COMPUTATIONAL SELECTION OF PROCESSING LOCATIONS IN VISION

by

Roger Browse

Department of Computing and Information Science
Department of Psychology
Queen's University, Kingston, Ontario, Canada

1. Introduction

Some of the issues addressed in computational vision research relate to the question of the structure of human intelligence in general, and are thereby subject to a Cognitive Science approach. This is particularly true of the traditional issue of how knowledge of specific objects might be applied towards visual recognition and understanding.

There are two advantages to the development of computer vision systems which maintain compatibility with selected aspects of human vision. First, clues to the underlying operational requirements of the vision system may be obtained from the characteristics of the human system, and second, the resulting computational system may act as an explanatory model for the human operations. The research reported here centers around a computational vision system which interprets line drawings of human-like body forms. This system maintains a number of compatibilities with human vision (see Browse 1981; 1982b).

This paper describes a mechanism for the interpretation-based integration of information available at different levels of resolution. From the integration, there follows a method for the intelligent selection of processing location within the image.

2. Multiple Resolution Systems

There is a variety of evidence in favor of approaching vision as a process which operates with information available from several different, but related, levels of resolution. The primate visual system has a distribution and structure of retinal receptors which leads to responses based on different receptive field sizes (see Hubel and Weisel, 1979). While several receptive field sizes may coexist at any specific point on the retina, a near-linear increase in field size (decrease in resolution) is exhibited towards the periphery (Wilson and Bergen, 1979).

Consideration of these aspects of the visual system appears in the formulation of contemporary computationally based theories of vision. Marr and Hildreth (1979) present a mathematical formulation of edge detection in which images smoothed with a variety of Gaussian filters, are convolved with the Laplacian operator. The zero-crossings of these convolutions are representative of the image intensity changes in different spatial frequency channels, depending on the value of the space constant of the Gaussian distribution.

One important theoretical question which confronts the development of such theories which maintain compatibility with human vision is: **What mechanism can provide useful interaction between information from different resolution levels?**

For most theories of computational vision, multiple resolution is a tool in the discovery of context-free image features such as edges. This notion has been developed through the introduction of "image pyramids" (Uhr, 1972; Hanson and Riseman, 1975) which represent an image as several interrelated layers, constructed at different resolutions. The base level is the regular digitized image, and the upper layers are successively smaller images, with pixel values derived by some averaging operation on four (or more) pixels at the level beneath it. A number of processing schemes have been devised to use these structures to aid in the detection of image features. The basic idea behind the use of pyramids is that indications of the existence of a feature may be found in a simple search of a smaller, coarser resolution version of the image. This initial detection can be used to direct the extraction of features from the finer levels (see Tanimoto, 1980).

Marr (1976) takes a similar approach. The process of combining the results from different spatial frequency channels relies on the idea that zero crossings at the same location at different scales are

probably the result of the same underlying physical phenomenon. So whenever the segments obtained from two or more (contiguous) channels agree in both position and orientation, an edge is hypothesized.

Other computer vision research attempts an interpretation-based interaction between levels of resolution, using knowledge of the class of objects which comprise the problem domain. Kelly (1971) devised a system which analyzed coarse level features in the context of what was expected for the outline of a head. Thus the subsequent examination of the fine features was able to ignore the other prominent edges produced by the background. Bajczy and Rosenthal (1980) have extended the interaction between world knowledge and image hierarchy in an inquiry-driven computer vision system which relies on the natural hierarchical relations of the problem domain. The methods are similar to those proposed by Palmer (1977).

Psychological research supports this idea of the involvement of interpretation in the interaction of resolution levels. Kinchla (1974) and Navon (1977) developed an experimental paradigm in which subjects are presented with the task of processing local and global information at the same time. The result from using this method demonstrate different properties of recognition based on visual information from the two levels. Miller (1981) has presented results which indicate the requirement for a model of perception in which information from different levels of resolution feed into a single decision process which integrates the results.

The approach taken in the body-drawing interpretation system is that separate representations are maintained for information from different resolution levels, and that each one has specialized capabilities for interacting with knowledge of the problem domain.

A second question which relates to the use of multiple resolutions is: **How can locations be selected for the application of high resolution visual processing?**

This question is not only of interest in simulations of human visual processing, but is also a concern to the pragmatics of engineering computer vision systems. The issue of intelligently ordering and restricting the processing of an image becomes more important as larger feature detection masks are being convolved with images of increasing total number of picture elements.

Saccadic eye movements are the most obvious of the visual selection processes. There are many excellent reviews emphasizing the role of cognition in the selection of fixation locations (see Rayner, 1978). It is generally accepted that expectation or conflict within the ongoing visual interpretation influences the choice of processing area (Loftus and Mackworth, 1978). One important aim of the body-drawing interpretation system was to develop a computational basis for these capabilities.

3. System Overview

A computational vision system has been implemented which interprets a class of line drawings of human-like body forms as shown in figure 1. The basic features available to the interpretation processes are representative of three different resolution levels:

(1) fine resolution: line segments and connections between line segments taken from a 1024x1024 image.

(2) coarse resolution: axis measurements of blobs detected in a 128x128 image.

(3) very coarse resolution: measure of the amount of detail that exists in each pixel of an 8x8 representation (shown as 32x32).

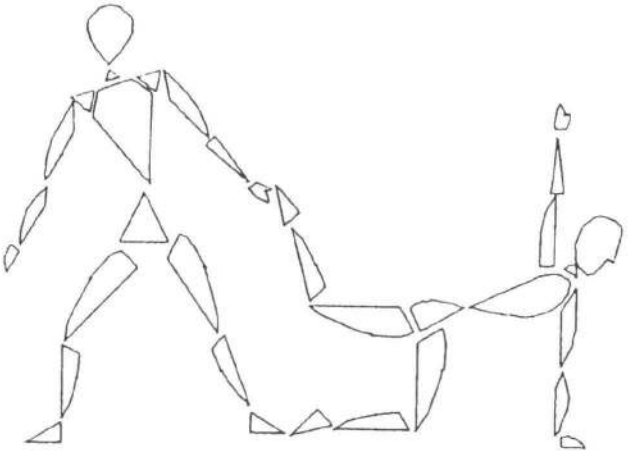

figure 1.

At any given point in the processing, these features are only available in limited diameters of the image, represented as overlapping concentric circles (see figure 2).

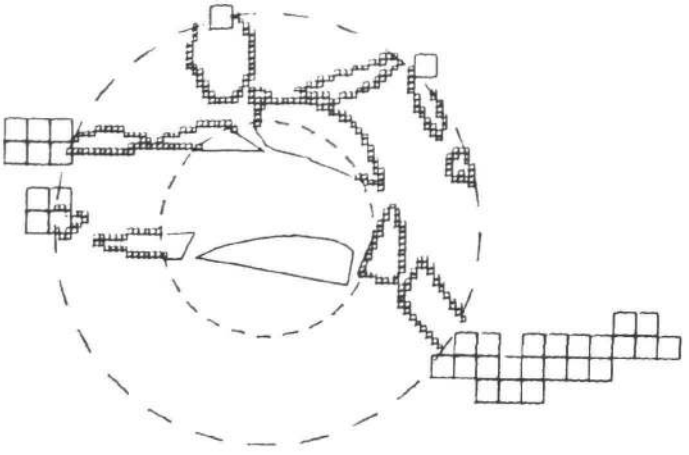

figure 2.

The system uses a schemata-based knowledge representation of the way features may compose to provide supportive evidence for body parts (see Browse 1980, 1982b). This description includes provision for the topologically distinct views of the body parts, and as well encodes the underlying structure of the human body form and its potential for deformations at the joints.

Control of possible interpretations is accomplished by using a cue/model technique adapted from Mackworth's (1977a) MAPSEE system. Through a series of steps involving the application of local consistency methods (Waltz, 1972; Mackworth, 1977b), both at the feature grouping level, and at the model invocation level, interpretation hypotheses are generated regarding the underlying body form being depicted in the image.

4. Interaction Between Resolution Levels

Within limited diameters of available features, complete interpretation of the line drawing is not possible. The results of processing the image during a single fixation (figure 3) is shown in the parse trees of figure 4.[1]

1. Other interpretations exist, providing an ambiguous context. These alternatives have been omitted for the sake of clarity.

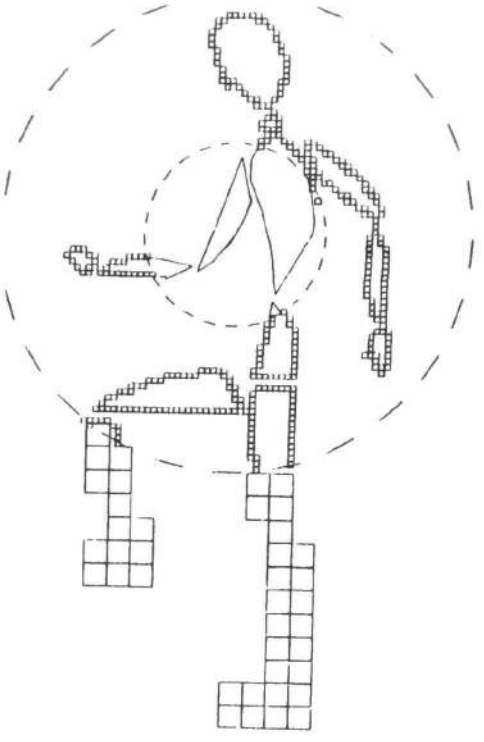

figure 3.

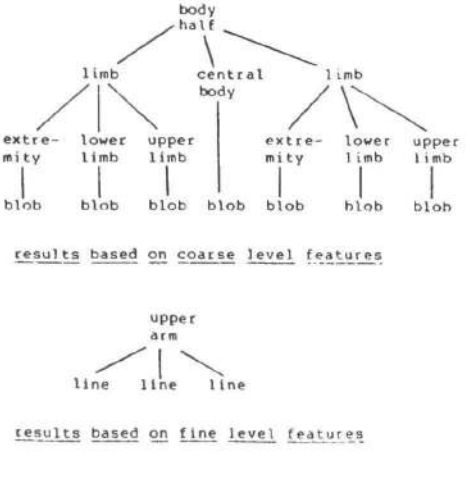

figure 4.

These results are available based on both the fine level features and on the coarse level. Note that a more specialized result is produced (in a smaller area) for the fine features.

One obvious objective of the interpretation process is to develop a detailed description for each area of the image. If the system were to rely solely on the processing of fine level features in the accomplishment of this goal, then every location in the image would have to be processed. This exhaustive operation can be avoided through the formation of **correspondences** between interpretations based on the different levels, and by propagating the detailed interpretation outward into the periphery.

Interpretation elements are said to **correspond** if (1) they are related by the specialization hierarchy, and (2) their attributes are similar (particularly location). If a course level interpretation object (such as "limb") is known, and one of its components (such as "upper-limb") has a correspondence at the fine level (such as "upper-arm"), then an **inferred correspondence** may be developed for the entire coarse level object: if one of the components has been identified in detail, then the entire structure can be known in terms of its detailed interpretation. These image locations to which detailed interpretation has been propagated may then initiate propagation even further into the periphery.

In order to summarize this mechanism, consider a real-life example: you walk into a room, and your eyes fall on a bookcase in front of you. At that moment, only a few books are actually within the foveal area of detailed vision. The rest of the bookcase is only observed in the lower resolution periphery, and provides an interpretation which may be consistent with several possibilities: records on a shelf, hanging racks of computer tapes, or books on a bookcase. The subjective experience is that of confirmation of "books" over the entire bookcase. The detailed interpretation of "books" at the fovea has propagated a more specific value to the related low resolution interpretation in the periphery.

This notion of correspondence between interpretations is also the mainstay of the system's capability to intelligently select fixation locations. Any coarse level interpretation object which (1) has components (such as limb does), and which (2) does not have a correspondence at the fine level interpretation, is a prime candidate for a subsequent fixation location. The reason it would be a good candidate is that, if fixated, one of its components will probably be well enough interpretated at the fine level to provide a correspondence adequate to propagate the detailed results to the entire structure. Such locations will provide the greatest possible evidence towards a final interpretation of the image.

The mechanism of consideration of correspondence possibilities provides a means of intelligent selection of fixation location based on the ongoing status of the interpretation process. The selection of fixation locations is also sensitive to properties of the image itself. In the absence of other requirements, locations are selected which are expected to expand the peripheral interpretation area. For the example in figure 3, only three fixations are necessary before an inferred correspondence is established for the entire body form. At that point the system will not have established the exact values of the relative orientations of all body parts, but later fixations may be used to obtain these details as required by the interpretation task.

5. References

Bajcsy, R. and Rosenthal, D.A. 1980, "Visual and Conceptual Focus of Attention," in Structured Computer Vision, S. Tanimoto and A. Klinger (eds.), Academic Press, New York, 133-149.

Browse, R.A. 1980, "Mediation Between Central and Peripheral Processes: Useful Knowledge Structures," Proc. Third Conf. of the Canadian Society for the Computational Studies of Intelligence, Victoria, Canada, 166-171.

Browse, R.A. 1981, "Relations Between Scemata-Based Computational Vision and Aspects of Visual Attention," Proc. Third Annual Conference of the Cognitive Science Society, Berkeley, California, 187-190.

Browse, R.A. 1982a, "Interpretation-Based Interaction Between Levels of Detail," Proc. Fourth Conf. of the Canadian Society for the Computational Studies of Intelligence, Saskatoon, Canada, 27-32.

Browse, R.A. 1982b, "Knowledge-Based Visual Interpretation Using Declarative Schmata," Ph.D. Thesis, Department of Computer Science, University of British Columbia. Technical Report 82-12.

Hanson, A.R. and Riseman, E.M. 1975, "The Design of a Semantically Directed Vision Processor," Technical Report No. 75C-1, University of Massachusetts, Amherst, Massachusetts.

Hochberg, J.E. and Brooks, V. 1978, "Film Cutting and Visual Momentum," in Eye Movements and the Higher Psychological Functions, J.W. Senders, D.F. Fisher, and R.A. Monty (eds.), Erlbaum, Hillsdale, New Jersey, 293-313.

Hubel, D.H. and Weisel, T.N. 1979, "Brain Mechanisms of Vision," Scientific American 241 (3), 150-162.

Kelly, M.D. 1971, "Edge Detection in Pictures by Computer Using Planning," in Machine Intelligence 6, B. Meltzer and D. Michie (eds.), American Elsevier, New York, 397-409.

Kinchla, R. 1974, "Detecting Target Elements in Multielement Arrays: A Confusability Model," Perception and Psychophysics 15, 149-158.

Loftus, G.R. and Mackworth, N.H. 1978, "Cognitive Determinants of Fixation Location During Picture Viewing," Journal of Experimental Psychology: Human Perception and Performance 4, 565-572.

Mackworth, A.K. 1977a, "On Reading Sketch Maps," Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Cambridge, Massachusetts, 598-606.

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

Marr, D. 1976, "Early Processing of Visual Information," Phil. Trans. Royal Society of London 275B (942), 483-524.

Marr, D. and Hildreth, E. 1980, "Theory of Edge Detection," Proc. Royal Soc. London B (207), 187-217.

Miller, J. 1981, "Global Precedence in Attention and Decision," Journal of Experimental Psychology: Human Perception and Performance 7, 1161-1174.

Navon, D. 1977, "Forest Before Trees: The Precedence of Global Features in Visual Perception," Cognitive Psychology 9, 353-383.

Palmer, S.E. 1977, "Hierarchical Structure in Perceptual Representation," Cognitive Psychology 9, 441-474.

Rayner, K. 1978, "Eye Movements in Reading and Information Processing," Psychological Bulletin 85 (3), 618-660.

Tanimoto, S.L. 1980, "Image Data Structures," in Structured Computer Vision, S. Tanimoto and A. Klinger (eds.), Academic Press, New York, 31-55.

Uhr, L. 1972, "Recognition Cone Networks that Preprocess, Classify, and Describe," IEEE Transactions on Computers 21, 758-768.

Wilson, H.R. and Bergen, J.R. 1979, "A Four Mechanism Model for Threshold Spatial Vision," Vision Research, 19, 19-32.