WHAT SHOULD BE COMPUTED IN LOW LEVEL VISION SYSTEMS

William B. Thompson
Albert Yonas

University of Minnesota
Minneapolis, Minnesota 55455

ABSTRACT

Recently, there has been a trend towards developing low level vision models based on an analysis of the mapping of a three dimensional scene into a two dimensional image. Emphasis has been placed on recovering precise metric spatial information about the scene. While we agree with this approach, we suggest that more attention be paid to what should be computed. Pschophysical scaling, adaptation, and direct determination of higher order relations may be as useful in the perception of spatial layout as in other perceptual domains. When applied to computer vision systems, such processes may reduce dependance on overly specific scene constraints.

1. Introduction

The following is a position paper directed at several aspects of low-level visual processing. The current trend towards focusing on the determination of exact, three-dimensional form in a scene is questioned. Both analysis of representative scene domains and experience with human vision suggest that less precise form properties may be sufficient for most problems. Several computational issues are also briefly discussed.

2. Alternate Approaches to "Low-Level" Analysis

Computer vision systems have traditionally been divided into segmentation and interpretation components. A multiplicity of image features have been investigated in the hope that they would facilitate the partitioning of an image into regions corresponding to "objects" or "surfaces" in the original scene. Only after this two-dimensional segmentation operation was completed would procedures be applied in an attempt to determine the original three-dimensional structure of the scene. Recently, an alternative approach to implementing the lower levels of a computational vision model has been developed. Its basic premise is that the determination of three-dimensional structure is such an integral part of the scene description processes that it should be carried out at all levels of the analysis [1,2,3,4,5].

Proponents of this approach usually employ a well structured methodology for developing computational models of form perception:

> Precisely describe a carefully constrained scene domain.
>
> Identify important scene properties.
>
> Determine the function which maps these scene properties into an image.
>
> Develop computationally feasible mechanisms for recovering the "important" scene properties from the image.

Great emphasis is placed on determining what scene properties are computable, given a set of constraints on the scene.

Scene properties normally considered essential to the analysis include object boundaries, three-dimensional position, and surface orientation. In many cases, the determination of these features requires that properties such as surface reflectance and illumination must also be found. A key distinction between the classical techniques and this newer approach is that in the latter, analysis procedures are developed analytically from an understanding of how scene properties affect the image rather than ad hoc assumptions about how image properties might relate to scene structure.

The representational structures which have been used to implement form based analysis have, for the most part, been iconic. The features represented are almost always metric properties of the corresponding point on the surface: distance from the observer, orientation with respect to either the observer or a ground plane, reflectance, incident illumination, and so on. To determine relative effects (eg. which of two points is farther away), absolute properties are compared.

The determination of these metric scene properties requires that the possible scenes be highly constrained. Usually, the analysis depends on restrictions both on the types of objects allowed and on the surface properties of the objects. For

This research was supported in part by the National Science Foundation under Grant MCS-78-20780 and by the National Insitute of Child Health and Human Development under Grants HD-01136 and HD-05027.

example, a "blocks world" assumption (or alternately, the assumption of a "Play-Doh" world made entirely of smooth surfaces) might be made. In addition, it is commonly assumed that surfaces are all lambertian reflectors and that, for a given surface, the coefficient of reflectance is constant. Illumination is often limited to a single distant point source, possibly coupled with a diffuse illuminator. Secondary illumination effects are usually presumed to be negligible.

3. Absolute Scene Properties Are Not Always Needed

The proponents of form based analysis presume the need for finding exact shape properties of a scene. They concentrate on investigating how constraints on scenes affect what properties are computable and how they can be determined. We suggest that more attention be paid towards what properties should be computed. We argue that for a wide variety of problem areas, absolute metric information about scene shape is not required. Instead, relative properties such as flat/curved, convex/concave, farther-away/closer, etc. are both sufficient and easier to compute.

Most tasks involving description of a visual environment depend on generalized shape properties. In fact, much effort has been spent searching for shape characterizations that embody those relationships useful for description but not the enormous amount of irrelevant detail contained in any representation based on specific position. Even in task domains such as object manipulation and obstacle avoidance, precise positional information is frequently not necessary. Both these task areas contain significant sub-problems involving object identification - a descriptive task often possible with approximate and/or relative information about shape. Even when actual position is needed, feedback control can be used to minimize the need for highly accurate positional determinations.

A second argument for emphasizing the difficulty of determining metric properties comes from our experience with human perception. The psychological literature contains many references to the effects of the scaling process that relates the physical domain to the psychological [6,7], the effects of adaptation to stimulation [8], and the effects of practice on variable error [9]. By investigating the competence of the human visual system in determining primitive shape effects, we can gain insight into sufficient (but not necessary) properties for more complex analysis. In our own work on perception of surfaces, preliminary results from one set of experiments seem relevant to the development of computational models.

We synthesized a frontal view of a surface the profile of which is shown in figure 1. Lighting was assumed to be a combination of a single distant point source and a perfectly diffuse source. A simple reflectance model was used and secondary illumination effects were not considered. A series of synthesized images was produced with the intention of examining the perception of single displays and the ability to determine differences between displays. The "object" in our images was an ellipsoid with semi-axes A, B, and C. (A was in the horizontal direction as seen by the viewer, B was in the vertical direction, and C was in the direction along the line of sight.) The object was presented against a black background, and thus no cast shadows were present. In one set of experiments, A and B were held constant producing a circular occluding contour. Subjects were asked to estimate the value of C for a number of different displays, with true values of C ranging from one half of A to four times A. On initial trials, subjects tended to see the same shape independently of the actual value of C. On subsequent trials, performance improved, but with a significant, systematic underestimation of the true value. As a final note, when subjects were asked to qualitatively describe the changes in the scene as C was varied, they often indicated that they felt that the change was due to differences in illumination, not shape.

It is certainly premature to make any definitive conclussions from our results. Nevertheless, we suggest the following conjecture: Subjects appear to see a specific shape (as opposed to simply a "round" object); however, the metric properties they estimate for that shape are not necessarily consistent with the "true" values. The subjects do appear to be better at ranking displays based on different values of C.

4. Non-metric Scene Properties

We suggest that requiring specific, accurate determination of scene properties may be unnecessarily restrictive. Less precise and/or purely relative qualities are sufficient for many situations. By concentrating on these characteristics, we may be able to significantly relax the constraints under which our computational vision models must operate. Finally, human vision is often quite inaccurate in determining metric values for these same properties. Rather than indicating a deficiency in human vision, this suggests that alternative (and presumably more useful) characteristics are being computed by the human perceiver.

Two approaches to structuring computer vision models based on these observations seem relevant. First of all, it may be possible to directly compute properties of interest, rather than deriving them from more "primitive" characteristics (see [10,11]). For example, we might look for ways of estimating surface curvature that do not depend on first determining depth and then taking the second derivative.

A second possibility is to presume that estimation of shape properties is subject to the same scaling processes as most other perceptual phenomena. Thus, our model would estimate some non-linear but monotonic transformation of characteristics such as depth. The transformations would be adaptive, but in general not known by higher level analysis procedures. Thus, the precise metric three-dimensional structure can not be recovered. For many tasks, the scaled values are sufficient and the need for highly constrained, photometric analysis of the image is reduced. With appropriate standardization, precise scene properties may be determined. Without standardization, relative characteristics are still computable. Ordinal relationships are detrminable over a wide range while quantatative comparisons are possible over a more limited range. (eg. it may be possible to judge that A is twice as far as B but not that C is 100 times as far as D.)

5. Computational Models

Recently, much attention has been focused on using parallel process models to specify the computational structure of low-level vision systems. An image is partitioned into a set of neighborhoods, with one process associated with each region. The processes compute an estimate of scene properties corresponding to the region using the image features in the region and whatever is known about surrounding scene structure. The circularity of form estimation for one point depending on the form of neighboring points can be dealt with in several ways. A variable resolution technique may be employed. First large, non-interacting neighborhoods are used. Then, progressively smaller neighborhoods are used, each depending on scene properties computed using previously analyzed, larger regions. (Marr's stereo model is an example [12].) Alternately, an iterative technique can be used to find crude estimates of scene properties and then those values are fed back into the process to produce more refined estimates. (Examples include many "relaxation labeling" applications [13].) In either case, the determination of absolute scene properties usually requires a set of boundary values - image points at which the scene constraints allow direct determination of the properties. The computational process must then propagate these constraints to other image regions.

The robustness of these parallel process models may be significantly increased if they are only required to compute relative properties. The need for accurately propagating scene information is greatly reduced. Furthermore, photometric analysis of the image will usually not be required. For instance, general characteristics of the intensity gradient may be all that is required for analysis. As an example, for a reasonably general class of scene types, a discontinuity in the luminence gradient will usually correspond to a shadow, an occlusion boundary, or the common boundary between two surfaces. Continuous but non-zero gradient values indicate either surface curvature or illumination variation. In neither case is the actual magnitude of the gradient required.

Finally, many of the problems in low-level vision are underspecified. No single "correct" solution exists because insufficient information is available to derive the original scene properties. Thus, computational models must either naturally embody default assumptions or allow for ambiguous representations. (There is reason to expect that both approaches are useful.) Even more important, the control structures used by the models must not impose any arbitrary input/output assumptions. For example, consider again the relationship between luminence gradient, illumination direction, and surface curvature. For a given gradient, knowing either illumination or curvature allows determination of the other. The model must be able to account for this symmetry.

6. Conclusions

When attempting to construct computational models of low-level vision systems, we need to pay as much attention to what should be computed as we do to how it is computed. We may investigate this problem in at least three ways. The first is a computational approach: we can determine what is computable given a set of constraints about the scene and the imaging process. The second is an ecological approach: we catalog the range of problem domains in which our system is expected to function and then determine the primitive scene properties needed for analysis. The third is metaphorical: study a working visual system (eg. human) in order to determine which low-level scene properties it is able to perceive. These properties then define a sufficient set for analysis.

Much current work focuses on estimating exact positional information about a scene. We argue that in many cases, these metric properties cannot be easily determined. Even more importantly, however, they often need not be determined. Simple relative properties may be sufficient for analysis and be much easier to compute.

BIBLIOGRAPHY

[1] D. Marr, "Representing and computing visual information", Artificial Intelligence: An MIT Perspective, P.H. Winston and R.H. Brown, ed., pp. 17-82, 1979.

[2] H.G. Barrow and J.M. Tennenbaum, "Recovering intrinsic scene characteristics from images," in Computer Vision Systems, A.R. Hanson and E.M. Riseman, eds., New York: Academic Press, 1978.

[3] B. Horn, "Obtaining shape from shading information," in The Psychology of Computer Vision, P.H. Winston, ed., New York: McGraw-Hill, 1975.

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

[5] K.A. Stevens, "Surface perception from local analysis of texture and contour", Ph.D. Thesis, MIT, Feb. 1979.

[6] G.T. Fechner, Elemente der Psychophysik, Leipzig: Breitkopf and Hartel, 1860. (Reissued Amsterdam: Bonset, 1964.)

[7] S.S. Stevens, "Perceptual magnitude and its measurement," in Handbook of Perception, Vol. II, Psychophysical Judgement and Measurement, Carterette and Friedman, eds., New York: Academic Press, 1974.

[8] H. Helson, Adaptation Level Theory, New York: Harper, 1964.

[9] E.J. Gibson, Perceptual Learning and Development, New York: Appleton-Century-Crofts, 1969.

[10] J.J. Gibson, The Senses Considered as Visual Systems, Boston: Houghton Mifflin, 1966.

[11] J.J. Gibson, The Ecological Approach to Visual Perception, Boston: Houghton Mifflin, 1979.

[12] D. Marr and T. Poggio, "A theory of human stereo vision," MIT AI Lab. MEMO 451, Nov. 1977.

[13] A. Rosenfeld, R. Hummel, and S. Zucker, "Scene labeling by relaxation operations," IEEE Trans. Systems, Man, and Cybernetics, vol. 6, pp. 420-433, June 1976.

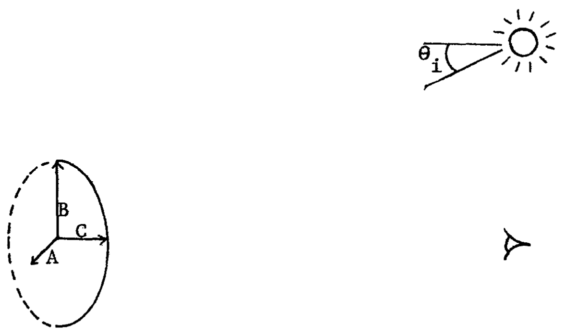

Figure 1.