VISION RESEARCH STRATEGY:
BLACK MAGIC, METAPHORS, MECHANISMS, MINIWORLDS AND MAPS*

Alan K. Mackworth

Department of Computer Science
University of British Columbia
Vancouver, B.C., Canada

Abstract

Machine vision will advance substantially only if it continues to develop a coherent theory. As with all fledgling sciences, the framework for such a paradigm has emerged as a result of restricting the scope of attention to limited but non-sterile domains that serve the current needs of the theory. An example of such a domain is the class of freehand sketches. These occupy a position in vision analogous to that of speech in that they are designed for person-to-person communication and thereby have a rich, conventional semantics which can be exploited. The goals of a project to understand sketches are given. A very brief description of a program, MAPSEE, that interprets sketch maps illustrates the argument. A conservative partial segmentation yields a variety of cues which invoke models that interact according to a uniform control structure: a network consistency algorithm. The necessary deficiences of the segmentation, their effect on the interpretation and using the interpretation to refine the segmentation are all mentioned. This example is used to focus discussion of a variety of vision issues such as the chicken-and-egg problem, the power of descriptive models and their corresponding weaknesses, the incremental nature of constraint methods, cue/model hierarchies, the modularity and generality problems and procedural adequacy. Finally a cyclic theory of perception is used to characterize a variety of vision programs.

1. Strategies for Vision

If we intend to continue developing a science of perception--for that is what I believe we are doing--our research must become self-conscious. We must be aware of our strategies and scholarly in the development of the field. Our efforts must be informed by what has been done, why it has been done and what has been learned. It should be clear that we have a developing paradigm [25,26,2,11,12] and that research goals should be based not on fashion but on the needs of that paradigm. It should also be clear that any science at this early stage must, perforce, close its eyes to almost all the allures and mysteries of nature and choose a highly circumscribed fragment of reality to examine. Indeed, glancing back at the history of science, the fragment chosen is usually not even part of nature as she is but merely an abstracted, stylized slice which can illuminate the murky recesses of current theory. Galileo chose, as his blocks world, bodies sliding down a friction-free inclined plane in a vacuum; Newton considered point masses of infinite density; Chomsky the ideal speaker-hearer's competence. In vision, Roberts [20] realized that the enormous effort being made to solve the problems of pattern classification was contributing little to the theory of machine perception. He then retreated from the "real-world" problems of character recognition to understand his blocks world: black velvet background, matte surfaces carefully lit and all. The decade of research inspired by his decision proved its correctness. The cumulative, puzzle-solving activity of viewing the world through polyhedral spectacles provided a theoretical base and practical support for the seeds of a new vision paradigm.

Perhaps the most important blocks world lesson is that only by patiently teasing out the semantics of a domain (that is, the relationship of representation [6]: the relationship between objects in the world and their pictorial traces) will we be able to write programs which interpret pictures in that domain. So we start by looking at the semantics of pictures.

2. Clean and Dirty Semantics: The Laws of Convention or the Laws of Physics?

It is still instructive to look for parallels between vision and natural language understanding without making a commitment to, say, a linguistic approach to vision or, even further, to the primacy of syntax in both domains or, on the other hand, to an imagery-based approach to language. One can establish an analogy between successions of task areas in the visual and aural domains ranging from perfect line diagrams through free-hand sketches to "real" images of natural scenes in the former and from perfect presegmented text through speech to arbitrary natural sounds in the latter. This admittedly crude analogy depends upon a variety of underlying factors. These include the nature of the representational medium and the presumption of

* The work reported here is supported in part by the National Research Council of Canada's Operating Grant A9281.

perfect segmentation but, primarily, the analogy depends upon the extent to which the laws of convention rather than the laws of physics dictate the relationship of representation.

The analogy demonstrates that vision researchers have largely ignored an area which has been the primary focus for our aural counterparts. We have ignored man-made images designed for person-to-person communication, images whose semantics are fixed by convention, and concentrated on images of natural scenes, images whose semantics are dictated by the laws of optics. Man-made images have, by their very nature and purpose, a rich, clean and useful semantics which can be codified and sensibly exploited. Again, this is not to say that we should discontinue work on recovering the relations ip between incident and reflected light, the nature of surfaces, edges, textures, and shadows and so on; as in speech understanding, progress will require a judicious admixture of both approaches.

3. Freehand Sketches and Maps

A common class of image designed for communication is the free-hand sketch diagram. For several years we have had the ability to draw such diagrams directly on graphical data tablets but this ability has not been heavily exploited. Most uses have been very mechanical and ad hoc. Only rarely [1,18] can a program be truly said to be interpreting the sketch.

In studying images sketched free-hand on a data tablet, this project has many goals. They include:

I) To see if we can broaden the scope of our vision programs by applying the theory developed in the blocks world decade to other domains. At the same time, reworking and extending the theory.

II) To explore the relationship between natural and conventional representations.

III) To determine the extent to which highly domain-specific knowledge can be factored out of

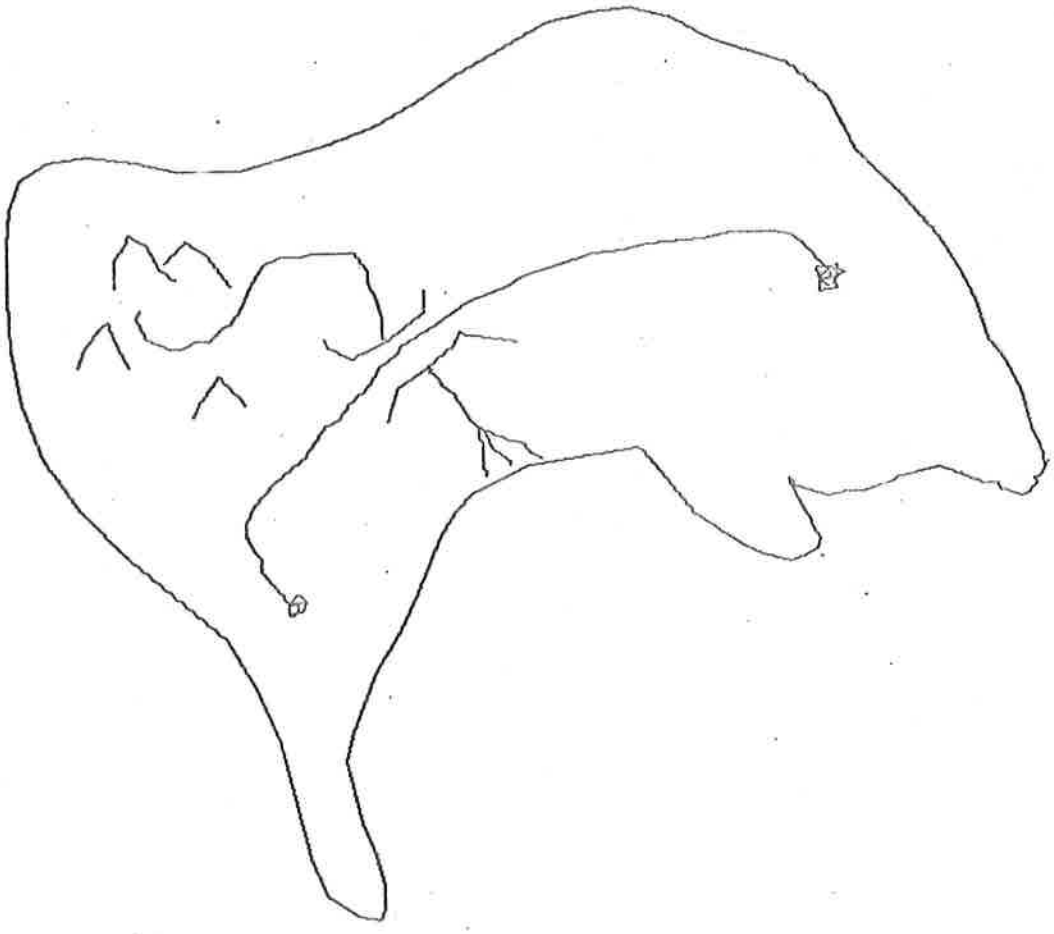


Figure 1. A Sketch Map of an Island

the image interpretation program, to be supplied by the user.

IV) To make available a useful interpretation program for some restricted but important classes of sketches.

V) To provide an experimental vehicle for studying the control structures required to implement schema-based theories of perception.

The initial domain chosen was a set of sketch maps drawn on a data tablet typified by the map shown in Fig. 1.

This is, deliberately, so badly sketched that many people have to be told, before they can see it, that it represents an island on which there is a road that connects two towns and crosses a bridge over a river that rises in a mountain range and ends in a delta.

This domain allows us to explore ways of satisfying the goals listed above. In particular, besides having a satisfying mixture of conventional and direct representations, such maps are related to the work we are doing on understanding LANDSAT (ERTS) images [22] which have primarily optical semantics. In the long run, such understanding would proceed more successfully if programs were able to accept advice, in the form of sketch maps, about the geography underlying the image.

4. How to Interpret Sketch Maps

In describing how to interpret sketch maps, I cannot here do justice to the current program, MAPSEE, an implementation in LISP recently completed. Without giving the details of its operation--these are presented elsewhere [13,14]--I will place the ideas behind MAPSEE in context. Furthermore, the short-range and long-range goals of the project will be distinguished as they are, at times, in apparent opposition. (But then, the tension between them establishes a productive dialectic.)

4.1 Cues and Models

In any world it is crucial to ask: what can various picture fragments depict? Here, it is clear that a line element can, in total isolation represent part of a road, a river, a bridge, a mountainside or a shoreline (of lake or the sea, with water on one side, land on the other or vice versa). An areal element could be land, lake or sea. The design of an interpretation scheme starts with the fact that, as in the blocks world, the enormous ambiguity of interpretation can be progressively reduced by considering picture fragments in wider and wider context. Individual picture fragments, or cues, invoke local models which serve to explain or interpret the immediate locale of the cue that invoked them. These models must talk to each other and agree on the interpretation of picture fragments that they mutually interpret. To discover the first level of model information the following experiment is recommended: cut a small hole in a sheet of paper and move it about Figure 1. Being familiar with the class of maps represented, you would discover a wide variety of informative local picture parts. The point clusters, the chain links (where a chain of line segments joins back on itself), the free ends, the sharp kinks in the chains and the various junctions all contain much interesting but totally ambiguous information. Alternatively, one can say that each part invokes a set of models for its environment. A catalogue of picture parts, known as "primary cues," and their possible models is not given here. Simply, note that each of the many possible interpretations of a primary cue places an interpretation on each of the line and region fragments that comprise the part. As shown in [14] the primary cue interpretation catalogue captures a wide variety of geographic and cartographic inferences.

4.2 Control Structure

If we suppose, for the sake of exposition, that our images were perfectly presegmented line drawings of maps then finding such cues in the picture and searching for a mutually compatible interpretation would be analogous to that process in the blocks world with some important exceptions and extensions. Mackworth [10] presents a series of algorithms designed to instantiate, in given domains, each of a set of variables that must satisfy a set of binary relations. Those algorithms, called there network consistency algorithms as exemplified by Waltz's arc consistency algorithm [24] and Montanari's path consistency algorithm [17], are often better than backtracking for such a task. In Waltz's case the variables or nodes are the junctions, the binary relations or arcs are the lines between the junctions, that is, the network of relations is isomorphic to the line drawing being interpreted. In MAPSEE the variables or nodes are the chains and the regions (which also must be interpreted--everything need not be packed into the chain labels) and the relations, no longer just binary, are generalized to n-ary relations that consist of the primary cue models. The control structure for the interpretation phase of MAPSEE is a new network consistency algorithm, NC. See [10] and [23] among others for other uses of the constraint satisfaction approach.

4.3 Representations

Pictures must have a variety of representations according to the needs of the various components of the task. In MAPSEE there are three: a procedural representation as, for example, originally created by the stylus tracking routines, a network representation of objects, relations between objects and local function definitions [7] and an array representation indexed by x-y coordinates.

Pictorial representations should encourage the use of a level of detail appropriate to the task at hand. Each of the three representations allows that. The primary cues, for example, are found by searching the most appropriate picture structure, exploiting the levels of detail to make the search effort as efficient as possible.

4.4 Conservative Segmentation

The earlier supposition that we have perfectly presegmented maps is totally wrong. The segmentation into chains, regions and the variety of primary cues is not given and cannot be done perfectly by any means. For example, as there can be substantial gaps between lines that were "intended" to meet, the region segmentation is difficult. Indeed, in a real sense, it cannot be done at all until the map has been interpreted! This is one of the many chicken and egg problems of scene analysis: segmentation is interpretation and vice versa[11]. However, an initial partial region segmentation is possible. A quick segmentation, using a Warnock-type algorithm in the tree of space occupation arrays followed by a merge of all adjacent regions can be done. The top-down tree search is stopped well before it could get into trouble, at a level whose resolution size is much greater than any unintentional gaps in the drawing. This guarantees no region leakage. No region so found corresponds to more than one "intended" region. But, of course, an intended region can be segmented into more than one found region. In Fig. 1, the large connected land region is split into three regions: one between the upper mountains and the river, one in the peninsula in the south-west and one consisting of the rest of the island except for the river delta. On the other hand, other intended regions are not represented by any found regions. In Fig. 1, the two small land regions in the river delta are not found.

The essential character of this approach to partial segmentation is its conservatism (or, if you prefer, it follows Marr's principle of least commitment [15]). Two other major aspects of the segmentation process are similarly conservative. In the search for primary cues there are many border-line cases of cue instances. These are all rejected: the criteria are always very tight. We must guarantee that no false cues are found. The obvious price is that many real cues are ignored. Finally, given a cue, it must be fleshed out with the picture fragments corresponding to its various subparts. The search for these in the picture is conservative. In looking for a region associated, in a certain direction, with a primary cue, for example, MAPSEE crawls carefully from a starting point in the given direction. If it finds a region within a very short distance, well and good, but if it doesn't it gives up even though the region may be found by continuing, because if it continued it could pick up the wrong one. If it gives up it creates a region ghost [4] which stands for the region that should be there but hasn't yet been found.

Thus, there are four classes of discrepancy between the partial segmentation and the segmentation intended by the user: the missing cues, the region ghosts, the missing regions and the extra regions. The effect of each of these discrepancies on the interpretation process is unique, but they have in common the vital property that they can not cause the elimination of interpretations that would remain if the segmentation were perfect. This is the true sense of the word "conservative" that has been used to characterize all aspects of this segmentation.

The missing cues have no serious effect on the consistency process, provided, of course, that sufficient remain. A missing cue simply fails to supply its extra constraints on the possible interpretations of the chains and regions. In this domain, however, there is such a welter of cues invoking consistent models that there is a multitude of partially independent but mutually confirming inference paths. Breaking a few of those inference paths causes no degredation in the interpretation. It is tempting to postulate that most perceptual tasks, in the real world as opposed to the psychological laboratory, have the rich semantics which give rise to this robustness property if we can but discover the appropriate language for the inferences and appropriate mechanisms for carrying them out.

The region ghosts are, if you like, region intensions while the found regions are (imperfect) region extensions [27]. A ghost is an intension in that it may be specified as, for example, "the region on the reflex angle side of this acute L." The intension/extension distinction forms a spectrum rather than a strict dichotomy here. Recall that a ghost arises when a cue fails to find an associated region. It may fail either because it stopped looking too soon even though there is a found region there or because there is no found region. The ghosts participate in the consistency process just as do the found regions. The single cue that created a region ghost constrains it and it is quite possible for interpretations of the ghost to be progressively ruled out. After the consistency process we still do not know the extension of a ghost but we may know more about it than before; for example, it may now be forced to have the interpretation "land".

The missing regions, as in the river delta, for example, also do not seriously affect the consistency process. The cues in the neighbourhood of a missing region will have used ghosts in its stead. But, standing in for a single missing region there will be several ghosts so the constraining effect will be weakened somewhat.

Similarly, the extra regions created by the splitting of a single intended region participate independently in the consistency process thereby exerting a weaker constraining effect than if the region had not been split. However, the semantic richness overcomes that weakening and forces the three found regions corresponding to the single intended land region to have that single interpretation. Again, as in the other cases, if the region splitting is so severe as to cut too many inference paths then the process will degrade gracefully. In that case the various found regions would not have the intended interpretation uniquely --it would simply be in the intersection of the possible interpretations of the found regions.

We can go further and use the results of the consistency process to refine the initial partial segmentation. There are four ways, currently being implemented, in which this can be done: a) establishing distinct ghosts with the same interpretation and location as co-extensive b) considering the merge of found regions with the same interpretation c) establishing a found region as the extension of a ghost with the same interpretation and d) discovering a new found region as the extension of one or more ghosts. These all involve revisiting the picture and segmenting more purposefully, more carefully and at a finer level of detail in the particular areas concerned.

The above description of MAPSEE is only a superficial sketch. For full details on the program, the cue-model structure, the n-ary network consistency algorithm and a trace of its interpretation of the sketch map of Figure 1 see [14]. The description given here, though, should suffice to indicate the power of this approach to vision.

One of the fundamental advantages of a cautious segmentation combined with cue-invoked descriptive models that are made to interpret the picture consistently is that the constraining effect of the picture objects discovered is additive and incremental not all-or-none. As additional information is discovered in the picture it contributes its own specialized constraints to the interpretation in a uniform way. As a result, picture objects missed in the segmentation, objects split in two, undiscovered relations between objects and picture objects hallucinated to stand in for ones that cannot be found all can cause, at best, a slower convergence to the same interpretation or, at worst, a graceful degradation [15] to a more ambiguous interpretation rather than a catastrophic failure.

5. The Search for Generality: Model Descriptions

One of the legitimate criticisms of the mini-world approach to vision or artificial intelligence advocated here is that it can degenerate into a series of implementations for a series of worlds with little transfer of theory (or code) from one to the next. This can be avoided if the worlds are chosen with regard to the needs of the theory, not vice versa. Moreover, in the search for generality, one should consider families of worlds which allow a high degree of theory-sharing. Here, for example, the family of sketch worlds and the organization of MAPSEE allow us to contemplate a PLANSEE, for sketch plans of a building, a FLOWCHARTSEE, . . . even a BLOCKSEE for sketches of blocks! To change to such a new world minimally requires a new primary cue interpretation table and, perhaps, extending the vocabulary of primary cues. The modularity of this paradigm is one of its encouraging aspects: the domain dependence is highly localized within the code.

Note that the primary cue interpretation table is implicitly compiled from a set of models of cartographic objects. To further explore the problem of generality that process of compilation must be made explicit and then automated. This would require a source language in which to specify the structure of the scene objects which could exist (here the cartographic objects: mountains, bridges, roads, towns, river systems, shorelines, lakes, seas, . . .) and their possible interaction in terms of a specified repertoire of primary cues. The compilation would essentially invert those descriptions to construct the primary cue interpretation table. Note that we would not then throw away the model description. The primary cue interpretation table only captures local knowledge. The primary cues serve as indices into the set of models: their complete interpretations could then impose more global constraints on the consistency process.

As a short range strategy, this would lead even further in the direction of satisfying goal III--factoring out the highly domain-specific knowledge. But, although network consistency algorithms are probably the best uniform procedure for satisfying descriptive models, we are, in the long run, going to be forced to abandon them if we want to explore goal V: exploring control strategies for schema-based theories of perception. This conflict will lead to a divergence in the project. One path will continue to explore descriptive models, network consistency and modular vision programs while the other will explore the concept of models as procedural schemata.

6. A Ptolemaic Theory of Perception

This paper started with an appeal to the history of science so it is appropriate, given that perception is a snake swallowing its tail, for it to end with such an appeal. I have always preferred the Ptolemaic description of the motion of the heavenly bodies to the Keplerian-Copernican view so, although Ptolemy's model is currently in disfavour for the [illegible]terial universe, I shall offer it, semi-seriously, as a metaphor for the universe of perception.

This approach to perception assumes that Helmholtz, Bartlett, Minsky, Clowes and Gregory are right! Although such knowledge-based theories of perception are riddled with large hol[illegible], hand-waving, errors and mystifications, there is enough evidence from both machines and humans to know that they are, in essence, correct. One view of Roberts' achievement in creating a machine vision paradigm is the realization that he established a working model of perception as an alternation of segmentation and interpretation or, in more detail, as a cycle of four processes: discovering cues, activating a hypothesis, testing the hypothesis and inferring the consequences of an established hypothesis.

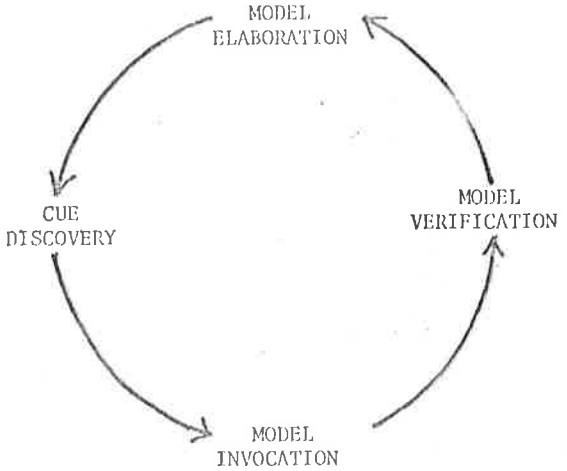
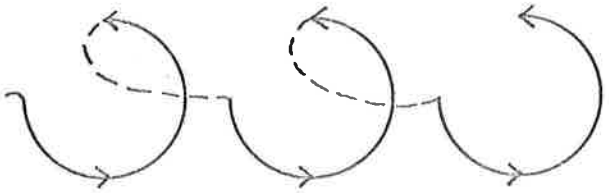


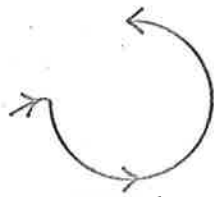
Figure 2. The cycle of perception

In Figure 2 these four processes are called cue discovery, model invocation, model verification and model elaboration. Everyday human perception is an ongoing equilibrium of similar processes. Although we can so do, we are rarely called upon to start up the cycle context-free in either bottom-up or top-down mode. But we place our programs in that situation all the time and then argue about whether bottom-up or top-down methods are more appropriate. In this metaphor, the chicken-and-egg problem simply reflects the fact that the circle is indeed unbroken.

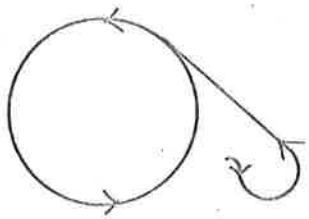
It is possible to characterize almost all vision programs by the way they treat the cycle of perception while ignoring many other issues of descriptive and procedural adequacy. Roberts' program starts with the cues and goes through the cycle several times---each time in a different area of the picture:



The Huffman-Clowes-Waltz approach starts with context-free cue discovery and does not complete the cycle:

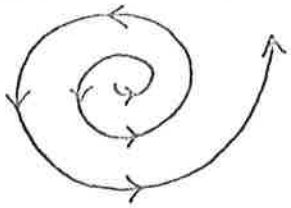


Among other things, MAPSEE has closed that gap. The several programs that use a planning approach such as Kelly's [8] and Shirai's [21] are bi-cycle theories or, as the English might say, penny-farthing theories:



in that the first cycle, on a reduced picture, provides the context for the second.

The semantics-driven region segmentation schemes invented by Yakimovsky and Feldman [28], generalized by Tenenbaum and Barrow [23] and modified by Starr and Mackworth [22] start with context-free segmentation of the "strongest" regions as cues. These regions are interpreted and their interpretations then provide the context for further segmentation and interpretation. This process continues until the entire picture is segmented/interpreted. This "island-driving" approach (which is strongly analogous to similar approaches in speech understanding [19] can be diagrammed as:



We showed [22] that a version of this technique is much more effective than traditional pattern recognition techniques in the interpretation of LANDSAT image data in that it allows 2D spatial and meaning contexts to guide the segmentation process.

Finally, in this metaphor, we need to discuss the use of hierarchies of cues and models. Mackworth [12] presents a variety of intelligent uses of composition (part-of) and generalization (is-a) hierarchies in the blocks world. In Minsky's [16] seductive vision of frame systems such hierarchies produce epicycles on the cyclic structure! (The image of this is left to your imagination.) From the metaphor it should be clear that the top-down control strategy with shared terminals transferred on failure suggested by Minsky and elaborated by Kuipers [9] is an attractive but inadequate control structure. Havens [7], in a contribution to the solution of the chicken-and-egg problem, has provided mechanisms in a programming language, MAYA, which allow the user to specify how bottom-up and top-down techniques are to intermingle in a perceptual task. Havens is pursuing the possibilities of this approach in a frame system for the blocks world. The procedural fork of our project will continue to explore the adequacy of control structures for schema-based theories of perception.

7. Conclusion

The thesis that vision research benefits most from choosing to understand limited but non-sterile domains that stretch the current theory has been supported by the example given from the sketch world. In that context, some light has been thrown on a wide variety of vision issues, such as a conservative partial segmentation, its effect on the interpretation, the possibility of a uniform control structure: network consistency with descriptive models, using the interpretation to refine the segmentation, the incremental nature of constraint methods, cue/model hierarchies, conventional versus optical semantics, and the modularity and generality problems that conflict with the procedural adequacy requirement placed on any theory of perception.

8. References

1. Anderson, R.H. Syntax-directed recognition of hand-printed two-dimensional mathematics. Ph.D. thesis, Div. Eng. and Appl. Phys., Harvard, 1968.

2. Barrow, H.G. and Tenenbaum, J.M. Representation and use of knowledge in vision. SIGART Newsletter, 52, June, 1975, 2-8.

3. Bobrow, D.G. and Collins, A.M. (Eds.) Representation and Understanding, Academic Press, N.Y., 1975.

4. Bobrow, D.G. and Winograd, T. An overview of KRL, a knowledge representation language. Journal of Cognitive Science, December, 1976.

5. Canadian Soc. Comp. Studies of Int. Proc. First National CSCSI/SCEIO Conf. Dept. of Comp. Sci., Univ. of B.C. Vancouver, B.C., August, 1976.

6. Clowes, M.B. On seeing things. Artificial Intelligence, 2, 1, 1971, 79-112.

7. Havens, W.S. Can frames solve the chicken-and-egg problem? in [5], pp. 232-242.

8. Kelly, M. Visual identification of people by computer. Memo AI-130, Comp. Sci. Dept., Stanford Univ., July, 1970.

9. Kuipers, B.J. A frame for frames: representing knowledge for recognition. in [3].

10. Mackworth, A.K. Consistency in networks of relations. TR 75-3, Dept. of Comp. Sci., Univ. of B.C., Vancouver, 1975, and Artificial Intelligence 8, 99-118, 1977 (in press).

11. Mackworth, A.K. How to see a simple world. in Machine Intelligence 8, E.W. Elcock and D. Michie (eds.) (in press) and TR 75-4, Dept. of Comp. Sci. Univ. of B.C., Vancouver, 1975, pp. 60.

12. Mackworth, A.K. Model-driven interpretation in intelligent vision systems. Perception 5, 1976, 349-370.

13. Mackworth, A.K. Making maps make sense. in [5], pp. 42-51.

14. Mackworth, A.K. On reading sketch maps. TR 77-2, Dept. Comp. Sci., Univ. of B.C., Vancouver, 1977, pp. 25.

15. Marr, D. Early processing of visual information. A.I. Memo 340, M.I.T., Cambridge, Mass., 1975.

16. Minsky, M.L. A framework for representing knowledge. in [25], pp. 211-277.

17. Montanari, U. Networks of constraints: fundamental properties and applications to picture processing. Information Sciences, 7, 1974, 95-132.

18. Negroponte, N. Recent advances in sketch recognition. AFIPS NCC Proc., 42, pp. 663-675.

19. Paxton, W.H. Experiments in speech understanding system control. in [5] pp. 1-21.

20. Roberts, L.G. Machine Perception of three-dimensional objects. in Optical and Electro-optical Information Processing, Tippett. et al. (eds.), M.I.T. Press, Cambridge, Mass., 1965, pp. 159-197.

21. Shirai, Y. A context sensitive line finder for recognition of polyhedra. Artificial Intelligence, 4, 2, 1973, 95-119.

22. Starr, D.W. and Mackworth, A.K. Interpretation-directed segmentation of ERTS images. Proc. ACM/CIPS Pacific Regional Symp. 1976, pp. 69-75.

23. Tenenbaum, M. and Barrow, H.G. IGS: a Paradigm for integrating image segmentation and interpretation. in Pattern Recognition and Artificial Intelligence, 1976, Academic Press.

24. Waltz, D.L. Generating Semantic Descriptions from Drawings of Scenes with Shadows. MAC AI-TR-271, M.I.T., Cambridge, Mass.

25. Winston, P.H. The MIT robot. in Machine Intelligence 7, Meltzer, B. and Michie, D. (eds.), Edin. Univ. Press, 1973, pp. 431-463.

26. Winston, P.H. The Psychology of Computer Vision. McGraw-Hill, N.Y., 1975.

27. Woods, W.A. What's in a link. in Representation and Understanding, D.G. Bobrow and A. Collins (eds.), Academic Press, 1975.

28. Yakimovsky, Y. and Feldman, J. A semantics-based decision-theoretic region analyzer. Proc. IJCAI, Stanford, Calif., 1973, pp. 580-588.