with respect to the top and sides of the head outline. If this search succeeded, the nose was sought in an area predicted with respect to the eyes. Then the mouth was sought in an area delimited vertically by the nostrils and bottom of the head outline and horizontally by the two eye centers.

A major shortcoming of Kelly's and many similar programs was that features were sought in a predetermined, contextually dependent sequence. Hence a single obscured feature could derail the entire analysis. A second, and related, drawback was lack of generality. Knowledge about objects was deeply embedded in procedures often in the form of ad hoc heuristics and thresholds. Developing these procedures was a tedious empirical process that had to be repeated for another narrow domain. Furthermore, even small modifications can involve extensive reprogramming; a single new object might substantially alter the features needed to distinguish previously defined objects as well as the contextual order in which objects should be acquired.

Fischler avoided the above difficulties by abandoning the notion of sequential search. Instead, he evaluated simultaneously the best overall embedding of a face prototype in the image. The facial model was used as data for a general optimization procedure that treated the model as an elastic template. In Fischler's program, unlike Kelly's program, the same optimization procedure could be used with different data to interpret a variety of domains.

Lieberman's program for interpreting landscape scenes used a sequential context-driven strategy, similar to Kelly's, while Yakimovsky's program used a global optimization approach similar to Fischler's.

We mention in passing that another way of overcoming the above limitations is to develop a system that can plan its own strategies based on global knowledge about a particular domain and specific knowledge about the current scene. Such a system could alter its strategy dynamically to capitalize on the strongest available evidence. Moreover, it would have the inherent generality of being able to function in any environment for which its knowledge base and perceptual primitives were adequate. Garvey's work on room scenes was the first attempt to automate the development of perceptual strategies in this way.

More importantly than particular programs, research to date has given us some understanding of some of the design principles for visual systems and an appreciation of the type of knowledge they must possess to function well.

What Knowledge Is Needed?

Relevant knowledge, used appropriately, can have only a positive effect upon the competence of a scene-analysis system. The more knowledge the better, apparently without limit. Adding (correct) information enables interpretation in some cases where interpretation was not previously possible, or corrects some erroneous analyses; deleting information has the reverse effect. There are caveats, of course, but the evidence indicates that there is enough redundancy in the real world so that the gains from adding knowledge outweigh the additional burdens when the knowledge is properly used. There appears to be no firm boundary between seeing and reasoning; in order to see, various amounts of reasoning may be necessary.

This precept finds support in the manifest disadvantages of a parsimonious description, that is, an object may be unrecognizable if a single crucial feature is obscured. It is generally more economical and reliable to characterize objects by a redundant juxtaposition of many, perhaps crude, features than by a detailed but nonredundant description of a single feature. Moreover, a redundant juxtaposition may provide the only feasible description for highly variable objects. A tree, for example, is more naturally characterized by the gross attributes "large," "green," and "irregularly shaped" than by a detailed description of any single attribute. The desirability of redundant data in the form of multiple or multisensory images follows as a consequence of the desirability of redundant description.[16]

A competent scene-analysis system must embody, explicitly or implicitly, knowledge about the (picture-taking) process that relates

the picture to the scene. An understanding of this process involves knowledge of projective geometry, topology, optics, and electro-optical transducers, as well as information about surfaces and lighting, including reflectivity, specularity, highlights, color, surface texture, shading, and the like. This concern with the physical basis of the picture is one of the principal distinctions between scene analysis and pictorial pattern recognition.

The above knowledge explains the relation between a patch of surface and the corresponding patch of picture, but patches of surface are not arbitrarily arrayed: surfaces may be continuous or discontinuous and may occlude one another. Structure is important and manifests itself in terms of objects and their features, for example, bumps and hollows. We normally describe the world in terms of objects, and we wish a scene-analysis system to do the same. Note that "object" is a fundamental notion, imposed by the beholder, and is context dependent. Often, the evidence for attributing two regions of a picture to different objects may be weaker than the evidence for further subdividing one of the regions. Only experience and context allow the right decision. For example, consider a black-and-white cat sleeping on a black fur rug. The system must know about likely objects and parts of objects; to deal with new cases of the same sort, it should have both generic and specific knowledge of cats.

Once the concept of object has been introduced, it follows that knowledge of the ways in which objects interact and behave, likely events, and so forth, helps in the analysis of pictures. For example, gravity and support help to judge relative sizes: Since two men cannot be floating in the air, their feet must be resting on the ground, which means that the man whose feet are higher in the picture is further away. The fact that this man also appears larger must be reconciled by concluding that the other man is short. Naturally, we are not suggesting that such reasoning occurs in precisely this form when we see; however, knowledge often has ramifications that lead the system to behave as though such reasoning had occurred.

Reddy[16] has listed some examples of knowledge that can be employed in scene analysis and has categorized them in an interesting way. A scene can be described hierarchically at the feature level as a vector for each pixel, for example, intensity, hue, saturation, depth, and the like; at the segment level as a contiguous grouping of pixels which do not differ significantly in one or more feature dimensions; at the region level as sets of segments grouped over shadows, occlusions, and highlights; at the object level as several regions belinging to one or more objects; at the structural level as ensembles of objects which obey known relationships (e.g., table and chairs); and finally at the

environment level as a single, meaningful description of the whole scene. In addition to specific knowledge mentioned by Reddy, there is also a substantial amount of metaknowledge concerned with applicability and reliability in particular contexts. For example, the rule "sky is blue" depends on implicit assumptions of time of day, season, weather, geography, location, and so forth.

Reddy views knowledge as a set of rewriting rules which transform an image represented at one level into a corresponding representation at another, higher level. Two examples of knowledge expressed in Reddy's hierarchy are "sky is blue" (structure/feature) and "city scenes involve vertically oriented rectangular shapes" (environment/segment). Note that knowledge may transcend several levels, and the logical structure of the scene does not dictate the processing sequence. A good example of this is the acknowledged interdependence of segmentation and interpretation, since the significance of pictorial entities can only be evaluated in the broader context of interpretation. In fact, it no longer seems appropriate to view segmentation and interpretation as distinct processes.

How Are Knowledge and Descriptions Represented?

A number of issues and nonissues have arisen over the representation of knowledge in general, not just visual knowledge. It seems desirable to have a redundant representation, on the grounds that data organized in only one way may be easily accessed for some queries but very awkwardly structured for others. To a large extent, however, this is a nonissue, a detail of implementation rather than fundamental science. The ability to manipulate easily the representation to extract implicit information as required and the necessity for cross-links so that relevant related information is readily available when needed are more important issues.

The nonissue of procedural versus assertional representation has been aired in domains other than the vision domain.[17,18] We simply remark here that one man's procedure is another man's data: An interpreter, a theorem prover, and a CPU all operate on data (assertions) or all run procedures. What is at stake are the practical issues of space and time. To the seeker of truth they are only important insofar as they affect his search. Compiling (procedural embedding) the knowledge makes running faster, at the expense of making modification slower. It seems that natural vision systems involve some compilation of frequently used processes and retain the fundamental ability to work from scratch using a reasoning ability: unfamiliar objects take some time to perceive until they become familiar.

A similar nonissue is that of iconic versus symbolic representation. Should the information be encoded by symbols that stand for generalized stereotypes or should it stand for itself? It is only by abstracting from the current particular case that general principles may be invoked, so symbols are clearly vital. Because symbols lose touch with details of reality, however, the original data must be maintained in some form to handle questions whose answers are not embodied in the original abstraction. What is very much an open question is the relation between symbols and images. It may be that one should convert to symbolic representation at a very early stage.[19] The evidence concerning human representation is mixed. Much of our visual imagery is in terms of stereotypes, rather than the original images, as though we reconstruct images from symbolic descriptions and associated iconic examples.[20] However, we are quite good at remembering details, which hints against highly abstracted symbolic descriptions.

Elementary considerations of flexibility in the use and acquisition of knowledge dictate that knowledge, whatever its form, be it procedural, assertional, or iconic, should be made explicit rather than buried in the code and forgotten.

How Should the Knowledge Be Used?

Because of noisy data and constraints that involve likelihoods rather than certainties, vision is not a purely deductive problem with a unique, "correct" solution. It requires determining a "best" interpretation of the data in the light of the available knowledge. The situation is exacerbated by the quantities of data and knowledge, by the fact that the knowledge sources are error prone and not unanimous in their recommendations, and by the strong interaction among knowledge sources. (This interaction is manifest, for example, in the ways in which segmentation and interpretation are inextricably interwined.) These considerations preclude a straightforward search for the best solution; ways must be found to circumvent the horrendous combinatorics.

Naive backtracking, that is, returning to a previous choice point and taking a different alternative, is utterly hopeless. It appears vital to take advantage of local independence of parts of the picture and undo only the direct consequences of an erroneous choice. (Perhaps vision has something important to contribute to problem-solving here.[21]) It is even better to avoid redundant work in the first place by eliminating possibilities as soon as possible rather than eliminating them many times over on different branches of the search tree. Waltz's filtering algorithm provides a dramatic illustration of this point. A generalization of this principle for optimization problems is to make locally

best choices which, if based on enough information, can be expected to lead to a near-optimal solution with minimal search.

The assessment of local evidence requires integrating many sources of knowledge, of different types and levels, which very often yield contradictory evidence. For example, two adjacent patches of a picture may have the same color, which suggests they belong to the same object, but may have different brightnesses, which suggests they do not; or, higher-level knowledge about shadows indicates that the patches are from an object with a shadow across it, but perhaps knowledge of illumination indicates that the shadow could not be cast in such a position, unless light had been reflected from some surface.

It would appear that some sort of consensus mechanism is necessary, an environment where knowledge sources can compete and cooperate to yield a result upon which they can largely agree. The relative weightings attached to evidence from different sources of knowledge depend dynamically on the current processing context and ultimately on the data, the expectations, and the goals.

Some of these principles are illustrated in a working program, MSYS, recently developed at SRI, for assigning interpretations to regions in previously segmented office scenes.[22] In MSYS, the problem was to find the set of region interpretations having the highest joint likelihood. Each possible interpretation is represented by a process which computes its global likelihood based on local region attributes and the likelihoods of other contextually related interpretations. These processes interact, compete, and cooperate, adjusting likelihoods and eventually achieving a state of equilibrium that corresponds to the best solution consistent with the evidence. The equilibrium state is by its very nature reversible and processes can be defined so that information is never lost, but can be recomputed if the consensus demands it. Because processes communicate via a global repository of data, it is possible to add or remove them (dynamically if desired) with incremental changes in system performance. Hence the system can readily be modified or extended in a modular way.

The use of an equilibrium model eliminates some issues of sequencing, but does not completely sidestep the need to strike a balance between data-driven and goal-driven organizations; the relative importance of observations and expectations still depends on goals. Data driving has the advantage of being less likely to get out of touch with reality but the disadvantage of considering too many possibilities. Goal driving considers only relevant possibilities, at the expense of ignoring the unexpected. Evidently some comination of the two organizations is necessary, such as a hypothesis-verify paradigm, but the best organization is still in doubt.

The organization issue also arises in activation of relevant knowledge and hypotheses within the network of equilibrium processes. As a practical matter selection and activation must be done dynamically as relevance emerges.

How Should Knowledge Be Acquired?

The substantial amount of world knowledge required for perception is most reasonably acquired incrementally as deficiencies are discovered, instead of exhaustively beforehand in a case analysis. Certainly as a matter of experimental convenience, incremental improvement ranks high. Humans certainly increment their visual skills, learning finer discriminations or broader generalizations as circumstances warrant.

Whether knowledge is acquired by supervised or unsupervised learning is not at issue here. As a practical matter, in the near term, it will be important to have good interactive facilities for empirically determining perceptual concepts and processes and communicating them to a machine.[23]

What Have We Learned?

Perhaps the most important thing we have learned from our experiences in vision research is that vision is hard. We hope we have learned why it is hard: All the key AI issues crop up--world models, representation, organization, reasoning, problem solving, even manipulation. The full range of reasoning, from simple association to complex deductions, with uncertain evidence and uncertain rules, is involved. The organizational problems of handling the large amounts of data and knowledge that arise with real pictures rule out toy solutions.

The limited abilities of current scene-analysis systems can be traced to compromises necessitated by the magnitude of the vision task.

The quantities of data are relatively huge but are still much less than is desirable. Many seemingly insurmountable problems encountered in processing $64 \times 64 \times 4$-bit pictures are greatly eased (though not necessarily eliminated) by increasing resolution and number of brightness levels, particularly if the digitization is performed with something other than a noisy television camera.

The knowledge possessed by existing systems is very limited. There are several blocks-world systems that specialize in some details of

the problem, for example, Waltz's catalogue and Shirai's block-finder, but there is no unified system that contains and uses all we currently know about the blocks domain.

The knowledge that systems have is frequently rigidly embedded in procedures which use knowledge in a limited way. For example, knowledge used passively, as a recognition test, may not be available for active use in searching for the object.[24]

Finally, despite 10 years of research, there is still a feeling of fundamental ignorance, of a few ad hoc successes which reflect little true understanding of visual perception. The field remains more Art than Science.

What Should Be Done?

Work in scene analysis is beginning to be directed toward some of the issues raised above. We, and others, are dissatisfied with the low-level abilities of existing scene-analysis systems, which seen overspecialized to particular domains. Because of the usual dependence of these systems upon one characteristic, to the exclusion of others, their performance is marginal and brittle; they are weak in forming descriptions of novel situations, which hampers learning; and they must be torn apart and rebuilt for each new domain. We greatly need a better foundation on which to build the higher functions of scene-analysis systems that are experts in their chosen domains.

In this prevailing climate, we believe the time is ripe for developing broadly based scene-analysis systems. By this we mean systems that have a great deal of low-level knowledge about similarity, continuity, surfaces, reflectivity, shading, color, illumination, occlusion, shadows, and so forth, that is common to many domains. Such systems would be able to make a good attempt at segmenting and interpreting (in simple terms) an arbitrary scene. Human beings can almost always make some sense of a scene, even when it comes from a domain with which they are unfamiliar, for example, photomicrographs or abstract art. They can at least identify significant lines and regions even when the evidence for them is quite weak, and despite their apparent lack of meaning. No existing machine can approach human performance in such situations.

We need a unified framework for experimenting with both general and domain-specific knowledge in scene interpretation, where the quality of the result can be progressively improved by incorporating additional knowledge. With low-level knowledge alone, the system should manifest behavior consistent with the Gestalt laws of human vision; augmented

by domain-specific knowledge, it should rival the competence of the best, current, special-purpose systems and be more readily extensible and considerably more robust.

The proper organization of such a broadly based system of many conflicting sources of knowledge is a key issue to which we have already devoted some attention above. This issue has also received considerable attention in the literature[14,25] as well as in other areas of AI, notably speech understanding.[26] A concensus has begun to form on some specific design criteria. Systems should be data driven; all knowledge should be modularly represented in a shared global data base; and competing hypotheses should be explicitly represented and freely available in the data base, rather than hidden in the internal backtracking variables.

This type of organization, as exemplified by MSYS and HEARSAY, has so far been used only with domain-specific knowledge for interpretation of segmented scenes. We believe it could be used with more general knowledge to assign consistent low-level interpretations of, for example, color, relative depth, orientation, and so forth, to edge fragments and small sets of picture elements, and thereby to integrate segmentation with interpretation.

In addition to the specific suggestions proposed above, we offer a few more general suggestions based on lessons of the past decade:

(1) Researchers concerned with the scientific issues of vision in general should acknowledge that the prodigious quantities of data and knowledge preclude any hope of near real-time operation on current serial computers. Researchers should therefore avoid the preoccupation with efficiency that characterized the ingenious but special-purpose solutions described earlier.

(2) A corollary of Suggestion 1 is to work always with the highest quality data that can be obtained. In the past, many researchers settled for degraded, low-resolution images which fit conveniently in core, and rationalized their acceptance with arguments that this data could be correctly perceived by humans. However, the elimination of local detail (which might allow local, rather than global interpretation), is probably making the problem unnecessarily difficult. Humans can cope with such bad pictures but perhaps it is premature to expect machines to do so.

(3) All assumptions used in the research should be carefully documented so that the work can be meaningfully evaluated,

both with respect to its own objectives and in comparison to related work with similar objectives. Examples of assumptions would include the quality of data, the types and quantities of knowledge required, processor requirements, whether the overall objective was general purpose or special system, as well as detailed assumptions about the types of objects and events in the scene domain.

Scientific research depends upon reproducibility of results: programs should therefore be much more widely exchanged and run. This demands a higher standard of documentation.

(4) Libraries of standardized scenes should facilitate realistic comparisons of alternative techniques.

(5) Work should proceed simultaneously in several pictorial domains to force generality and provide opportunities for cross-fertilization.

(6) Motion perception has been seriously neglected, despite its paramount importance in human perception. Sequences of pictures should be processed to explore the constraint of continuity.

(7) Systems should be designed to assimilate advice or new knowledge provided incrementally by the experimenter to prevent recurrence of errors, preferably via pictorial examples. It is more natural to designate pictorial concepts by pointing with a cursor than by editing a program.

III CONCLUSION

We conclude by reiterating some of the major premises underlying this paper:

- The more knowledge, the better.
- The more data, the better.
- Vision is a gigantic optimization problem.
- Segmentation is low-level interpretation using general knowledge.
- Knowledge is incrementally acquired.
- Research should pursue Truth, not Efficiency.

A further decade will determine our skill as visionaries ....

IV ACKNOWLEDGMENTS

The authors wish to acknowledge the collaboration of Martin A. Fischler during the formative stages of this paper. Dick Duda and Mike Brady provided many helpful comments.

Preparation of this paper was supported in part by the Office of Naval Research under contract N00014-71-C-0294. The research upon which this paper was based was supported in part by the Advanced Research Projects Agency under contract DAHC04-72-C-0008 and by the National Aeronautics and Space Administration under contract NASW-2086.

REFERENCES

1. L. G. Roberts, "Machine Perception of Three-Dimensional Solids," Technical Report 315, MIT Lincoln Laboratories, Lexington, Massachusetts (May 1963). Also in Optical and Electro-Optical Information Processing, J. T. Tippett et al., eds. (MIT Press, Cambridge, Massachusetts, 1965).

2. G. Falk, "Interpretation of Imperfect Line Data as a Three-Dimensional Scene," Artificial Intelligence, Vol. 3, No. 2, pp. 101-144 (Summer 1972).

3. Y. Shirai, "A Context Sensitive Line Finder for Recognition of Polyhedra," Artificial Intelligence, Vol. 4, No. 2, pp. 95-119 (Summer 1973).

4. D. G. Waltz, "Generating Semantic Descriptions from Drawings of Scenes with Shadows," AI Technical Report 271, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts (August 1972).

5. G. J. Agin and T. O. Binford, "Computer Description of Curved Objects," Advanced Papers of the Third International Conference on Artificial Intelligence, Stanford University, Stanford, California (August 1973).

6. B.K.P. Horn, "Shape from Shading: a Method for Obtaining the Shape of a Smooth Opaque Object from One View," MAC Technical Report 79, Project MAC, Massachusetts Institute of Technology, Cambridge, Massachusetts (November 1970).

7. K. J. Turner, "Computer Perception of Curved Objects Using a Television Camera," Ph.D. thesis, School of Artificial Intelligence, Edinburgh University, Edinburgh, Scotland (November 1974).

8. M. Kelly, "Visual Identification of People by Computer," Memo AI-130, Computer Science Department, Stanford Univeristy, Stanford, California (July 1970).

9. M. A. Fischler and R. A. Elschlager, "The Representation and Matching of Pictorial Structures," IEEE Trans. Comput., Vol. C-22, No. 1, pp. 67-92 (January 1973).

10. T. D. Garvey and J. M. Tenenbaum, "On the Automatic Generation of Programs for Locating Objects in Office Scenes," Proceedings of the Second International Joint Conference on Pattern Recognition, pp. 162-168, Copenhagen (August 1974).

11. R. Bajcsy and L. I. Lieberman, "Computer Description of Real Outdoor Scenes," Proceedings of the Second International Joint Conference on Recognition, pp. 174-179, Copenhagen (August 1974).

12. Y. Yakimovsky and J. A. Feldman, "A Semantics-Based Decision Theory Region Analyzer," Proceedings of the Third International Conference on Artificial Intelligence, Stanford University, Stanford, California (August 1973).

13. R. Ohlander, "Analysis of Natural Scenes," Ph.D. thesis, Computer Science Department, Carnegie-Mellon University, Pittsburgh, Pennsylvania (April 1975).

14. P. H. Winston, "The MIT Robot," in Machine Intelligence, Vol. 7, pp. 431-463, B. Meltzer and D. Michie, eds. (Edinburgh University Press, Edinburgh, and American Elsevier Publishing Company, 1972).

15. J. M. Tenenbaum, "On Locating Objects by their Distinguishing Features in Multi-Sensory Images," Computer Graphics and Image Processing, Vol. 2, Nos. 3 and 4 (December 1973).

16. A. Newell and D. R. Reddy, "Image Understanding: Some Notes," Minutes of the ARPA Image Understanding Workshop, Science Applications Incorporated (March 1975).

17. T. Winograd, "Five Lectures on Artificial Intelligence," AI Memo 246, Stanford University Artificial Intelligence Laboratory, Stanford, California (September 1974).

18. G. J. Sussman, "A Computational Model of Skill Acquisition," AI Technical Report 297, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts (August 1973).

19 D. Marr, "On the Purpose of Low-Level Vision," AI Memo 324, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts (December 1974).

20. D. A. Norman and D. E. Rumelhart, Explorations in Cognition (Freeman and Company, San Francisco, California 1975).

21. M. Minsky, "A Framework for Representing Knowledge," in The Psychology of Computer Vision, P. H. Winston, ed., pp. 211-277 (McGraw-Hill Book Company, New York, New York, 1975).

22. J. M. Tenenbaum and H. G. Barrow, "MSYS: A System for Reasoning about Scenes," SRI Artificial Intelligence Center Technical Report, Stanford Research Institute, Menlo Park, California (to be published).

23. J. M. Tenenbaum et al., "An Interactive Facility for Scene Analysis Research," SRI Artificial Intelligence Center Technical Note 84, Stanford Research Institute, Menlo Park, California (September 1973).

24. E. C. Freuder, "Active Knowledge," Vision Flash 53, Artificial Intelligence Laboratory, Massachusetts Institute of Technology Cambridge, Massachusetts (October 1973).

25. S. E. Fahlman, "A Hypothesis-Frame System for Recognition Problems," Working Paper 57, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts (December 1973).

26. L. Erman and V. Lesser, "A Multi-Level Organization for Problem Solving Using Many Diverse, Cooperating Sources of Knowledge," Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Pennsylvania (March 1975).