MODELING AND USING PHYSICAL CONSTRAINTS IN SCENE ANALYSIS*

M. A. Fischler, S. T. Barnard, R. C. Bolles, M. Lowry,
L. Quam**, G. Smith, and A. Witkin**

SRI International, Menlo Park, California 94025

ABSTRACT

This paper describes the results obtained in a research program ultimately concerned with deriving a physical sketch of a scene from one or more images. Our approach involves modeling physically meaningful information that can be used to constrain the interpretation process, as well as modeling the actual scene content. In particular, we address the problems of modeling the imaging process (camera and illumination), the scene geometry (edge classification and surface reconstruction), and elements of scene content (material composition and skyline delineation).

I INTRODUCTION

Images are inherently ambiguous representations of the scenes they depict: images are 2-D views of 3-D space, they are single slices in time of ongoing physical and semantic processes, and the light waves from which the images are constructed convey limited information about the surfaces from which these waves are reflected. Therefore, interpretation cannot be strictly based on information contained in the image; it must involve, additionally, some combination of a priori models, constraints, and assumptions. In current machine-vision systems this additional information is usually not made explicit as part of the machine's data base, but rather resides in the human operator who choses the particular techniques and parameter settings to reflect his understanding of the scene context. This paper describes a portion of the SRI program in machine vision research that is concerned with identifying and modeling physically meaningful information that can be used to automatically constrain the interpretation process. In particular, as an adjunct to any autonomous system with a generalized competence to analyze imaged data of 3-D real-world scenes, we believe that it is necessary to explicitly model and use the following types of knowledge:

(1) Camera model and geometric constraints (location and orientation in space from which the image was acquired, vanishing points, ground plane, geometric horizon, geometric distortion).

(2) Photometric and illumination models (atmospheric and image-processing system intensity-transfer functions, location and spectrum of sources of illumination, shadows, highlights).

(3) Physical surface models (description of the 3-D geometry and physical characteristics of the visible surfaces; e.g., orientation, depth, reflectance, material composition).

(4) Edge classification (physical nature of detected edges; e.g., occlusion edge, shadow edge, surface intersection edge, material boundary edge, surface marking edge).

(5) Delineation of the visible horizon (skyline)

(6) Semantic context (e.g., urban or rural scene, presence of roads, buildings, forests, mountains, clouds, large water bodies, etc.).

In the remainder of this paper, we will describe in greater detail the nature of the above models, our research results concerning how the parameters for some of these models can be automatically derived from image data, and how the models can be used to constrain the interpretation process in such tasks as stereo compilation and image matching.

If we categorize constraints according to the scope of their influence, then the work we describe is primarily concerned with global and extended constraints rather than with constraints having only a local influence. To the extent that constraints can be categorized as geometric, photometric, or semantic and scene dependent, it would appear that we have made the most progress in understanding and modeling the geometric constraints.

---

* The research reported herein was supported by the Defense Advanced Research Projects Agency under Contract Nos. MDA903-79-C-0588 and DAAG29-79-C-0216; these contracts are monitored by the U. S. Army Engineer Topographic Laboratory and by the U. S. Army Research Office.

** Current affiliation: Fairchild Artificial Intelligence Laboratory, Palo Alto, California.

II CAMERA MODELS AND GEOMETRIC CONSTRAINTS

The camera model describes the relationship between the imaging device and the scene; e.g., where the camera is in the scene, where it is looking, and more specifically, the precise mapping from points in the scene to points in the image. In attempting to match two views of the same scene taken from different locations in space, the camera model provides essential information needed to contend with the projective differences between the resulting images.

In the case of stereo reconstruction, where depth (the distance from the camera to a point in the scene) is determined by finding the corresponding scene point in the two images and using triangulation, the camera models (or more precisely, the relative camera model) limit the search for corresponding points to one dimension in the image via the "epipolar" constraint. The plane passing through a given scene point and the two lens centers intersects the two image planes along straight lines; thus a point in one image must lie along the corresponding (epipolar) line in the second image, and one need only search along this line, rather than the whole image to find a match.

When human interaction is permissible, the camera model can be found by having the human identify a number of corresponding points in the two images and using a least-squares technique to solve for the parameters of the model [5]. If finding the corresponding points must be carried out without human intervention, then the differences in appearance of local features from the two viewpoints will cause a significant percentage of false matches to be made; under these conditions, least squares is not a reliable method for model fitting. Our approach to this problem [3] is based on a philosophy directly opposite to that of least-squares -- rather than using the full collection of matches in an attempt to "average out" errors in the model-fitting process, we randomly select the smallest number of points needed to solve for the camera model and then enlarge this set with additional correspondences that are compatible with the derived model. If the size of the enlarged compatibility set is greater than a bound determined by simple statistical arguments, the resulting point set is passed to a least-squares routine for a more precise solution. We have been able to show that as few as three correspondences are sufficient to directly solve for the camera parameters when the three-space relationships of the corresponding points are known; a recent result [13] indicates that 5 to 8 points are necessary to solve for the relative camera model parameters when three space information is not available a priori.

The perspective imaging process (the formation of images by lenses) introduces global constraints that are independent of the explicit availability of a camera model; particularly important are the detection and use of "vanishing points." A set of parallel lines in 3-D space, such as the vertical edges of buildings in an urban scene, will project onto the image plane as a set of straight lines intersecting at a common point. Thus, for example, if we can locate the vertical vanishing point, we can strongly constrain the search for vertical objects such as telephone or power poles or building edges, and we can also verify conjectures about the 3-D geometric configuration of objects with straight edges by observing which vanishing points these edges pass through. The two horizontal vanishing points corresponding to the rectangular layout of urban areas, the vanishing point associated with a point of illumination [8], and the vanishing point of shadow edges projected onto a plane surface in the scene, provide additional constraints with special semantic significance. The detection of clusters of straight parallel lines by finding their vanishing points can also be used to automatically screen large amounts of imagery for the presence of man-made structures.

The technique we have employed to detect potential vanishing points involves local edge detection by finding zero-crossings in the image convolved with both Gaussian and Laplacian operators [9], fitting straight line segments to the closed zero-crossing contours, and then finding clusters of intersection points of these straight lines. In order to avoid the combinatorial problem of computing intersection points for all pairs of lines, or the even more unreasonable approach of plotting the infinite extension of all detected line segments and noting those locations where they cluster, we have implemented the following technique. Consider a unit radius sphere physically positioned in space somewhere over the image plane (there are certain advantages to locating the center of the sphere at the camera focal point if this is known, in which case it becomes the Gaussian sphere [6,7], but any location is acceptable for the purpose under consideration here). Each line segment in the image plane and the center of the sphere define a plane that intersects the sphere in a great circle -- if two or more straight lines intersect at the same point on the image plane, their great circles will intersect at two common points on the surface of the sphere, and the line passing through the center of the sphere and the two intersection points on the surface of the sphere will also pass through the intersection point in the image plane.

III EDGE CLASSIFICATION

An intensity discontinuity in an image can correspond to many different physical events in the scene, some very significant for a particular purpose, and some merely confusing artifacts. For example, in matching two images taken under different lighting conditions, we would not want to use shadow edges as features; on the other hand, shadow edges are very important cues in looking for (say) thin raised objects. In stereo matching, occlusion edges are boundaries that area correlation patches should not cross (there will also be a region on the "far" side of an occlusion edge in which no matches can be found); occlusion edges also define a natural distance progression in

an image even in the absence of stereo information. If it is possible to assign labels to detected edges describing their physical nature, then those interpretation processes that use them can be made much more robust.

We have implemented an approach to detecting and identifying both shadow and occlusion edges, based on the following general assumptions about images of real scenes:

(1) The major portion of the area in an image (at some reasonable resolution for interpretation) represents continuous surfaces.

(2) Spatially separated parts of a scene are independent, and their image projections are therefore uncorrelated.

(3) Nature does not conspire to fool us; if some systematic effect is observed that we normally would anticipate as caused by an expected phenomena due to imaging or lighting, then it is likely that our expectations provide the correct explanation; e.g., coherence in the image reflects real coherence in the scene, rather than a coincidence of the structure and alignment of distinct scene constituents.

Consider a curve overlayed on an image as representing the location of a potential occlusion edge in the scene. If we construct a series of curves parallel to the given one, then we would expect that for an occlusion edge, there would be a high correlation between adjacent curves on both sides of the given curve, but not across this curve. That is, on each side, the surface continuity assumption should produce the required correlation, but across the reference curve the assumption of remote parts of the scene being independent should produce a low correlation score. In a case where the reference curve overlays a shadow edge, we would expect a continuous high (normalized) correlation between adjacent curves on both sides and across the reference curve, but the regression coefficients should show a discontinuity as we cross the reference curve. This technique is described in greater detail in [14]. Figure 1 shows experimental results for an occlusion edge.

IV INTENSITY MODELING (and Material Classification

Given that there is a reasonably consistent transform between surface reflectance and image intensity, the exact nature of this transform is not required to recover rather extensive information about the geometric configuration of the scene. It is even reasonable to assume that shadows and highlights can be detected without more precise knowledge of the intensity mapping from surface to image; but if we wish to recover information about actual surface reflectance or physical composition of the scene, then the problem of intensity modeling must be addressed.

Even relatively simple intensity modeling must address three issues: (1) the relationship between the incident and reflected light from the surface of an object in the scene as a function of the material composition and orientation of the surface; (2) the light that reaches the camera lens from sources other than the surface being viewed (e.g., light reflected from the atmosphere); and (3) the relationship between the light reaching the film surface and the intensity value ultimately recorded in the digital image array.

Our approach to intensity modeling assumes that we have no scene-specific information available to us other than the image data. We use a model of the imaging process that incorporates our knowledge of the behavior of the recording medium, the properties of atmospheric transmission, and the reflective properties of the scene materials. In particular for aerial imagery recorded on film, we use an atmospheric model that assumes a constant amount of atmospheric reflectance independent of scene radiation, a film model that assumes a logrithmic relation between incoming radiation intensity and film density, and a surface reflectance model that assumes Lambertian behavior (the reflected light is proportional to the incident light; the constant of proportionality is a function of the surface material; and the relative brightness of the surface is independent of the location of the viewer). We identify a few regions of known material in a scene -- three materials are sufficient -- to calibrate our model to the particular image. The resultant model is used to transform the given image into a new image depicting the actual scene reflectances.

Our intensity model has the form

$$d = a*\log(r+b) + c$$

where d is the image intensity, r the scene reflectance, a and c parameters associated with the film process, and b is the ratio of atmospheric backscatter to scene illumination. We determine a,b and c by fitting our model to the identified (d,r) pairs. The fitting is achieved by guessing b -- we know b lies in the range 0 to 1 -- applying the least squares method to the resultant linear equation to calculate a,c, and the residual sum, and adjusting b to minimize this residual sum.

The resultant reflectance image has allowed reasonable material labeling and image segmentation to be achieved on the basis of the reflectance information alone.

V SHADOW DETECTION (and Raised Object Cueing)

The ability to detect and properly identify shadows is a major asset in scene analysis. For certain types of features, such as thin raised objects in a vertical aerial image, it is often the case that only the shadow is visible. Knowledge of the sun's location and shadow dimensions frequently allows us to recover geometric information about the 3-D structure of the objects casting the shadows, even in the absence of stereo data [8,10];

but perhaps just as important, distinguishing shadows from other intensity variations eliminates a major source of confusion in the interpretation process.

Given an intensity discontinuity in an image, we can employ the edge labeling technique described earlier to determine if it is a shadow edge. However, some thin shadow edges are difficult to find, and if there are lots of edges, we might not want to have to test all of them to locate the shadows. We have developed a number of techniques for locating shadow edges directly, and will now describe a simple but effective method for finding the shadows cast by thin raised objects (and thus locating the objects as well).

We assume we either know the approximate sun direction, or equivalently, the shadow vanishing point. We first employ a thin line detector oriented parallel to the sun direction at every location in the image, and then apply a moving-window averaging technique in the sun's direction to further enhance the line detector's response and reduce noise. The result of these operations is to smear both the noise and the thin shadow lines (Figure 2). We next thin the shadow lines, eliminate all weak responses, and overlay the result on the original image (Figure 3). The foot of each shadow line now points to the base of the thin raised object casting the shadow. Given the results from two (or more) images taken at different times, the intersections of shadow lines locates the objects more precisely and also eliminates false alarms.

VI VISUAL SKYLINE DELINEATION

Although not always a well defined problem, delineation of the land-sky boundary provides important constraining information for further analysis of the image. Its very existence in an image tells us something about the location of the camera relative to the scene (i.e., that the scene is being viewed at a high-oblique angle), allows us to estimate visibility (i.e., how far we can see -- both as a function of atmospheric viewing conditions, and as a function of the scene content), provides a source of good landmarks for (autonomous) navigation, and defines the boundary beyond which the image no longer depicts portions of the scene having fixed geometric structure.

In our analysis, we generally assume that we have a single right-side-up image in which a (remote) skyline is present. Confusing factors include clouds, haze, snow-covered land structures, close-in raised objects, and bright buildings or rocks that have intensity values identical to those of the sky (a casual inspection of an image will often provide a misleading opinion about the difficulty of skyline delineation for the given case). Our initial approach to this problem was to investigate the use of slightly modified methods for linear delineation [4] and histogram partitioning based on intensity and texture measures; we employ fairly simple models of the relationship between land, sky, and cloud brightness and texture.

Our experience to date, on a data base of 15 scenes, leads us to believe that although we can obtain reasonably good results when the confusing factors mentioned above do not dominate the scene, we still make local mistakes which will require more sophisticated reasoning to eliminate; to the extent that the confusing factors become more prominent, the problem can be made arbitrarily hard. Figure 4 shows a typical image and the skyline delineation we have obtained for it. Prior work on this topic, employing considerably more semantic knowledge then in our approach, is contained in Sloan [11].

VII SURFACE MODELING

Obtaining a detailed representation of the visible surfaces of the scene, as (say) a set of point arrays depicting surface orientation, depth, reflectance, material composition, etc., is possible from even a single black and white image [12,2]. A large body of work now exists on this topic, and although directly relevant to our efforts, it is not practical to attempt a discussion of this material here. There is, however, one key difference between surface modeling and the other topics we have discussed -- the extent to which the particular physical knowledge modeled constrains the analysis of other parts of the scene. In this paper we have been primarily concerned with physical models that provide global or extended constraints on the analysis; surface modeling via point arrays provides a very localized constraining influence.

VIII CONSTRAINT-BASED STEREO COMPILATION

The computational stereo paradigm encompasses many of the important task domains currently being addressed by the machine-vision research community [1]; it is also the key to an application area of significant commercial and military importance -- automated stereo compilation. Conventional approaches to stereo compilation, based on finding dense matches in a stereo image pair by area correlation, fail to provide acceptable performance in the presence of the following conditions typically encountered in mapping cultural or urban sites: widely separated views (in space or time), wide angle views, oblique views, occlusions, featureless areas, repeated or periodic structures. As an integrative focus for our research, and because of its potential to deal with the factors that cause failure in the conventional approach, we are constructing a constraint-based stereo system that encompasses many of the physical modeling techniques discussed above. It is not our intent to discuss this system here, but rather to indicate the framework in which the distinct geometric, photometric, and semantic constraints will interact; Figure 5 shows some examples of this interaction.

IX CONCLUDING COMMENTS

When a person views a scene, he has an appreciation of where he is relative to the scene, which way is up, the general geometric configuration of the surfaces (especially the support and barrier surfaces), and the overall semantic context of the scene. The research effort we have described is intended to provide similar information to constrain the more detailed interpretation requirements of machine vision (e.g., such tasks as stereo compilation and image matching).

REFERENCES

1. S. T. Barnard and M. A. Fischler, "Computational Stereo," ACM Surveys, 1982 (in press).

2. H. G. Barrow and J. M. Tenenbaum, "Recovering intrinsic scene characteristics from images," in Computer Vision Systems (A. Hanson and E. Riseman ed.) Academic Press, pp. 3-26 (1978).

3. M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography," CACM, Vol. 24(6), pp. 381-395 (June 1981).

4. M. A. Fischler, J. M. Tenenbaum, and H. C. Wolf, "Detection of roads and linear structures in low-resolution aerial imagery using a multisource knowledge integration technique," Computer Graphics and Image Processing, Vol. 15(3), pp. 201-223 (March 1981).

5. D. B. Gennery, "Least-squares stereo-camera calibration," Stanford Artificial Intelligence Project Internal Memo, Stanford University (1975).

6. D. A. Huffman, "Realizable configuration of lines in pictures of polyhedra," in Machine Intelligence (Elcock and Michie, ed.), Edinburgh University Press, Edinburgh, Scotland, pp. 493-509 (1977).

7. J. R. Kender, "Shape from Texture," (Ph.D. Thesis, Report No. CMU-CS-81-102) Carnegie-Mellon University, Pittsburgh, Pennsylvania (November 1980).

8. D. G. Lowe and T. O. Binford, "The interpretation of three-dimensional structure from image curves," IJCAI-81, pp. 613-618 (1981).

9. D. Marr and E. C. Hildreth, "Theory of edge detection," MIT AI Lab Memo 518 (1979).

10. S. Shafer and T. Kanade, "Using shadows in finding surface orientations," (Report No. CMU-CS-82-100) Carnegie-Mellon University, Pittsburgh, Pennsylvania (January 1982).

11. K. R. Sloan, Jr., "World model driven recognition of natural scenes," University of Pennslyvania, Philadelphia, Pennsylvania (June 1977).

12. J. M. Tenenbaum, M. A. Fischler, and H. G. Barrow, "Scene Modeling: A structural basis for image description," Computer Graphics and Image Processing," Vol. 12(4), pp. 407-425 (April 1980).

13. R. Y. Tsai and T. S. Huang, "Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces," University of Illinois, Urbana, Illinois (August 1981).

14. A. Witkin, "Recovering intrinsic scene characteristics from images," SRI Project 1019, Interim Technical Report, SRI International, Menlo Park, California (September 1981).

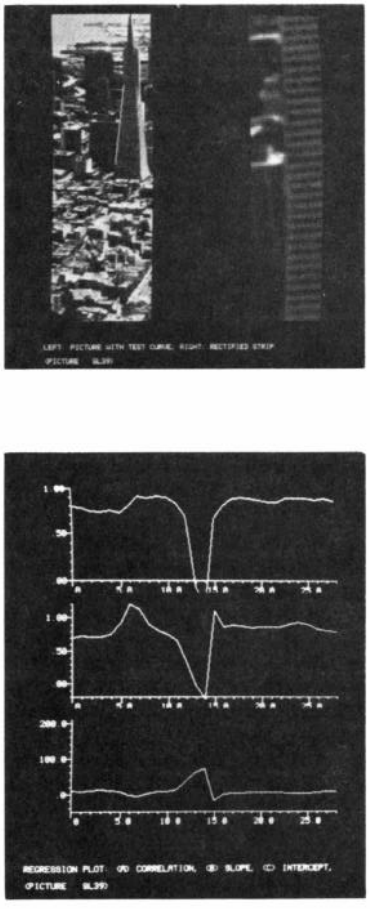


Figure 1

Example of an Occlusion Edge

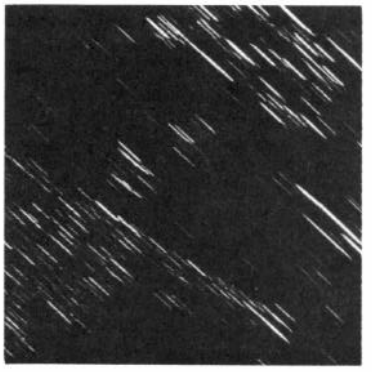


Figure 2

Detection of Thin Shadow Lines (result of line detection and integration)

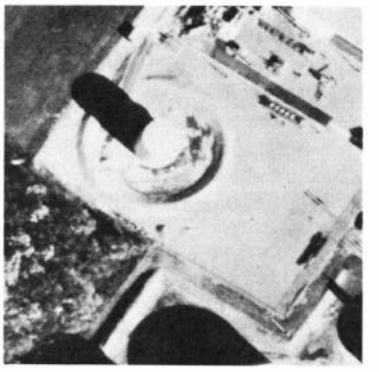
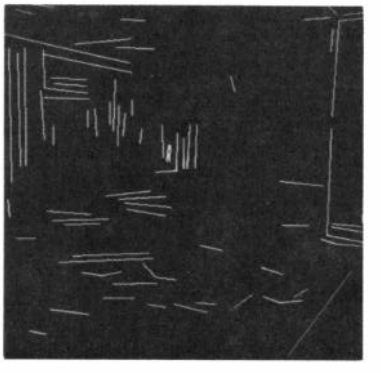


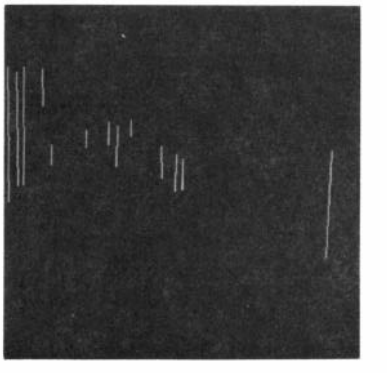
Figure 3

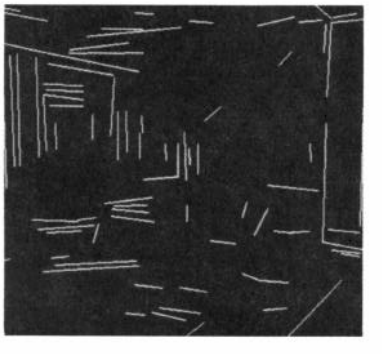
Highest Likelihood Shadow Lines Overlayed on Original Image

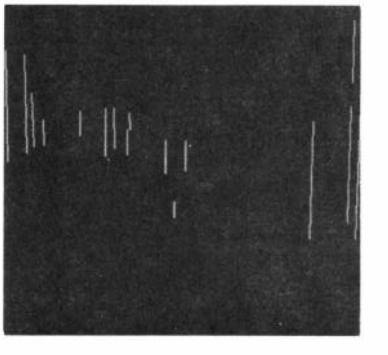
Figure 4

Results of Skyline Delineation









(a) (b) (c)

Figure 5

Constraint-Based Detection and Matching of Vertical Edges

(a) stereo-images-top image left view, bottom image right view

(b) lines found

(c) vertical lines found