Recognition of Real Objects

VISION FLASH 33

by

Eugene C. Freuder

Massachusetts Institute of Technology

Artificial Intelligence Laboratory

Robotics Section

October 1972

Abstract

High level semantic knowledge will be employed in the development of a machine vision program flexible enough to deal with a class of "everyday objects" in varied environments.

This report is in the nature of a thesis proposal for future work.

Work reported herein was conducted at the Artificial Intelligence Laboratory, a Massachusetts Institute of Technology research program supported in part by the Advanced Research Projects Agency of the Department of Defense and monitored by the Office of Naval Research under Contract Number N00014-70-A-0362-003.

This memo is located in TJ6-able form on file VIS;VF33 >.

Forward

This material is not presented as a discovery but rather as a journey on the path laid down by Professors Minsky, Papert and Winston in Vision. Please ignore any dogmatic tone which may appear in the efforts to overcome my natural tendency to the opposite "I think, perhaps, maybe..." school of rhetoric.

Abstract

A proposed program of research in Machine Vision is described. The first section, the Scenario, lays out succinctly and informally the aims of the project, summarizes the background of the work, the problems to faced and the approach to be taken. An extended description of the Problem follows, its background and its formulation for the present work. We next discuss the Artificial Intelligence Issues involved, and the Approach we will take. Benchmarks are laid down for specific progress. Our research interests are once again carefully formulated in the Summary.

Scenario

Picture the following scenario.

We walk into the street, and hijack the first passerby. We have him go home, retrieve his toolbox and return to the AI Robot. He grabs a handful of tools and dumps them in front of the eye. The Robot identifies these objects.

The key element here is the flexibility to deal with real objects in real situations (surreal perhaps in this instance.)

The <u>flexibility</u> to deal with <u>reality</u>. This will be the basic aim of our research.

Why has previous work in visual recognition often been sadly lacking in this regard? Why does this effort have a better chance of success?

Consider first some recent contributions.

1. Minsky and Papert--The heterarchical approach essential to the flexibility sought (4,5,6,7).

2. Winston--Network structure, data driven control, basic structural description, learning by failure, grouping, quantitative analysis of relational concepts (11,12,13).

3. Shirai--A.I., and vision research in particular, as an <u>experimental</u> science, as demonstrated most recently by Shirai(8).

4. Waltz--The extent to which completeness of description can replace flexibility of control (9).

5. Charniak--The role and extent of knowledge in understanding (1).

6. Winograd--Semantic/syntactic integration. Frontal attack on

global problem; large system structure. Visual/linguistic analogies (10).

7. Hewitt--The power of a problem oriented control structure. The ease of a pattern directed data base. Two way flow of control; flexibility of non destructive failure (3).

As a framework for these developments, we have the Vision System. This has provided the experience and facilities for a broad based background in vision research. (Complemented by the more formal seminar structure.) Plus the specific technical and organizational results of the System.

Perhaps most importantly we want to deal specifically with the problem suggested in the first paragraph. This is very different from most previous efforts at visual recognition.

Often the descriptive problem has been dealt with separately as a pattern recognition issue. When real visual input has been used, the aim has been generally to transform this input, with a single low level predicate, into a "preprocessed" data base. Features could then be extracted from this data base and dealt with as a pattern recognition or tree search problem.

In fact, we maintain that there is no real problem, on the macroscopic level, of distinguishing a respectable set of real objects. In some respects, the more complex and unusual the objects the better. It should not be at all difficult to "separate these objects in parameter space" with any number of redundantly distinguishing features.

The problem is to get some of these gross features:

a. from amid all this redundant mass of information

b. in particular, from the mass of "meaningless" low level data

c. in differing:

1. individual samples of objects (variations in size, shape, orientation, texture, etc.)

2. scene contexts (occlusions, shadows, etc.).

On a local level these changing conditions create great differences. If we deal only from the "bottom up" we cannot hope to deal with such chaotic variation.

However with higher level interaction we can hope for global guidance at levels that can deal with such variation. The old saw: "It's easier to find something if you know what you're looking for". We need to know what we are looking for to find the features.

I think my goals can be approached in two stages.

I.

First I will pick a single object of some semantic richness, e.g. a hammer. I will then write the world's greatest hammer recognizer. This is not as trivial as it may sound.

The program will needless to say not merely reply "hammer" when faced with any object, on the grounds that hammer is the only object it knows about. It will have to decide whether the object is a hammer or not a hammer.

It should be able to make this decision about your roommate's favorite hammer (or screwdriver). It should be able to decide after a stranger is allowed to place the object in the field of view and arrange

the lights. It should decide correctly when visiting Cub Scouts come in and are allowed to empty their pockets over and around the object to be studied.

This is a hard problem.

II.

Assuming Part I is not a thesis in itself, the next stage will be to integrate knowledge of several objects into a system. Hopefully the experience of part I will provide principles for encoding visual knowledge that will facilitate adding the initial ability to deal with several new objects.

We will want to design the analytical structure to take advantage of the hypothesis and verification techniques developed in Part I. Note that the problem domain of Part I encourages us to explore the practical implications of our philosphical bias: application of specific higher level knowledge.

Problem

Machine Vision involves both "low level" image processing and "high level" descriptive analysis. The failure of Machine Vision to date has been to adequately link these levels.

Much of the work in this field has restricted itself rather strongly to one or the other of these levels. There are several reasons for this. In the beginning there was a tendency for image processing people to believe they could "do it all" with a clever enough set of Fourier Transforms. The descriptive analysts, on the other hand, tended to underestimate the problem of preprocessing a suitable data base for their descriptive schemes.

Problems in each of these areas have been formidable enough to perhaps demand single minded concentration. But now that a groundwork has been laid we are in a position to take a more comprehensive approach. It has in fact become clear that such an approach is mandatory for continued progress in Machine Vision (4-7,11-13). The pressures for thesis work at a single level of the problem have begun to lead away from the overall goal. Search for the most sensitive line finder sidesteps the natural context in which the problem is to decide which lines to ignore, not which obscure lines to find. Yet work continues on idealized "picture grammars" which assume the preprocessor has given them just the data base they need (without any knowledge of those needs).

Part of the reason that work on these two aspects of the vision problem has often remained unnaturally isolated must be the distinction

in interests and skills required for work in the two fields. We therefore think it might be advantageous for two students to pool their experience in these separate areas. In any event a "pincer" attack is called for specifically on the interface between high level and low level analysis.

This does not mean regarding present accomplishments in these areas as "black boxes" and wiring them together. Rather, processes at each level must be organized with the requisite cooperation in mind.

Seing at anything beyond the feature point level is a matter of organizing visual information according to some visual model of the world. In the past we arbitrarily used only the most elemental physical units of the model to process the visual data--the most basic physical regularities and anomalies such as lines and points. The rest of the higher level knowledge was stored separately to be "matched" against data that could hopefully be organized into meaningful structures without knowing what structures were meaningful. This knowledge too must be part of the visual machinery that organizes the raw visual input and thus "sees". Instead this information has been so independent of any inherently visual process that researchers have been able to point pridefully to their ability to perceive and program this "recognition" apparatus as instances of highly general abstract mathematical models.

In point of fact, higher level context assumptions, a parallelipiped environment, for example, have long been implicit, or even explicit, in much of our work. Experience has shown the utter necessity for such guidance in order to make sense out of a chaotic sensory reality. It is

time to fully integrate our higher level processing into the organization of our visual data.

We have seen how difficult it is to obtain even a line drawing of a cube, without suspecting that we are looking at a cube. How can we expect to obtain some idealized data representation of a horse and then recognize that representation as a horse. We cannot choose the ideal descriptive model of a horse analytically and expect some syntactic predicate to produce this description for us.

The conclusion we come to is the need to eliminate the old distinctions between preprocessing, description and recognition. One cannot obtain a preprocessed result which is then analyzed to produce a description which is then matched with a model to be recognized. Rather seeing is a homogenized process, the "model" is built into the structure of the processor, the description of a data set is the protocol of its processing. Recognition is coincident with description.

Recognition is not a matching of templates against a processed data structure. Rather it is a many layered silk screen process that begins with the raw input and leads to the richly patterned perceptual world we seek.

Issues

We feel that an approach which is predicated on the intimate interaction of descriptive apparatus and descriptor predicates is the best hope for practical progress in Machine Vision. Beyond that we believe that such a project would provide an appropriate laboratory for the illumination of several current A.I. conceptual concerns.

1. Heterarchy

Problems in reconciling low level results with high level idealizations, even in a very simplified domain, provided one of the early motivations for the heterarchy concept (4-7,12). Return of recognition level advice to the preprocessor is perhaps the most salient example of heterarchy practice (4). It is significant that, except on a most general level, even this case study of heterarchy has not been implemented and its implications certainly not fully explored.

Our investigation of heterarchy in vision would emphasize the crucial role of recognition level knowledge.

We do not analyze heterarchy in terms of interaction or advice between major discrete processing modules or stages. We do not view "high level" knowledge as criticizing the descriptive structure resulting from low level "preprocessing". Rather we believe that in general no useful structure can be derived without high level knowledge directly involved in the construction process from the beginning.

In any case, a working fully ramified case study should provide useful insight into the theory and practice of heterarchical interaction.

In particular:

2. Sensory/perceptual systems.

The heterarchical interaction between low level sensory systems and high level perceptual systems is of particular interest to A.I. research at present. Having worked "down" from chessplaying and integration, A.I. is now facing the far more obcure problem of enabling machines to duplicate the essential "automatic" processing of real world sensory data. Techniques and processes abstracted from our results in melding low level and high level visual processing should prove relevant when we provide our robot with other sensory apparatus. The possibility of higher level semantic intervention in the auditory analysis of speech, for example, is already a live issue.

3. Knowledge--"the size of infinity"

In developing a system that can deal with a real world context we will face certain problems analogous to those faced by Charniak in his work on understanding childrens' stories. Previous successes in machine vision, or even machine pseudo-vision, i.e. analyses of hypothetical input, have dealt largely with a severely restricted domain. The recognition set has been either highly stylized or simply very small. There is an analogy here to syntactic uni-directional theorem proving systems, which break down on a non-restricted data base. First we will have to get a practical idea of what the size of infinity is in the visual domain, and then we will need techniques for organizing it. In particular:

4. Knowledge--as procedures

Our knowledge--about shapes as well as subjects--will be organized as procedures. This is more than a fashionable device in a system where knowledge will often consist of knowing the right questions to ask of some other module, preprocessing modules in particular. ("Preprocessing" is obviously a misnomer here, reflecting an organizational mode we hope to replace.)

Recognition will not consist of preparing a description of an input scene and matching it with a description of a model. Rather description and recognition will occur simultaneously during the processing of the scene. A description will be essentially the protocol of a process, not a result.

5. Parallel processing

If parallel processing techniques are to prove significant, this would certainly seem to be an area in which they might demonstrate their value. We envision a system in which processing would proceed simultaneously on visual subunits or on conceptual subprocesses. Proceedings would be dependent continually on the results of intermodule communications and questions, both as to the results being attained in different "higher" level investigations, and the results of additional low level processing. It would seem plausible that dialogues of the following style could take place profitably:

a: I think these four things are legs. Is anybody looking at anything above them all?

b: Yes, I am.

a: Can you tell me if it's planar or bumpy?

b: Not yet.

a: Well, look, work on it; take our time and wake us when you have something.

a: (Yawn) Wazzat? Bumps huh? Probably an animal. We'll activate hoof, paw and knee searches. You send your stuff to an animal body parser.

c: I can't find and subdivisions on this thing for head or tail. Can anybody make head or tail out of any nearby blobs of relative size x and position n?

.

.

.

This particular dialogue may be vacuous. investigation of many possible dialogues may reveal some essential usefulness of this type of organization.

6. Grouping

As even the brief dialogue abouve indicated, grouping problems will be central. Guzman like techniques will not suffice, particularly for higher level organization. Rather than very sensitive local predicates capable of distinguishing fine variations, we will require broader based cruder predicates capable of determining relative homogeneity. Contextual, conceptual cues will be needed to distinguish overlapping objects from functional groups. A program that knows about collars will not dismayed to find a dog's head separated from its body; it will be capable of seeking the "severed head". Alternative hypotheses will be

explored for conjoining parts of the picture, ignoring occlusions, or melding sequences into single units. The same subscene may be viewed in each of these ways, as linked units, overlapping distinct units, single textured units, successful interpretation directing the choice.

7. Organization

Direction will be available from both the lower level and the upper level. That is knowledge about a shape, as well as an object or class of objects, will be embedded in procedures which in turn direct the flow of processing through other procedures. This flow will not be locked into any piece by piece, or even level by level, decision tree or network search. The system can skip to high level hypotheses, plan an approach on that basis, fail, review what it found in the process and use this as a basis for some more ground up investigation. Details can be verified on a hoof before during or after the search for a head, whatever is expedient.

Approach

We expect our progress on this project to reflect the structure of the results. That is both will proceed by a process of "successive approximation".

Professor Papert has characterized one approach to problem solving as "neglect all complication, try something". The results may serve as a "plan" on which to base further refinements or a basis for "debugging".

There is a distinction between simplifying the problem and simplifying the solution. One technique involves splitting the problem into a number of separate or succesive simplified problems. E.g. ignore shadows, texture, etc., just consider ideally lighted ideal planes. Then consider shadows. Then consider texture, but no shadows. Etc. This approach can be very fruitful. However, it may leave us far from a solution to our total problem, particularly if the various aspects of the problem are related in a non-linear fashion. In this case it may prove necessary at times to approach the total problem with a simplified solution. We can then rework this solution successively to approach a more adequate solution. The intermediate results serve to indicate directions for improvement. We may "project" structural aspects of the problem from the solution, rather than relying on an a priori division of the problem.

We expect a continuing dialogue during the research that reflects and suggests the dialogue that will be built into the system:

>Is that side straight?

>I can't tell; it peters out in spots.

>Oh really? Maybe they're lined up wrinkles. Are there more of them?

>Could be.

>Can you characterize their profile?

>Yes, I can give you a "possible wrinkle" predicate; but that predicate will also work on soft corners.

>Put the surrounding planes for a wrinkle will be similar?

>Yes, and aligned shadows often occur with wrinkles.

>Now can you ...

>No but I can ...

>Well then if I tell you ..., then could you ...

>...

The important thing is that many of the technical realities of the system will flow from real experience with the problems of processing real data. This does not mean that the results will lack theoretical content, but that the theory will bear a valid relationship to the reality of the vision problem. We want the theory to flow from and reflect the structure of visual experience, rather than attempt to distort visual data to conform to preconceived and inappropriate analytic theories.

Benchmarks

We might proceed in several phases. A possible sequence would be as follows.

As a first stage exercise, we could consider the development of a system that dealt with simple geometric models, e.g. cube, cylinder. This system would provide experience in melding high level knowledge with a suitably varied set of low level predicates. The idea is not to push any one predicate or approach to its limit but to allow the extent of higher level knowledge and variety of low level approaches to work back and forth, to zero in on the correct perception.

A brief example, a simple cube. We avoid working very hard to find the precise edges from a feature point analysis. Rather we obtain some rough regions with a crude homogeneity predicate (that averages out not only surface noise but textures). We massage these a bit and get a rough idea of their shape, if we suspect a parallelipiped, we may apply a finer test to verify this shape. We find enough to guess a cube or at least a planar object. We then apply a crude line verifier over the wide band between regions. Meeting success we may home in on the lines with a sensitive line verifier in a narrow region. (If some regions were lumped into one at the start we use our models to guide predicates in a parsing attempt (4).)

This example already illustrates several interesting points. Unlike a pass oriented system, we do not have to apply all our predicates at once, but only as (and where) needed. We make use of broad based, region

oriented predicates to avoid the problems of high sensitivity until we have some hypotheses to guide the search for finer features. By working down from the general to the specific we avoid losing the forest for the weeds. And even at the lowest level we are still dealing with a broad based predicate, a line verifier. We also avoid an initial commitment to a world model of straight line planar objects.

The next stage would provide experience in dealing with a complex of surfaces in non linear terms. A limited set of "real" objects, a set of tools for example, would provide a miniworld. The second stage system would be able to function in this world. Here we would have to incorporate a greater variety of data types and predicates or procedures.

Another simple example, a hammer. We find an initial region. We suspect and verify a roughly rectangular shape. Relative length versus width prompts a handle hypothesis (or vice versa). This initiates searches for regions at either end, either crosswise or colinear to the handle axis (we could have a screwdriver). We find a chunky region crosswise at one end. We hypothesize a hammer with handle and head. We move back down to verify a few fine details to assure our success.

Notice here some further principles emerging. Precisely what is required to see a hammer head varies depending on where or from what direction you "enter" the observation. Think, for example, how carefully you would have to draw a hammer head to make it identifiable in the following contexts: at the end of a roughly drawn hammer handle, standing alone by itself, at the end of a carefully drawn hammer handle, in a picture of the head alone to be viewed only partially and in

isolated pieces. The requirements are different if we move "down" from a knowledge, or hypothesis, of "something at the end of a hammer", or if we move "up" from a detailed picture of the contour of a piece of the head. Generally we have a range of redundant information that specifies an object; any particular set of details, e.g. of shape, are not always required for identification, and may be easier to verify than to obtain by "preprocessing".

A flexibly programmed "understanding" of hammer heads should be able to be "entered" at several levels. Processing should proceed in parallel, with mutual calls, modified by the current state of knowledge.

It may be appropriate here to initiate a Charniak-type study of "everything we need to know" about, e.g. hammers, in order to perceive them in varied contexts. Why should such extensive knowledge not be necessary to "understand" in visual contexts, just as it is required in verbal contexts?

Another miniworld would be established to provide experience in dealing with classes of objects. A set of saws, for example, or a set of doll house furniture, could form such such a class. The third stage effort would provide a system that could move easily through the perceptual space of this class.

Some of Winograd's ideas on organizing knowledge might be useful here. A systemic grammar might be a useful metaphor for at least one aspect of the organization. Also some convenient method of inputting knowledge would be useful at this point. Ideally the system would be such that someone could eventually hook up the output of a Winston

learning program, to the input of this system. (And the output of this system to the input of a Winston learning program, of course.)

Finally a rich miniworld would be chosen to combine our previous experience in a more varied environment. A cutaway doll house, or a multipurpose workbench, for example. By the completion of this last stage our general principles for an integrated perceptual system should have been demonstrated, and new principles for implementing this integration should have emerged.

This is rather an ambitious program. We could only hope to lay the groundwork really, to build an instructive testimonial to the possibilities derivable from a proper conjunction of high and low level knowledge.

Summary

In summary, in order to make a quantum jump in vision research, a number of bold steps are required. The thinking of preeminent theoreticians in the field, has long tried to push us in these directions. However "practical" considerations have too long held us back.

Instead of studying the parts and then deriving the "glue", we must study the glue and derive the parts.

It has for some time now been recognized that the real problems in vision lie in understanding the cooperation of the various subprocesses. In practice, it has been easier to define and deal with specific pieces of the spectrum of visual knowledge. However, we eventually end up sitting around bemoaning the difficulty of putting the pieces together.

The solution is not simply to learn more about more pieces. Neither is it to treat the pieces we have as "black boxes" to be tied together. Aside from the theoretical arguments that could be mounted against this approach, it has proven rather barren so far in practice. To learn something about the mysterious "binding energies" as it were, we must simply grit our teeth and attack the problem directly. This approach will often mean messy, tentative, ad hoc progress. We must be willing to pay that price. We may utilize our hard won piecemeal knowledge, of course. However, when we have finally won our way to some understanding of the interactive process of vision, we may find ourselves also in a better position to identify the essential subprocesses involved.

Instead of generating data types and predicates to handle an idealized data base we must generate idealized data types and predicates to handle a real data base.

Becoming absolute experts on line drawings will only qualify us as experts in graph theory. The humor of this approach is evident when we consider that "real" physical data is manufactured at great cost and effort to match this idealized data base, and even then the predicates derived for the line drawing model cannot succeed in producing a satisfactory translation of the physical objects into line drawing data types.

Our line drawing studies have provided us with some useful methodologies and results, that should provide guidance and submodules for a more general study. However it is time to return for inspiration and guidance to more realistic data. Not to set up another panacea predicate. But to extract whatever information required to organize the sensory input. Organize not simply and dogmatically as lines or corners, but as recognizable perceptions, as appropriate for the data.

Most fundamentally, the easy, but artificial, distinction between seeing, describing, and recognizing must be broken down. We cannot merely organize the visual input according to some "neutral" data base. We must not partition out part of our knowledge to function as a template to "match" our results. The only way we can hope to organize the visual data to "match" a high level form is to use that high level knowledge to perform the organization. The concept of a "model" as such is outdated. The medium is the message. Recognition is the process of description.

Description is the process of recognition.

Bibliography

AIMIT = Artificial Intelligence Laboratory, M.I.T.

1) Charniak, E., "Toward a Model of Children's Story Comprehension," D.Sc. Dissertation, Department of Electrical Engineering, M.I.T., August, 1972.

2) Freuder, E., "Views on Vision," Vision Flash 5, AIMIT, February, 1971.

3) Hewitt, C., "Description and Theoretical Analysis (Using Schemata) of Planner: A Language for Proving Theorems and Manipulating Models in a Robot," A.I. Technical Report 258, AIMIT, April, 1972.

4) Minsky, M. and Papert, S., "Status Report II: Research on Intelligent Automata," Project Mac, M.I.T., 1967. (Abstracted in "Progress Report IV" Project MAC, M.I.T., 1967.)

5) Minsky, M. and Papert S., "Proposal to ARPA for Research on Artificial Intelligence at MIT 1970-1971," A.I. Memo. 185, AIMIT, December, 1970.

6) Minsky, M. and Papert S., "1968-1969 Progress Report, A.I. Memo. 200, AIMIT.

7) Minsky M. and Papert S., "Progress Report," A.I. Memo. 252, AIMIT, January, 1972.

8) Shirai, Y., "A Heterarchical Program for Recognition of Polyhedra," A.I. Memo. 263, AIMIT, June, 1972.

9) Waltz, D., "Generating Semantic Descriptions from Drawings of Scenes with Shadows," Ph.D. Dissertation, Department of Electrical Engineering,

M.I.T., September, 1972.

10) Winograd, T., "Procedures as a Representation for Data in a Compter Program for Understanding Natural Language," A.I. Technical Report 235, AIMIT, February, 1971.

11) Winston, P. H., "Learning Structural Descriptions from Examples," A.I. Technical Report 231, AIMIT, September, 1970.

12) Winston, P. H., "Heterarchy in the M.I.T. Robot," Vision Flash 8, AIMIT.

13) Winston, P. H., "Summary of Selected Vision Topics," Vision Flash 30, AIMIT, July, 1972.