ON SEEING THINGS, AGAIN

Alan Mackworth
Laboratory for Computational Vision
Department of Computer Science
University of British Columbia
Vancouver, B.C., Canada V6T 1W5

ABSTRACT

Computational vision has developed as a distinct scientific field in the last decade with a shared paradigm, a research strategy and a collection of results within a common theory. That development has focussed on an analysis of the visual task itself; the task requires the unpacking of a collection of confounding processes. Explicit intermediate representations of the confounded domains must be constructed, with certain characteristics. Constraint-based representations and processes provide a common methodology at all levels of the visual system. Adequacy criteria may be applied to the various representations of visual knowledge, both implicit and explicit, that have been proposed. Finally, nine broad areas of research progress are summarized. The agenda for the next decade must include understanding meta-knowledge computation such as the representation and use of the hierarchies of default assumptions that our microtheories require.

I THE TASK

The vision problem is characterized by information loss in the image formation process. The intensity values in an image are the result of the interaction of many factors, including the intensity, colour, location and nature of the light sources, the position, reflectance, transparency and opaqueness of the objects in the scene, the transmission, refractance, absorption and scattering properties of the light transmission media, the optical properties of the imaging device, the response characteristics of the sensor and so forth.

The difficulty of computational vision arises not because we do not understand or cannot model these processes and their interactions; we do and we can. The difficulty lies in representing the uncertainty caused by the many-to-one confounding process itself and in resolving that uncertainty through the acquisition of more imagery or the addition of other constraints on the elements of the confounded domains.

II THE EMERGENCE OF A SCIENCE

The emergence of the science of computational vision over the last decade has been characterized by this focus on the nature of the visual task itself. Any vision system must make explicit descriptions of the sets of elements of the implicit, confounded domains that could have produced the image under certain specific *a priori* assumptions (Mackworth, 1983a).

Marr (1982) insisted on three levels of analysis: a *computational* theory of the task that specifies the purposes of the system, the domains involved and the input/output predicates for processes operating on those domains, an *algorithmic* level giving detailed descriptions of those processes and a *mechanism* level that describes the implementation of those algorithms on the real hardware available (or on virtual hardware, since this analysis may be recursive). The research conducted in the style of this dictum has had several effects. First, we see that there can be a genuine science of vision with a common body of theory at the computational level. This science can be applied either to biological systems or to man-made artefacts. Second, an obsession with particular algorithms or mechanisms without justification or analysis at the computational task level is sterile. Some historically early work concentrated on performance and one-off demonstrations, even going so far as to make a virtue of necessity, claiming "The program *is* the theory". The days of the theory-shy vision hacker are over. Equally, of course, we must avoid Charybdis: theory that exists for its own sake, masquerading as pure science, with all its seductive allure. Highly mathematical microtheories embodying artificial assumptions, not embedded in a realistic system design, may not be very useful.

Since the computational level of analysis is, in large measure, dictated by the task and the mechanism level dictated, in some sense, by the hardware, the widest choice with the weakest evaluation measures occurs at the algorithmic level (Poggio, 1983).

III INTERMEDIATE REPRESENTATIONS

Since the confounding process must be unpacked as a series of confounding processes we

must have explicit representations of the intermediate stages. Such intermediate representations must have certain characteristics. They must be capable of finitely representing all and only the (usually) infinite number of configurations in the intermediate domains that could have produced the image. They are best thought of as intensional descriptions of equivalence classes in some configuration space for the intermediate domain. They must be capable of being refined as additional information is available in the form of additional imagery of the assumed common underlying scene whether that information comes from stereo, colour, structured light, a rangefinder, motion or some other source. They must be efficiently computable from the information available in the image and the other intermediate representations. They must be capable of representing the tradeoffs between intermediate domains such as the shape/perspective tradeoff or the reflectance map/shape tradeoff (Woodham, 1983) that allow one factor to be varied while simultaneously varying the other to maintain a constant image. Finally, they must be capable of representing and enabling access to model constraints (and their logical consequences) such as "this surface is convex", "all surfaces are 'smooth' or ruled", "all surfaces have Lambertian reflectance maps", "all objects in this domain are opaque polyhedra", "this object is a banana" and the like.

Intermediate representations such as the primal sketch (Marr, 1982), intrinsic images (Barrow and Tenenbaum, 1978), the gradient space (Huffman, 1971; Mackworth, 1973), the 2½D sketch (Marr, 1982), the reflectance map (Horn, 1977) and generalized cylinders (Binford, 1971) have been developed with some of these criteria in mind if not explicitly stated, although none of them properly satisfy all those criteria.

Intermediate representations are arranged on a spectrum based on the degree of egocentricity of their coordinate frames. The ones noted are so ordered. Representations can be categorized as retina-centred, viewer-centred, object-centred and world-centred.

IV CONSTRAINTS

This characterization of the vision task is based on the fact that a single image underconstrains the scene; it asks how to represent the equivalence class of possible scenes. It then asks what additional constraints are necessary to specify the scene uniquely. The additional constraints may arise from a priori general knowledge of the scene domain, the imaging process, lighting and the like or from contingent knowledge of context or from general or specific object models or from additional images together with assumptions about how they interrelate in the scene domain.

The constraint-based view has proven useful at the computational, algorithmic and mechanism levels of analysis (Marr, 1982; Mackworth, 1973, 1977a; Zucker, 1981). The intermediate representations can also be judged on their ability to support this view of the process by allowing representation of such constraints at all levels of the visual system.

V VISUAL KNOWLEDGE REPRESENTATIONS

The vision task requires knowledge-based processing thus placing it firmly in the artificial intelligence paradigm. A set of criteria for descriptive adequacy and procedural adequacy must be used to evaluate proposed visual knowledge representations (Clowes, 1971; Havens and Mackworth, 1983). Parenthetically, this does not prejudge the goal-driven/data-driven distinction which is simply one of the procedural adequacy issues for the particular task and level of the vision system involved.

The constraint-based approach is not committed to a thoroughgoing proceduralization of the knowledge necessary for interpretation. As a consequence it enables the design of image-based systems where there may be no rigid distinction between input and output (Barrow and Tenenbaum, 1978). Information and constraints are simply propagated from whichever source is able to supply them. In overdetermined visual tasks such as industrial inspection or remote sensing, where part geometry, digital terrain, reflectance and lighting models are available, this may be most fruitful approach (Woodham, 1983). It also suggests a rapprochement of the long-divorced vision and graphics communities. Other knowledge representations including logic (Kowalski, 1979), grammars (Browse, 1982) and schema systems (Havens and Mackworth, 1983) share this procedural adequacy advantage (Mackworth, 1983b).

VI THE ACHIEVEMENTS

In summary, there are nine main areas of achievement of the past decade of computational vision. First, there has been a convergence on a paradigm for a science of computational vision (Brady, 1981). Second, there has been an emphasis on understanding, modelling and exploiting the physics of image formation - its geometry and radiometry (Horn, 1977). Third, the constraint-based nature of vision has become clear. Fourth, the importance of various intermediate representations and criteria for judging them have emerged. Fifth, there has been a flurry of microtheories, results and algorithms for shape recovery from various information sources including shading (Ikeuchi and Horn, 1981), texture (Witkin, 1981), motion (Ullman, 1979; Longuet-Higgins and Prazdny, 1981), stereo (Grimson, 1981; Baker and Binford, 1981), contour (Clowes, 1971; Huffman, 1971; Mackworth, 1973; Barrow and Tenenbaum, 1981; Kanade, 1981) and photometric stereo (Woodham, 1981). Sixth, some of the new insights have been applied to understanding the neurophysiology and psychophysics of mammalian vision

systems (Poggio, 1983; Ullman and Hildreth, 1983; Zucker, 1983). Seventh, new non-von Neumann models of computation of the cooperative (Marr, 1982), connectionist (Hinton, 1981; Ballard, 1981), relaxation (Zucker, 1981), and consistency (Mackworth, 1977b) style are emerging and suggesting new architectures. Eighth, criteria of descriptive and procedural adequacy have guided the development of schema-based knowledge representations for high level vision (Hanson and Riseman, 1978; Brady and Wielinga, 1978; Browse, 1982; Havens, 1983; Havens and Mackworth, 1983). Engineering applications of the new science in industrial inspection, robotics and remote sensing are flourishing at CMU, SRI, Machine Intelligence, Fairchild, Hitachi, GM, Stanford, MIT, McGill, UBC and many other labs around the world.

VII THE AGENDA

The agenda for the next decade has the same item at the top of the list as the agenda of the last decade, "How is vision possible?" The answers to that question will continue to flow from pushing hard at the boundaries of all the nine areas outlined above.

One additional research topic must be added to that agenda. The current standard theory has within it a large collection of microtheories concerned, for example, with shape recovery. Each of these microtheories has its own set of assumptions and restrictions, domain of applicability, requisite inputs and attached methods, many of which are implicit both in the theories and the programs that implement them. We must now learn how to represent these knowledge "bundles" explicitly in our programs and have well-structured ways of indexing, invoking, using and coordinating them.

As an example of this, our cooperative interpretation paradigm suggests that two (or more) knowledge sources can work synergistically to interpret an image when either alone would do poorly (Mackworth, 1978). Glicksman's (1983) Missee system uses the qualitative spatial scene knowledge provided by the Mapsee2 interpretation of a sketchmap to augment the spectral and spatial constraints crudely extracted from an aerial image of the same scene, integrating information from both sources into a single schema-based representation.

More generally, meta-knowledge computation must be part of the vision system. One approach to this is to apply default logic theory (Reiter, 1980). Many logical implications flow from the scene domain to the image domain but can only be reversed under various contingent "general viewpoint", "general light source" or "smooth surface" assumptions analogous to the "closed world" or "circumscription" (McCarthy, 1980) assumptions. As a trivial example, a straight edge in a 3D scene is depicted as a straight line in the image. The abductive inference step required to reverse that logical impliciation does not follow without explicitly invoking the default "general viewpoint" assumption. Current vision systems (Lowe and Binford, 1981; Binford, 1981) abound with examples of this style of reasoning. We must develop a logic of depiction.

The importance of hierarchical descriptions both in the image domain and in the scene domain is well understood. Spatial frequency channels, for example, simply encode a hierarchy of image detail useful for edge detection (Marr and Hildreth, 1980) and stereo matching (Grimson, 1981). Within a particular scene domain, specialization and composition hierarchies serve both descriptive and procedural adequacy (Mackworth and Havens, 1981). Hierarchical descriptions will also prove to be useful for organizing collections of default and domain assumptions of widely varying power and generality.

ACKNOWLEDGEMENTS

The suggestions and support of all my colleagues, especially Bave Brent, Roger Browse, Rachel Gelbart, Jay Glicksman, Bill Havens, Jan Mulder, Ray Reiter and Bob Woodham, are gratefully acknowledged. Max Clowes was a mentor and a friend.

REFERENCES

[1] Baker, H.H. and T.O. Binford "Depth from Edge and Intensity Based Stereo" Proc. IJCAI-81 Vancouver, Canada, August, 1981, pp. 631-636.

[2] Ballard, D.H. "Parameter Networks: Towards a Theory of Low-Level Vision" Proc. IJCAI-81, Vancouver, Canada, August, 1981, pp. 1068-1078.

[3] Barrow, H.G. and J.M. Tenenbaum "Recovering Intrinsic Scene Characteristics From Images" In A.R. Hanson and E.M. Riseman (eds.) Computer Vision Systems, New York: Academic Press, 1978, pp. 3-26.

[4] Barrow, H.G. and J.M. Tenenbaum "Interpreting Line Drawings as Three-Dimensional Surfaces" Artificial Intelligence 17 (1981) 75-116.

[5] Binford, T.O. "Visual Perception by Computer" Presented to IEEE Conf. on Systems and Control, Miami, December, 1971.

[6] Binford, T.O. "Inferring Surfaces from Images" Artificial Intelligence 17 (1981) 205-244.

[7] Brady, J.M. and B.J. Wielinga "Reading the Writing on the Wall" In A.R. Hanson and E.M. Riseman (eds.) Computer Vision Systems, New York: Academic Press, 1978, pp. 283-301.

[8] Brady, J.M. "Preface-Changing Shape of Computer Vision" Artificial Intelligence 17

(1981) 1-15.

[9] Browse, R.A. "Knowledge-based Visual Interpretation Using Declarative Schemata" Ph.D. Thesis, Dept. of Computer Science, University of B.C., Vancouver, Canada, 1982.

[10] Clowes, M.B. "On Seeing Things" Artificial Intelligence 2:1 (1971) 79-112.

[11] Glicksman, J. "Using Multiple Information Sources in a Computational Vision System" Proc. IJCAI-83 Karslruhe, Germany, August, 1983 (this volume).

[12] Grimson, W.E.L. From Images to Surfaces Cambridge, MA: MIT Press, 1981.

[13] Hanson, A.R. and Riseman, E.M. "VISIONS: A Computer System for Interpretating Scenes" In A.R. Hanson and E.M. Riseman (eds.) Computer Vision Systems, New York: Academic Press, 1978, pp. 303-333.

[14] Havens, W.S. "Recognition Mechanisms for Hierarchical Schemata Knowledge Representations" Int. J. of Comouters and Mathematics, 9:1 (1983) 185-200.

[15] Havens, W.S. and Mackworth, A.K. "Representing Knowledge of the Visual World" IEEE Computer (accepted for publication).

[16] Hinton, G.E. "Shape Representation in Parallel Systems" Proc. IJCAI-81, Vancouver, Canada, August, 1981, pp. 1088-1096.

[17] Horn, B.K.P. "Understanding Image Intensities" Artificial Intelligence 8:2 (1977) 201-231.

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

[19] Ikeuchi, K. and B.K.P. Horn "Numerical Shape from Shading and Occluding Boundaries" Artificial Intelligence 17 (1981) 141-184.

[20] Kanade, T. "Recovery of the Three-Dimensional Shape of an Object from a Single View" Artificial Intelligence 17 (1981) 409-460.

[21] Kowalski, R. Logic for Problem Solving Amsterdam: North-Holland 1979.

[22] Longuet-Higgins, H.C. and K. Prazdny "The Interpretation of Moving Retinal Image" Proc. Royal Soc. B 208 (1980) 385-387.

[23] Lowe, D.G. and T.O. Binford "The Interpretation of Three-Dimensional Structure from Image Curves" Proc. IJCAI-81 Vancouver, Canada, August, 1981, pp. 613-618.

[24] McCarthy, J. "Circumscription - A Form of Non-Monotonic Reasoning" Artificial Intelligence 13:1,2 (1980) 27-39.

[25] Mackworth, A.K. "Interpreting Pictures of Polyhedral Scenes" Artificial Intelligence 4:2 (1973) 121-137.

[26] Mackworth, A.K. "On Reading Sketch Maps" Proc. IJCAI-77 Cambridge, MA, August, 1977a, pp. 598-606.

[27] Mackworth, A.K. "Consistency in Networks of Relations" Artificial Intelligence 8:1 (1977b) 99-118.

[28] Mackworth, A.K. "Vision Research Strategy: Black Magic, Metaphors, Mechanisms, Miniworlds and Maps" In A.R. Hanson and E.M. Riseman (eds.) Computer Vision Systems, New York: Academic Press, 1978, pp. 53-61.

[29] Mackworth, A.K. and W.S. Havens "Structuring Domain Knowledge for Visual Perception" Proc. IJCAI-81 Vancouver, Canada, August, 1981, pp. 625-627.

[30] Mackworth, A.K. "Constraints, Descriptions and Domain Mappings in Computational Vision" In O.J. Braddick and A.C. Sleigh (eds.) Physical and Biological Processing of Images Berlin: Springer-Verlag, 1983a, pp. 33-40.

[31] Mackworth, A.K. "Recovering the Meaning of Diagrams and Sketches" Proc. Graphics Interface '83, Edmonton, Canada, May, 1983b, (in press).

[32] Marr, D. and E.C. Hildreth "Theory of Edge Detection" Proc. Royal Soc. B 200 (1980) 187-217.

[33] Marr, D. Vision San Francisco: W.H. Freeman, 1982.

[34] Poggio, T. "Visual Algorithms" In O.J. Braddick and A.C. Sleigh (eds.) Physical and Biological Processing of Images Berlin: Springer-Verlag, 1983, pp. 128-153.

[35] Reiter, R. "A Logic for Default Reasoning" Artificial Intelligence 13:1,2 (1980) 81-132.

[36] Ullman, S. The Interpretation of Visual Motion Cambridge, MA: MIT Press, 1979.

[37] Ullman, S. and E.C. Hildreth "The Measurement of Visual Motion" In O.J. Braddick and A.C. Sleigh (eds.) Physical and Biological Processing of Images Berlin: Springer-Verlag, 1983, pp. 154-176.

[38] Witkin, A.P. "Recovering Surface Shape and Orientation With Texture" Artificial Intelligence 17 (1981) 17-45.

[39] Woodham, R.J. "Analysing Images of Curved Surfaces" Artificial Intelligence 17 (1981) 117-140.

[40] Woodham, R.J. "Viewer-Centered Intensity Computations" In O.J. Braddick and A.C. Sleigh (eds.) Physical and Biological Pro-

cessing of Images Berlin: Springer-Verlag, 1983, pp. 217-229.

[41] Zucker, S.W. "Computer Vision and Human Perception" Proc. IJCAI-81 Vancouver, Canada, August, 1981, pp. 1102-1116.

[42] Zucker, S.W. "Cooperative Grouping and Early Orientation Selection" In O.J. Braddick and A.C. Sleigh (eds.) Physical and Biological Processing of Images Berlin: Springer-Verlag, 1983, pp. 326-334.