A Computational Paradigm for Three Dimensional Scene Analysis

James L. Crowley

CMU-RI-TR-84-11

The Robotics Institute
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

April 1984

A version of this paper appeared in the Conference Proceedings of the IEEE Computer Society, Workshop on Computer Vision: Representation and Control, May 1984, Annapolis, Maryland.

Table of Contents

List of Figures

Abstract

This paper presents a computational paradigm for a system which will dynamically model the contents of a three dimensional scene. The dynamic scene model may be made available to processes which analyze and interpret the scene as a composition of objects, and processes which plan and execute actions based on the composition of the surfaces or objects in the scene. This computational paradigm is presented as a collection of processes and data structures, many of which are currently areas of active research.

The system receives information in the form of a time sequence of stereo images. These images are immediately converted into an "initial representation" which facilitates the processing of later stages. The "initial representation" is then passed to a number of independent processes called "shape experts". Each shape expert extracts information about three dimensional surfaces from a different source. Surface information is integrated with the information obtained over time to maintain a "Composite Surface Model". The Composite Surface Model is then made available to processes for planning, analysis, or object recognition.

The framework is introduced, and then each of the components are examined. The problems associated with each component are discussed, and a brief description is given of current research in that area.

1 Introduction

Vision has evolved over millions of years as a mechanism for providing animals with a means of sensing their environment. Vision provides an internal model of the environment that is constantly updated to reflect the current state of the environment. This internal model provides the basis for such tasks as:

- Planning actions (reasoning through "mental simulation")
- Supervising the execution of actions (visual feedback)
- Learning about places, objects, and other animals
- Recognizing pre-learned places, objects, and animals.

The common basis for all of these tasks is a reliable, dynamically updated description of the three dimensional objects in the animal's environment. This paper is an exploration and exposition of the computational problems involved in maintaining such a dynamic description.

This paper is not a description of an existing system; it is an attempt to elucidate the components of an idealized vision system. The first part of the paper presents a computational framework for dynamic three dimensional scene analysis. The problems involved in each of the components are then briefly examined and some of the research approaches to their solution are presented.

1.1 Dynamic Scene Analysis

A scene is a three dimensional (3-D) setting composed of physical objects. Modeling a 3-D scene is a process of constructing a description for the surfaces of the objects of which the scene is composed. Dynamic modeling is a process of maintaining the veracity of the surface model in the presence of changes in the scene. This paper presents a computational paradigm for a system which will dynamically model the contents of a three dimensional scene. The dynamic scene model may be made available to processes which analyze and interpret the scene as a composition of objects, and processes which plan and execute actions based on the composition of the surfaces or objects in the scene.

An important concept presented in this framework is the "Composite Surface Model". The composite surface model is a data structure which represents the surface information obtained, over time, by a large number of independent processes. A composite surface model plays two fundamental roles in a scene analysis system:

1. The composite surface model is the data structure in which surface information from different sources and different views obtained over time is integrated.

2. The composite surface model is the data structure on which 3-D object recognition, action planning, process monitoring and other task oriented processes operate.

The composite surface model serves as the view of the world for processes which must plan actions, monitor the execution of actions, recognize objects, or otherwise interpret the surface data. When such processes operate directly on the output of sensors or sensor interpretation processes they are vulnerable to a number of problems.

- They are sensitive to transient errors or "illusions" from the sensor or the sensor interpretation process.
- They are limited in scope to what the sensor can see "now".
- They can not easily reconcile inconsistent information from different sources.

The problems of obtaining a consistent interpretation over time, from different views, and in the face of information with varying degrees of confidence is are fundamental to perception. These problems are common to a variety of different tasks which are based on sensing the surfaces in a scene. It is important to decouple these problems from the variety of processes which may operate on the sensor information.

1.2 A Framework for a System for Dynamic Scene Modeling

Figure 1 shows the components for a system which can describe a 3-D scene in terms of the surfaces of physical objects. Images are first expressed in a symbolic representation called the initial representation, which is efficient for interpretation. Shape information is then extracted from the initial representation by simple stereo matching, generalized stereo, shape from shading, shape from texture, occluding boundaries and other "shape experts". The result is passed to a process which maintains a viewer-centered description of the surfaces in the scene, called the "composite surface model". It is composite in the sense that it is composed over time, from many views and from information from a number of independent processes.

The composite surface model is available for a variety of tasks. One popular use for the composite surface model is to construct and maintain a "scene description". A scene description is a data structure which represents a scene as a composition of labeled physical objects at relative spatial positions. The overall problem in scene description is to develop algorithms and data structures which enable a program to locate and identify the physical objects in a scene from 2-D gray-scale images. A scene description may be produced by matching collections of surfaces in the composite surface model to models of the shape of objects from a data base of known objects. The

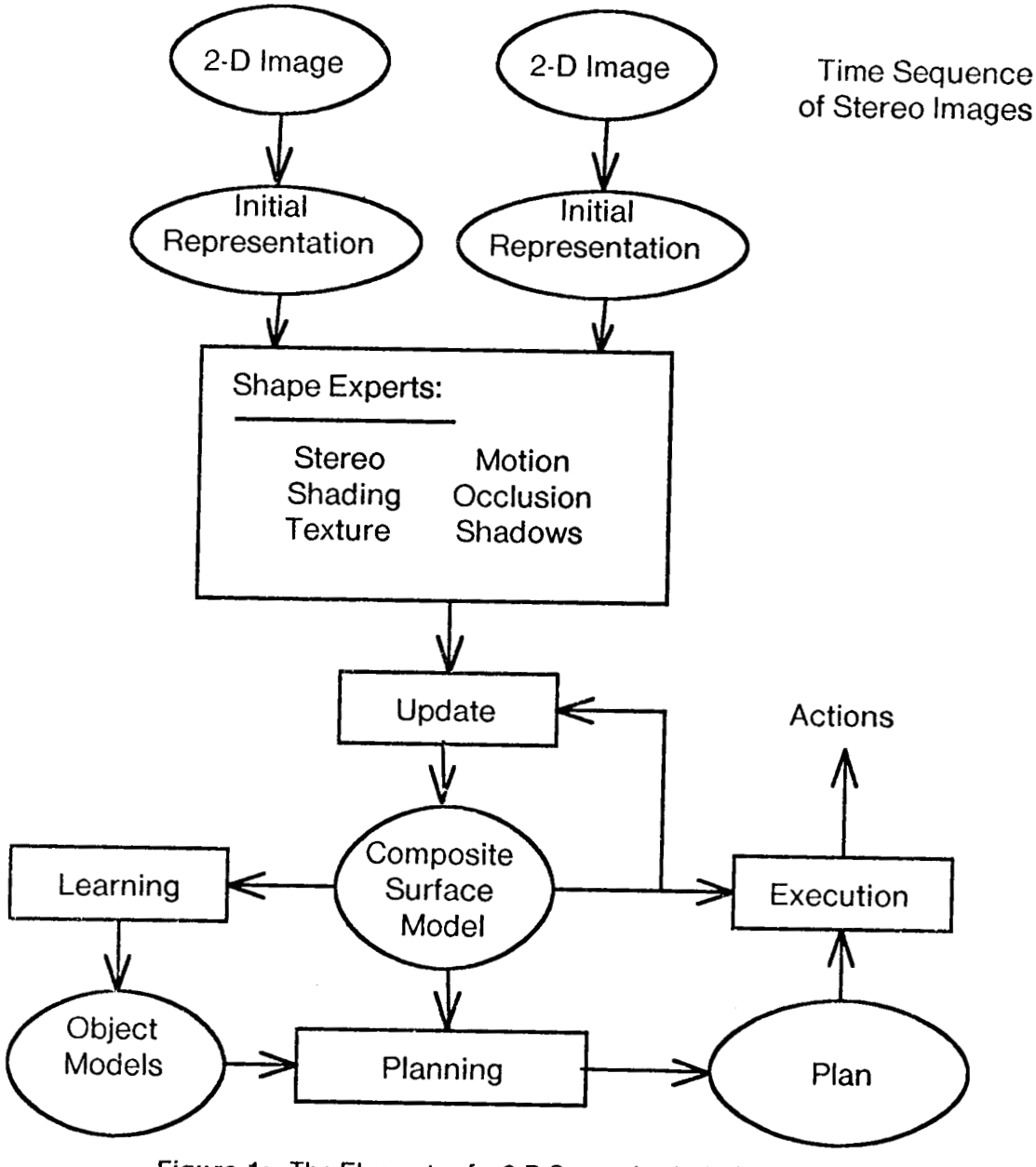

Figure 1: The Elements of a 3-D Scene Analysis System

configuration of objects which are identified in the scene, and their 3-D positions are then represented in the scene description.

The composite surface model may also be used by a system which plans and executes actions by a robot arm or a mobile robot. Such a system may be designed to move an arm or a mobile robot body among the objects in the scene without touching them or to manipulate the objects in a scene to accomplish a specified goal. A composite surface model is useful for both planning and monitoring the execution of a task.

The following sections examine each of these components in greater detail and describes the research problems presented by each component.

2 The Initial Representation

The initial representation must represent the information in a gray scale image in such a manner as to make the task of the various shape experts easier. Processes which obtain surface shape from simple stereo, generalized stereo, occlusion contours, motion and texture are all based on the matching of gray scale shapes between images. Updating the composite surface model also requires matching information from the current composite surface model to information obtained from the shape experts. Thus the initial representation must facilitate the determination of the correspondence of gray scale patterns between images.

Many researchers have found that correspondence matching can be made more efficient if it is performed at multiple resolutions. [24], [23]. Marr and Grimson [14] have developed stereo matching based on the zero crossings in band-pass images at four resolutions. These band-pass images are formed by convolution with a Difference of Gaussian filter, which is defined to approximate a form of Laplacian of Gaussians.

We have developed a multiple resolution representation which is efficient for correspondence matching. This representation is based on the peaks and ridges in a Difference of Low-Pass (DOLP) transform [7]. The DOLP transform is a reversible transform which expresses an image (or signal) as a sequence of band-pass images (or signals) [8]. This sequence of band-pass images is sometimes referred to as a "Laplacian Pyramid". We have also defined a fast computation algorithm for the DOLP transform, based on the techniques of resampling and of cascade filtering using Gaussian filters. The result of applying this algorithm to an N by N image is $2\ \mathrm{Log}_2(N)$ band-pass images which are close approximation to the size scaled "Laplacian" images. The difference of Gaussian impulse

responses of these images are copies of a prototype difference of Gaussian filter, scaled in size at powers of the square root of 2.

The local peaks and ridges from each band-pass image express gray-scale shape in a multi-resolution tree, which has the property of permitting shapes to be matched despite changes in size, orientation, or position of forms (gray-scale patterns) in an image [9]. This representation is useful for measuring surface shape from simple stereo [6], generalized stereo, occlusion contours, and optical flow [4]. This representation is also relatively immune to degradation due to noise, and separates image forms based on their size. Thus it is useful for detecting and measuring shape from textured patterns.

The patterns which are described by this representation are "gray-scale shapes" or "forms". It is not necessary for a pattern to have a uniform intensity for it to have a well defined description in this representation. In this representation, a form is described by a tree of symbols which represent the structure of the form at every resolution. There are four types of symbols { M, L, P, R } which mark locations (x, y, k) in the DOLP three space where a band-pass filter of radius R_k is a local "best-fit" to the form.

Figure 2 shows an example of the use of peaks and ridges for representing a uniform intensity form. This figure shows the outline of a dark rhomboid on a light background. Circles illustrate the position and radii of band-pass filters whose positive center lobes best fit the rhomboid. Below the rhomboid is part of the graph produced by detecting and linking peaks and ridges in the sampled DOLP transform.

A description in this representation contains a small number of symbols at the root. These symbols describe the global (or low-frequency) structure of a form. At lower levels, this tree contains increasingly larger numbers of symbols which represent more local details. The correspondence between symbols at one level in the tree constrains the possible set of correspondences at the next higher resolution level.

The description is created by detecting local positive maxima and negative minima in one dimension (ridges) and two dimensions (peaks) in each band-pass image of a DOLP transform. Local peaks in the DOLP three space define locations and sizes at which a DOLP band-pass filter best fits a gray-scale pattern. These points are encoded as symbols which serve as landmarks for matching the information in images. Peaks of the same sign which are in adjacent positions in adjacent band-pass images are linked to form a tree. During the linking process, the largest peak along each branch is detected. This largest peak serves as a landmark which marks the position and size of a gray-scale

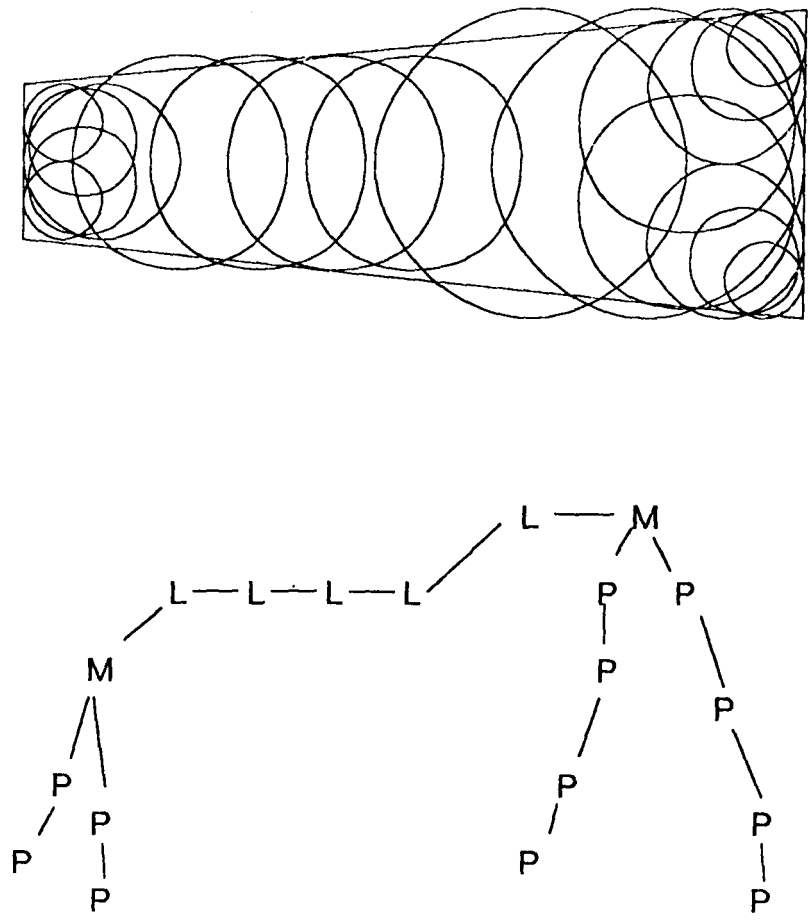

Figure 2: A Rhomboidal Form and its Representation:

In the upper part of this figure the rhomboidal form is outlined in solid straight lines. The description is for such a form which is dark on a light background. Circles indicate the locations and sizes where the band-pass filters from a sampled DOLP transform produced 3-Space peaks (M-nodes), 2-Space peaks (P-nodes), and 3-Space ridges (L-nodes). The structure of the resulting description is shown in the lower part of the figure. The description of the "negative shape" which surrounds this form is not presented.

form. The paths of the other peaks, which are attached to such landmarks, provide further description of the form, as well as continuity with structure at other resolutions. Further information is encoded by detecting and linking two-dimensional ridge points in each band-pass image, and three-dimensional ridge points within the DOLP three space. The ridges in each band-pass image link the peaks in that image which are part of the same form. The three-dimensional ridges link the largest peaks that are part of the same form and provide a description of elongated forms.

In several of the shape experts described below, the fundamental problem is to determine the

correspondence between forms in two or more images. This representation has properties which greatly simplify the process of determining the correspondence of patterns in two images.

1. Only peaks correspond to peaks. The existence of peaks or P-nodes provides a set of landmarks which can be used as tokens in the matching process.

2. The multi-resolution structure of the representation permits the correspondence process to commence with the most global M-nodes for each form. Since very few such symbols exist at the coarsest resolution, the complexity of this process is kept small.

3. The connectivity of P-paths permits the match information from a coarse resolution to constrain the possible set of matches at the next higher-resolution level. Thus what could be a very large graph matching problem is repeatedly partitioned into several small problems.

3 The Shape Experts

Viewer centered information of the 3-D shape of surfaces may be obtained from a sequence of images by a variety of techniques. All of these techniques provide surface information for the process which maintains the "Composite Surface Model". The following sections examine a number of techniques for obtaining surface information from gray scale-images of a scene. In general, these shape experts operate on a particular subset of the phenomena which occur in an image. In many cases the different shape experts complement each other nicely, with each shape expert providing information in cases where the others fail.

3.1 Depth from Simple Stereo

Obtaining depth information from a stereo pair of images involves two problems:

1. Finding the difference in position of pixels in the two images which correspond to the same physical (3-D) point, and

2. Converting this difference of position to an estimate of depth.

When all of the physical parameters concerning imaging and the positions of the two cameras are known, the second problem is reduced to a problem of geometry. The most difficult part of this problem is usually seen as calibrating the system to learn these camera parameters. Gennery [13] has developed a method for discovering the relative camera model from a few sparse matches. The development of techniques to discover the camera models for a pair of stereo cameras continues to be an area of active research. The first problem, known as the "stereo correspondence problem" is more difficult. The stereo correspondence problem requires matching of signals in the two camera images based on grays-scale patterns.

3.1.1 Conditions for Simple Stereo

It is well known that a pair of stereo cameras can be configured so that stereo matching may be accomplished as a 1-D matching problem [10]. We refer to such stereo matching as "simple stereo". Simple stereo is possible whenever:

1. The two images are taken at approximately the same time.

2. The two cameras are approximately the same distance from the scene.

3. Knowledge of the camera positions provides knowledge of the epipolar planes. That is, the planes passing through the shared image point and the centers of the two cameras.

The lines along which the epipolar planes pass through the image are known as the epipolar lines. If the conditions for simple stereo are true, the gray-scale shapes along the epipolar lines in one image are constrained to match to shapes along the epipolar lines in the second stereo image. If the equation of the epipolar lines in the two images are known, then stereo matching is reduced to a series of 1-D matching problems. The simplest such case is for two cameras with

- identical optics

- co-planar image planes, and

- a displacement that has no vertical (y axis) component.

In such a case, the matching may be performed on a row by row basis. However, these conditions are not absolutely necessary to obtain surface information from simple stereo; whenever the conditions for simple stereo which were cited above are true, 1-D matching techniques will work for any pair of signals from the same epipolar plane.

3.1.2 1-D Correspondence Matching

Many techniques for solving for the stereo correspondence involve matching selected points or regions in two (or more) stereo images. A popular approach is to detect small patterns which appear especially matchable, and only search for matches for these points. Moravec's stereo matching algorithm is an example of this approach [24]. Moravec developed an "interest operator" which selects small patterns where the variance is high. Ideally, with this approach, the matching should be performed on points whose shape and position are invariant to the viewing angle and position. Regardless, the result is a depth estimate at a discrete set of points. Depth at intermediate points must be supplied by assumptions about the surface geometry, and often a costly interpolation step is required.

Kanade and Ohta have developed a matching algorithm for simple stereo based on dynamic

programming [17]. The algorithm was first developed to perform an optimal 1-D match between scan rows of the two stereo images. The algorithm was then generalized to perform a 2-D match, in which the second dimension is consistency across the scan rows. This algorithm uses a continuity constrained to find the most likely collection of surfaces for a pair of stereo image.

Much work in stereo matching has been done with random dot stereo-gram images. Using such images, Marr [22] and Grimson [14] have developed algorithms that make use of the zero crossings in the convolution of the images with Difference of Gaussian (DOG) filters. Difference of Gaussian filters at four resolutions are used, and the result of matching at low resolution is used to constrain the matching at higher resolutions. An interpolation step, which is based on a surface consistency constraint, is used to provide an estimate of the depth at points which are not on zero-crossings.

When stereo matching is based on a multiple resolution representation, such as the initial representation described above, it becomes possible to obtain an estimate of depth over most of the image [6]. These depth estimates have a variable resolution. The resolution of the depth estimate depends on the size of the patterns that can be matched at a location in the picture, where size refers to the size of a difference of low-pass filter which is used to represent the pattern in the image. Where no match can be found, the depth estimate may be found from the other shape experts.

This depth estimate can be made with no assumptions about the continuity of surfaces that produced the images. However, some assumptions about the surfaces in the scene can facilitate the correspondence matching process. The depth estimate from this method tends to have a lower resolution at points away from sharp edges, but these are the points where "shape from shading" techniques are the most useful. In this sense, this stereo matching technique complements other sources of depth information.

A good review of stereo matching research, for both the simple stereo and generalized stereo cases is presented by Barnard and Fischler [2]. Much of the early work in stereo was done with aerial images. The survey papers by Konecy and Pape [19] and Case [5] provide a good review of the efforts in this area.

3.2 Generalized Stereo

When the relative positions of the two cameras are not constrained so that the image planes are co-planar, epipolar lines will not necessarily be parallel. This occurs when a camera is moving along an arbitrary 3-D line in the scene. In this case, correspondence must be made between gray-scale shapes that can be displaced in any direction in the image plane and that can even change size. This is called "Generalized Stereo" or "Motion Stereo".

A popular approach to correspondence matching in the case where the camera is moving is to determine the lines of "optical flow" [4]. Given reliable feature points, the correspondence algorithms developed by Ullman [28] and by Marr and Poggio [22] can be used to find correspondence. Lucas has developed an algorithm that uses the spatial intensity gradient of the images to iteratively hill climb to a correspondence [20].

When a camera moves forward or backward in a scene, the size of the gray-scale forms in the scene change. Thus, general stereo matching is facilitated by a representation which can be matched despite changes in size, as well as position and orientation. The multi-resolution representation given by peaks and ridges in the DOLP transform has been shown to have this property [9]. We have recently begun an effort to demonstrate stereo matching for the case where the camera has moved in an arbitrary three dimensional direction, using the representation based on peak and ridges in the DOLP transform. The result of such stereo matching is a change in position and a change in size for forms in the two images. These two measurements can be used to determine the distance to the objects which the forms represent. These measurements can also be used to determine the direction of travel of the camera, by detecting the "optical flow" of the forms.

A successful solution to the generalized stereo problem will open applications in mobile robot navigation, as well as in dynamic 3-D scene analysis.

3.3 Contours of Occlusion

Control of camera motion, and freedom to move in arbitrary directions, provides an opportunity to measure another important source of information about 3-D shape: the lines of discontinuity that bound objects. Marr has noted that there are four basic ways that contours of gray level discontinuity (edges) can result in a single image:

1. Discontinuities in distance from the viewer
2. Discontinuities in surface orientation
3. Changes in surface reflectance
4. Illumination effects such as shadows, light sources and highlights.

Interpreting a scene involves determining which source is responsible for each such edge contour.

Discontinuities in distance from the viewer and in surface orientation are particularly important for the composite surface model. These contours mark the boundary of surface patches in the 3-D scene, and figure prominently in many of the representations for composite surface model. Occlusion

contours may be detected by detecting patterns in one image that are occluded in a second image. Gray-level discontinuities resulting from occlusion contours are particularly easy to discriminate from illumination effects and changes in surface reflectance if the system has the ability to move the cameras in a direction perpendicular to the contour. In the case of an occlusion contour, such motion hides or exposes a part of the background pattern against which the contour is defined. Such an occlusion detection process can also detect discontinuities in surface orientation which do not result in image edges. Detecting the occluded or exposed background pattern requires the same sort of correspondence matching required for generalized stereo.

3.4 Shape from Shading

The orientation of a patch of a 3-D surface may be described by the two parameters of its gradient often referred to as (p,q) [3]. The (p,q) plane is known as the "gradient space". When a scene is illuminated by a single source, the observed intensity at a pixel is a function of: the illumination intensity, the angle between the illumination source and the surface patch normal (i), and the angle between the line from the camera and the surface patch normal (e).

For a given configuration of camera, illumination source, and 3-D surface, there are a closed connected set of values of (p,q) that correspond to a given intensity at a pixel. This set of values correspond to a contour in the space of (p,q). If the scene is observed with illumination from a second light source, then a second contour in (p,q) results from the observed intensity. In general there will be two values for (p,q) that lies on both contours. A third image with illumination from yet another 3-D point will disambiguate the situation and give a unique value of (p,q). This shape measurement technique is known as "photometric stereo". Alternatively, knowledge about the possible shapes of objects can be used to disambiguate the shape from two images.

Photometric stereo is not usually practical, but the contours of equal reflectance in (p,q) space can be used in another way to obtain local shape information. Given knowledge of the values of (p,q) for a point, the gray levels along any line from that point map into a contour in (p,q) space, provided that no discontinuities in surface orientation are crossed, that the values of (p,q) are assumed continuous and smoothly varying and that the surface reflectance ("albedo") is constant. Thus knowledge of (p,q) at a point allows knowledge of relative (p,q) at adjacent points.

Pentland [25] has recently shown how to make such estimates from the Laplacian of the image intensity. Pentland's technique assumes that the surface has a lambertian albedo (reflects light equally in all directions), but can permit multiple and extended light sources.

3.5 Shape from Texture

A visual texture is a image region composed of many instances of the same or similar small patterns (texture elements or "texels"). The shape of a surface can be inferred from the pattern of a texture on the surface in at least three ways.

First, if the exact shape of the texture element is known, then transformations in that shape identify a local surface orientation to within 2 values of (p,q).

Second, if we assume that the elements of a texture are all of the same size, then the directional derivatives of the size of the texture elements can be used to determine the the orientation of the surface. Even if the texture elements are not all exactly the same shape, the local derivatives can be based on local ensembles of sizes. This happens, for example, when a human looks at a field of grass.

Third, the assumption that texture elements are co-linear on a surface allows measurement of a 3-D contour on the surface as a 2-D contour in the image.

Stevens has shown how to recover surface shape from co-linear contours of texture elements [27]. Kender has shown a computational paradigm that allows local surface orientation to be recovered from all three of these cases [18].

3.6 Shape from Motion

A rigid object may be described by a set of landmark points on its surface, and a distance and orientation of the vector between these landmark points. When a rigid object moves in a scene, the 3-D length and relative 3-D orientations of these vectors remain constant, while the observed 2-D lengths and relative 2-D orientations change. A sequence of images of a moving rigid object in which a set of keypoints are detected can allow the 3-D lengths and orientations of the lines between keypoints to be determined. The first major study of this ability in humans was performed by Wallach and O'Connell [29]. Subjects were shown the shadow of a rotating wire figure and reported seeing a rigid wire figure turning in space. Johansson showed human subjects films of persons moving around in a dark room with reflective tape on major joints. Subjects could interpret what they saw and could easily tell what the observed person was doing [16].

Ullman [28] undertook a computational study of structure from motion. He showed that under an assumption of orthography (no perspective, i.e. an infinite focal length) it was computationally possible to determine structure from at least three views of at least four points. Roach and Aggarwal

[26] derived similar results for perspective projection, showing that either two views of five points or three views of four points were sufficient. These investigations sought to prove the possibility of shape from motion; the resulting algorithms were quite sensitive to even minor noise. Webb and Aggarwal [30] developed a procedure which used many views and an assumption of rigid rotation about a common point to determine shape from motion.

4 The Composite Surface Model

The composite surface model serves as a common data structure into which the ensemble of shape experts place their interpretation of surface shape in the scene. It also permits each shape expert to read the interpretation from the other shape experts so that it may modify or guide its interpretation in ambiguous situations. The composite surface model is the data structure in which a description is built up from many views taken over time. It is also the data structure in which inconsistent information about surfaces from the different knowledge sources is resolved.

In addition to integrating a surface description from many knowledge sources, the composite surface model is also the data structure on which a variety of "higher level" processes operate. Such processes perform tasks such as:

- Recognize 3-D objects and construct a scene description
- Plan actions with respect to the scene
- Monitor the execution of actions
- Learn the shape of objects, or collections of objects.

The composite surface model is "viewer centered"; it contains a description of surfaces as seen by the viewer from a particular perspective, perhaps including surfaces which are temporarily occluded. It is referred to as composite because it is a composition of information obtained, over time, from different views, and from all of the shape experts. The development of a representation for the composite surface model is currently an important research issue in scene analysis.

The most obvious implementation for a composite surface model is as a "depth map", that is a 2-D array of the form $z = f(x,y)$. The most obvious problem with this representation is how to represent surfaces that are vertical with respect to the viewer. In this case there are multiple surface points at a given location (x,y). A second problem is an inability to represent surfaces which are temporarily occluded.

One possible implementation for the composite surface model is to represent surface regions as patches which are enclosed in closed contours where the surface shape is discontinuous. Each such patch could be approximated by a plane or a second order curve. A surface description of this form can be implemented as a graph of "surface patch elements", with each element linked to its adjacent neighbors. A planar representation composed of triangles has been developed by Fuchs [12]. A more general scheme which employs polyhedral approximations to 3-D objects has been developed by Faugeras et. al. [11]. This group developed an algorithm for segmenting a collection of surface points into a collection of planar patches.

An alternative to representing each patch as a planar or second order element is to represent each patch as a network of surface normals. Such a set of normals can be represented by a spatial proximity graph [15].

A Composite surface model based on representing surface patches also offers an additional attractive property: many matching algorithms devised for edge-based descriptions of images can be generalized for a surface patch description. Matching is fundamental to the integration of information from different shape experts and to the integration of surface information over time.

5 Object Centered Models

Objects have 3-D shapes. In the general case there is no way to know a-priori the 3-D angle or distance from which an object is likely to be seen. Thus a model is needed which describes the complete 3-D shape of an object. The system must be able to determine from this model what surfaces and surface features will be seen from a given viewing angle. Because such a model is represented independent of viewing angle it should be "object centered". That is surfaces in the model are represented relative to an internal coordinate system.

A powerful method for representing the 3-D shapes of objects is the technique known as "generalized cylinders" developed by Agin [1]. A generalized cylinder is described by three components:

1. A spine, or 3-D curve which is the center axis of the object,

2. A cross section, and

3. A sweeping rule which transforms the cross section as it is swept along the spine.

Objects which have more than one spine are described as a configuration of generalized cylinder models. Agin used parametric functions to represent cross sections. Thus a generalized cylinder model of an object was a simplification which ignored many small shapes on a surface.

Marr has proposed a scheme for representing shapes as a hierarchy of generalized cylinders [21]. In this representation scheme, the description at each level is kept very simple. A pointer refers to a more detailed description of each component. Brooks has demonstrated a model driven visual interpretation system named ACRONYM which uses such a representation [3].

Generalized cylinders give a good volumetric description for many types of objects. However, there are many classes of objects for which cylinders are inappropriate as a representation. (For example, a desk surface, or a sheet of paper.) There are also cases where there are several possible ways to fit a generalized cylinder to a shape (consider a cube, for example). A more general representation with a richer vocabulary of shape primitives is needed for a general scene analysis system.

6 Object Matching

Construction of a scene description from a composite surface model involves matching object centered 3-D models to the viewer centered composite surface model. Matching 3-D object models to the composite surface model may proceed from one of two approaches. The so-called "bottom up" approach involves detecting landmarks in the composite surface model, finding corresponding landmarks in the 3-D models, and then using this correspondence to hypothesize the orientation and size parameters for projecting the 3-D model onto the composite surface model for matching. Alternatively, a "top down" strategy may be employed, which uses some "high level expectations" to predict the expected size and 3-D orientation at which an object is likely to be found.

In either case, some number of candidate 3-D models must then be more closely compared to the surfaces in the composite surface model. Such matching requires projecting from the object centered 3-D models into the viewer centered surface model.

7 Summary and Conclusion

This paper has presented a computational paradigm for dynamic 3-D scene analysis. The heart of this paradigm is the construction and dynamic maintenance of an internal description of the surfaces in a scene, called the "composite surface model". A process called "update" combines information from independent "shape experts" with the current contents of the composite surface model.

The components of a dynamic 3-D Scene analysis system were presented as:

- The Initial Representation
- The Shape Experts

- The Composite Surface Model
- Updating the Composite Surface Model
- Object Models

This paper was written to explore the various problems of dynamic three dimensional scene analysis. The problems of the "initial representation", the "shape experts" and the representation of surface information have received increasing attention in the literature recently. The problems of resolving inconsistency between the shape experts and the problem of representing uncertainty in the surface model have not been as widely recognized. This paper was written to present these elements in a combined system, and to briefly review the state of research in each aspect of the system.

References

[1] Agin, G. J. and T. O. Binford.
Computer Description of Curved Objects.
IEEE Trans. Comp. C-25(4):439-448, April , 1976.

[2] Barnard, T. B. and M. A. Fischler.
Computational Stereo.
Computing Surveys 14(4):553-572, December, 1982.

[3] Brooks, R. A.
Symbolic Reasoning among 3-D Objects and 2-D Models.
AI Journal 16:285-348, 1981.

[4] Burt, P. J., X. Xu, and C. Yen.
Multi-Resolution Flow-Through Motion Analysis.
IEEE Trans on PAMI submitted for publication, 1984.

[5] Case, J. B.
Automation in Photogrametry.
Photogram. Eng. Remote Sensing 47(3):335-341, March, 1981.

[6] Crowley, J. L. and A. L. Lowrie.
Multi-Resolution Stereo Correspondence Matching Using the 1-D DOLP Transform.
In Preparation , 1984.

[7] Crowley, J. L.
A Representation for Visual Information.
PhD thesis, Dept. of Elect. Eng., Nov., 1981.

[8] Crowley, J. L. and R. Stern.
Fast Computation of the Difference of Low-Pass Transform.
IEEE Trans on PAMI 6, 1984.

[9] Crowley, J. L. and A. C. Parker.
A Representation for Shape Based on Peaks and Ridges in the Difference of Low-Pass Transform.
IEEE Trans on PAMI 6, March, 1984.

[10] Duda, R. O. and P. E. Hart.
Pattern Classification and Scene Analysis.
Wiley, New York, 1973.

[11] Faugeras, O. D. Hebert, M, and Pauchon, E.
Segmentation fo Range Data into Planar and Quadratic Surfaces.
Proceedings of the Conf. on CVPR-83 20(10):8-13, June, 1983.

[12] Fuchs, H., Kedem, Z. and Uselton, S. P.
Optimal Surface Reconstruction from Planar Contours.
CACM 20(10):693-702, October, 1977.

[13] Gennery, D. B.
Stereo Camera Calibration.
Proceedings of the Image Understanding Workshop :101-107, 1979.

[14] Grimson, E.
From Images to Surfaces.
MIT Press, Cambridge, Mass., 1981.

[15] Henderson, T. C.
Efficient 3-D Object Representation for Industrial Vision Systems.
IEEE Trans. on PAMI PAMI-5(6):609-618, November, 1983.

[16] Johansson, G.
Visual Motion Perception.
Scientific American 232(6):76-88, June, 1975.

[17] Ohta, Y. and T. Kanade.
Stereo by Intra- and Inter-Scanline Search Using Dynamic Programming.
Technical Report CMU-CS-83-162, Dept. Comp. Science, Carnegie-Mellon University, October, 1983.

[18] Kender, J.
Shape from Texture.
PhD thesis, Dept. of Comp. Sci. C-MU, 1980.

[19] Konecny, C. and D. Pape.
Correlation Techniques and Devices.
Photogram. Eng. Remote Sensing 47(3):323-333, March, 1981.

[20] Lucas, B. and T. Kanade.
An Iterative Image Registration Technique with an Application to Stereo Vision.
Procedings of the Image Understanding Workshop, 1981 :121-130, 1981.

[21] Marr, D. and H. K. Nishihara.
Representation and Recognition of the Spatial Organization of Three Dimensional Structure.
Proc. of R. Soc. London (B) 200:269-294, 1978.

[22] Marr, D. and Poggio, T.
A Computational Theory of Human Vision.
Proc. R. Soc. Lond. B , 1979.

[23] Marr, David.
Vision.
W. H. Freeman and Co., San Francisco, 1982.

[24] Moravec, H. P.
Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover.
PhD thesis, Stanford University, September, 1980.

[25] Pentland, A. P.
Local Shading Analysis.
IEEE Trans on PAMI 6, 1984.

[26] Roach, J. W. and J. K. Aggarwal.
Determining the Movement of Objects from a Sequence of Images.
IEEE PAMI 2(6):554-562, June, 1980.

[27] Stevens, K. A.
Surface Perception from Local Analysis of Texture and Contour.
PhD thesis, MIT Dept of EE, 1979.

[28] Ullman, S.
The Interpretation of Visual Motion.
MIT Press, Cambridge, Mass, 1979.

[29] Wallach, H. and D. N. O'Connell.
The Kinetic Depth Effect.
J. Exp. Psych. 45(4):205-217, 1953.

[30] Webb, Jon A. and J. K. Aggarwal.
Structure from Motion of Rigid and Jointed Objects.
Artificial Intelligence 19:107-130, 1982.