Active Stereo: Integrating Disparity, Vergence, Focus, Aperture, and Calibration for Surface Estimation

Narendra Ahuja, *Fellow, IEEE,* and A. Lynn Abbott, *Member, IEEE*

Abstract—**Much research has emphasized stereo disparity as a source of depth information. To a lesser extent, camera vergence and lens focus have also been investigated for their utility in depth recovery. Each of these visual cues exhibits shortcomings when used individually in the sense that none alone can be used to reconstruct surfaces for real scenes that often cover a wide field of view and a large range of depth. This paper presents an approach to *integration* of these cues that attempts to exploit their complementary strengths and weaknesses through active control of camera focus and orientations. In addition, the aperture and zoom settings of the cameras are controlled. The result is an *active* vision system that dynamically and cooperatively interleaves image acquisition with surface estimation. A dense composite map of a single contiguous surface is synthesized by automatically scanning the surface and combining estimates of adjacent, local surface patches. This problem is formulated as one of minimizing a pair of objective functions. The first such function is concerned with the selection of a target for fixation. The second objective function guides the surface estimation process in the vicinity of the fixation point. Calibration parameters of the cameras are treated as variables during optimization, thus making camera calibration an integral, flexible component of surface estimation. An implementation of this method is described, and a performance evaluation of the system is presented. An average absolute error of less than 0.15% in estimated depth was achieved for a large surface having a depth of approximately 2 m.**

Index Terms—**Active vision, camera calibration, fixation, range from focus, range from stereo, range from vergence, surface estimation, visual cue integration, visual target selection.**

I. Introduction

An arbitrary point in a 3-D scene projects onto different locations in stereo images. When the imaging geometry is known, the disparity between these two locations provides an estimate of the corresponding 3-D position. Many algorithms have been developed for estimating surfaces from stereo images of a scene. Most of these algorithms assume that the images are acquired from known viewpoints with compatible camera orientations and, of course, with the area of interest in proper focus. However, for real scenes that are deep and wide, no single imaging configuration can obtain stereo images of the entire scene suitable for surface reconstruction. This is because the cameras capture visual fields of limited size and depth. To reconstruct the surface for an entire scene, the camera configuration must be varied to sequentially capture different parts of the scene. Like human eyes, the cameras must pan and tilt, converge and diverge, and focus on near or far objects. For overall surface reconstruction, the surface parts estimated from different configurations must be merged. This implies that stereo-based surface reconstruction and the control of imaging must take place in a cooperative mode. The two processes must be interleaved in time. The surface reconstructed for a given part of the visual field must be added to the cumulative surface data, which in turn must be used to predict the part of the unmapped surface that will be reconstructed next.

The above observations are in agreement with the tenets of the active, intelligent data acquisition approach Bajcsy has emphasized [1], [2]. Even though there have been a number of computational studies on such active approaches, there has been only limited use made of the different cues in developing detailed computational approaches and implementations for surface estimation from stereo images, especially in a mutually cooperative mode such as that discussed in [1] and [3]–[5].

This paper describes an approach for active surface reconstruction that integrates the use of stereo with the control of camera focusing and vergence. Thus, image data acquisition is integrated with surface estimation. Our implementation of this method produces dense depth maps for scenes that are deep in extent and are wider than the field of view of the cameras. The system autonomously selects new locations for fixation to smoothly extend the evolving surface map.

A pragmatic side effect of such active control of cameras is a continuous degradation of calibration due to inaccuracies in the mechanical control system. The method described here automatically compensates for this by integrating system calibration with the surface reconstruction, thus resulting in adaptive self-calibration. This also obviates the need for frequent calibration processes requiring special calibration patterns, which are often very elaborate and therefore infeasible while processing real scenes. Effectively, in such adaptive calibration, the role of the external calibration patterns is fulfilled by the partially reconstructed surface map of the scene.

Manuscript received October 29, 1990; revised October 31, 1991. This work was supported by the National Science Foundation under grant IRI-89-11942, Army Research Office under grant DAAL 03-87-K-0006, and the Rockwell Corporation. Recommended for acceptance by Associate Editor E. Grimson.

N. Ahuja is with the Beckman Instiute, Coordinated Science Laboratory and Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801.

A. L. Abbott is with the the Bradley Department of Electrical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.

IEEE Log Number 9211199.

0162–8828/93$03.00 © 1993 IEEE

The integration task in our approach is formulated as one of optimization. In Section II, we discuss the stereo, vergence and focus processes individually as sources of 3-D information. Section III argues that these sources have complementary performance characteristics, and they should be used in an integrated mode. A control structure for such integrated operation is also given, which consists of two steps: target selection and surface reconstruction. Some previous work on each of these steps, both in physiological and computational contexts, is reviewed in Sections IV and V. Section VI presents the basic computational approach that we have developed for achieving the desired integration. Section VII describes a specific integration algorithm that we have developed following the approach of Section VI and its implementation on an active camera system that we have in our laboratory. Experimental results for two physical objects are presented. For the first object (a barrel), a performance analysis that confirms a high level of accuracy for this method is presented. Over eight fixations, an average error of less than 0.15% was observed for an object distance of approximately 2 m. For the second surface (a chair), the camera system completed 36 fixations in an autonomous mode and produced a surface map that represents about a third of the chair. Section VIII presents a summary.

The approach presented in this paper is for a scene that contains a single, smooth (continuous-depth) surface, e.g., a single object. When there are sharp depth discontinuities present in the visual field (e.g., due to self-occlusion or multiple objects), the integration paradigm assumes greater complexity. In particular, when the entire surface of a fixated object has been scanned, a new target point, which is visible from both cameras, must be selected on a hitherto unmapped object. To select a target appropriately requires knowledge of the structure of the unmapped objects. However, estimation of the surface map of new objects is the original objective. This leads to a circular dependence problem wherein the structure of the unexplored scene is required before the same is acquired. This general case involving multiple objects is beyond the scope of the present paper. An approach to this general case that integrates coarse-to-fine image acquisition with coarse-to-fine target selection and surface reconstruction for unmapped scenes is introduced in [6]. Our work thus far on surface reconstruction from active stereo is in two distinct parts: multicue integration for a single object reconstruction as described in this paper and integration of image acquisition with surface reconstruction as described in [6]. Both of these active stereo algorithms use an extension of the passive stereo algorithm reported in [7] and [8] for surface reconstruction from stereo image pairs for each target area.

The following, then, are some salient characteristics of the approach presented in this paper:

1) It is capable of autonomously scanning and reconstructing continuous surfaces having an arbitrary depth range. The capability to scan autonomously incorporates dynamic control of the values of camera vergence angles, focus settings, aperture settings, and zoom settings. Except for zoom, for which only two settings (high and low) are used, the rest of the integrated parameters take on continuous values.
2) It is capable of working with standard video cameras and with the calibration errors associated with camera reconfiguration. External calibration is done only once before the reconstruction begins; thereafter, the system performs self-calibration by using the partially reconstructed surface as a reference. The calibration process is an integral part of surface reconstruction, just as is the control of vergence, focus, aperture, and zoom settings.
3) It integrates the processes of image acquisition and surface reconstruction, thus making the system active. This is true because imaging parameters are selected optimally for surface reconstruction on the basis of observed image data, and the resulting reconstructed surface is used to select new imaging parameters so that the cameras continue to scan the scene.
4) It delivers a dense surface map as output instead of 3-D locations for isolated feature points. As described earlier, a performance analysis has shown a high degree of accuracy for the resulting map.

II. Stereo Disparity, Vergence, and Focus as Depth Cues

Stereo disparity and camera vergence have long been recognized as important binocular sources of 3-D information. Changes in focus directly contribute to the degree of image blur and may serve as a monocular cue to distance information. Since each plays an important role in the approach presented in this paper, we will now discuss each of these individually.

A. Depth from Stereo

Many algorithms have been developed for estimating surfaces from two stereo images of a scene acquired using a fixed, known camera configuration. The paradigm used by most early algorithms consists of three steps:

1) Detect suitable features in each image.
2) Find corresponding features in the two images.
3) Determine the 3-D locations associated with corresponding pairs of image features, and fit a surface to these 3-D points.

The features used are typically derived from intensity edges or from image regions. Edge-based algorithms attempt to match individual edge points or linear edge segments that consist of chains of aligned edge points. Most area-based approaches use as features image regions based on absolute intensity values and apply cross-correlation measures to evaluate the quality of the match between the regions. Most of these algorithms complete the matching process before surface interpolation is performed to obtain a dense depth map. Uniqueness of matching is enforced only by conditions that involve simple local relationships among disparity values, e.g., constancy of disparity.

In recent years, integration of the different steps of the stereo paradigm has been emphasized [7], [9]–[12]. The approach proposed in [12] integrates all three steps: feature detection, feature matching, and surface interpolation using an analog formulation. The stereo algorithm used in the work reported

in this paper is an extension of the Hoff and Ahuja algorithm discussed in [13], [8], and [14]. This algorithm integrates the last two steps of the stereo paradigm: those of feature matching and surface interpolation. Integration is performed using a model of the real world in which objects are viewed as having smooth surfaces in the sense that the normal direction varies slowly except across relatively rare creases and ridges. The surface characteristics are used to resolve matching ambiguities, and matching decisions are made so that the resulting surfaces are piecewise smooth. This is in contrast with previous approaches wherein matching is done without considering its impact on the quality of the resulting surface. This algorithm operates at a hierarchy of image resolutions, starting at a coarse level and proceeding toward levels of finer resolution. Each succeeding stage in the hierarchy uses the results of the previous, coarser stage to guide the search for correspondences.

Most stereo algorithms, including this one, require an externally specified, coarse, initial surface estimate that is refined using stereo analysis to obtain a more accurate surface description. Without such an estimate, an exhaustive search for correspondences would be required. For example, in the Hoff and Ahuja algorithm, a frontal surface having a depth halfway between the nearest and the farthest points of the scene is externally provided as the initial estimate. The algorithm can recover the "true" surface map using stereo if the depth range of the scene is not too large.

Since this reconstruction method assumes that surfaces are piecewise smooth, some method is needed to quantify the degree of surface smoothness. The square of the quadratic variation

$$E_s = \int\int \left[\left(\frac{\partial^2 S}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 S}{\partial xy} \right)^2 + \left(\frac{\partial^2 S}{\partial y^2} \right)^2 \right] dxdy \quad (1)$$

is one measure of the smoothness of a surface S. It has been argued that the function S that minimizes E_s is most consistent with perception and is uniquely determined [15].

B. Depth from Vergence

When the optical axes of a stereo-camera pair intersect, the point of intersection is known as the *point of vergence*. From the knowledge of the distance separating the cameras (the *baseline*) and the rotation angles, it is possible to determine the vergence angle and the 3-D location of the vergence point.

When the point of vergence is known to lie on some surface in the scene, it is called the *point of fixation*, and it is then possible to use vergence information to estimate the location of the surface. To compute surface depth, therefore, it is necessary to verify that both cameras are in fact aimed at the same location. One approach to this problem is to determine the disparity present at the image centers and then reduce this disparity to zero by vergence movements. An efficient method is therefore needed to obtain a measurement of binocular disparity at a single point.

In one method for fast vergence control [16], the two stereo images are placed side by side and processed as a single image. If it is assumed that the second image is approximately a translated version of the first image, then Fourier-transform techniques can be used to locate the spatial "echo" represented by the second image.

Another approach to fast one-point disparity measurement is to use translational registration methods wherein one image is shifted with respect to another until the overlapping regions are most similar. The most common similarity measures for this minimize intensity differences or maximize cross-correlation over a window. The work reported in this paper uses a normalized cross-correlation measure for registration. This measure d^2 as a function of the translation (s,t) is given by

$$d^2(s,t) = \frac{\left[\int\int I_L(x,y) I_R(x+s,y+t) dxdy\right]^2}{\left[\int\int I_L^2(x,y) dxdy\right]\left[\int\int I_R^2(x+s,y+t) dxdy\right]} \quad (2)$$

where I_L and I_R are left and right image intensities.

These similarity measures may fail when the surface gradient is sufficiently large relative to the image planes. Other problems, such as insufficient detail, can also lead to incorrect registration. However, this possibility is ruled out since the whole approach of this paper is contingent on the availability of surface detail. Fixation may be impossible for another reason: occlusion of the desired fixation point from one or both viewpoints. This problem has received little attention. In general, it is not possible to fixate every scene point because an object may self-occlude its far side. An example of this is shown in Fig. 1. (Even though two objects are shown, the near object could well be connected to the more distant object, resulting in a single, self-occluding object.) The left optical axis is aimed at point $\boldsymbol{p}$ (Fig. 1(a)). If the right camera is also aimed at this point (Fig. 1(b)), point $\boldsymbol{q}$ obstructs the view. Point $\boldsymbol{p}$ therefore cannot be used as a point of fixation. For vergence information to be useful, the two cameras can either aim at point $\boldsymbol{q}$ (Fig. 1(c)) or try to fixate another distant point $\boldsymbol{r}$ (Fig. 1(d)). This paper refers to the active process of detecting and avoiding occlusions to achieve fixation as *exploratory fixation*. This will be addressed again in Section VII.

C. Depth from Focus

Changes in the focus setting of a lens result in a varying degree of blur in the image. By minimizing the blur, the surface depth can be estimated. Assume that an object at distance u from the lens center forms an image at distance v on the opposite side of the lens. These two distances are related by the lens law, which is formulated as

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f} \quad (3)$$

where f is the *focal length* of the lens. The focusing mechanism for a lens changes the distance separating the lens center and the image plane of the camera. If an image point is in focus, and if v and f are known, then from the lens equation, it is possible in principle to determine the distance to the object. When a scene point does not satisfy the lens equation, the image is blurred. The effect of defocusing can be modeled, to a first approximation, by the convolution of the image with a low-pass filter. This causes the loss of high spatial frequency components in the image. Measures of high-frequency content

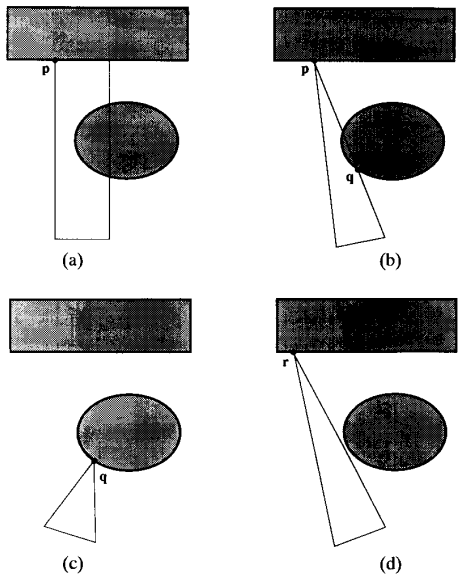


Fig. 1. Demonstration of occlusion using top views of verging cameras: (a) Initially, scene point $\boldsymbol{p}$ projects onto the center of the left image; (b) attempt to fixate $\boldsymbol{p}$ brings the image of point $\boldsymbol{q}$ on the circular object into the center of the right image. Because of this occlusion, the system could now attempt to fixate either (c) the near point $\boldsymbol{q}$ or (d) a point $\boldsymbol{r}$ on the distant surface.

can therefore be used to develop an objective function that assumes its optimum value when the image blur is minimized, and the image is therefore in sharpest focus.

Several autofocus methods have been proposed in the past. Horn [17] describes a Fourier-transform method in which the normalized high-frequency energy from a 1-D FFT is used as the objective criterion. Sperling suggests the squared Laplacian as a measure of image blur [3]. Tenenbaum uses a thresholded gradient magnitude method in which Sobel operators are used to estimate the gradient [18]. The criterion function used is the sum, over some image window, of the gradient energy that exceeds a certain threshold. This has also been used by Krotkov [19]. Jarvis suggests sharpness measures based on entropy, variance, and gradient [20]. A survey and comparison of several criterion functions for focus is presented in [21]. The criterion functions described there make use of such measures as signal power, gray-level standard deviation, thresholded pixel counts, and summation of squared gradient in 1-D. Most focus-ranging methods search the set of possible focus settings to minimize the image blur and then use the lens equation to calculate the range to the imaged surface. The algorithms presented in this paper use a gradient method similar to that of Tenenbaum, based on the energy of the brightness gradient in the images:

$$E_f = \int\int \left[\|\nabla I_L\|^2 + \|\nabla I_R\|^2\right] dxdy. \tag{4}$$

In the ideal, noiseless case, this objective function is unimodal.

Focus is an attractive source for depth information since it is monocular, having no analog to the correspondence problem of stereopsis,[1] and because it is relatively simple to use. However, any physical imaging system has limitations in resolving details. In particular, it may not be possible to detect small changes in image sharpness. This gives rise to the phenomena of *depth of focus* and *depth of field*. The image of a point may blur into a circle of diameter C before a loss in sharpness is detected. The circle having this diameter is known as the *circle of confusion*. For an object at a distance u_0, the image is theoretically formed at v_0, as determined by the lens equation. Because of the limited ability of the system to discern image features, however, the image plane can move through a range of locations about the distance v_0, and the image will appear equally sharp. This range is the depth of focus. Conversely, for a given location v_0 of the image plane, objects within the range of depths $[u_1, u_2]$ will appear equally sharp in the image. This interval is the depth of field. Object points at the two extremes of this interval form perfectly focused images at image locations v_1 and v_2, and both objects at these two extremes form circles of diameter C on the image plane at v_0. The size of the depth of field depends on the aperture A, the focal length f, and the image plane location v_0 and is given by

$$u_2 - u_1 = \frac{2ACu_0f(u_0 - f)}{A^2f^2 - C^2(u_0 - f)^2}. \tag{5}$$

In this equation, u_0 is determined by the focus distance v_0. The effect of depth of field is to provide an upper bound on the accuracy that is possible for depth estimates from focus. To see this, observe from (5) that the depth of field becomes infinite when

$$u_0 = f\left(\frac{A}{C} + 1\right). \tag{6}$$

III. The Need for Integration

Each process described in the previous section can provide an estimate of scene depth independently. This section discusses the benefits of cooperation among these processes, and the need for camera movements in surface reconstruction for real scenes. Thus, this section presents the basic motivation for the approach to active surface estimation presented in this paper.

A. Strengths and Limitations of Individual Cues

If fundamental limitations of one depth cue can be offset by depth information from a different cue, then the two cues can be used as *complementary* sources of surface information. This section takes a comparative look at each cue in this regard. The criteria for comparing their performances include required image characteristics, capability, complexity, and accuracy.

1) Required Image Characteristics: Point feature-based (e.g., edge-based) stereo methods require localized features or high-frequency detail in the images. This occurs only when scene surfaces are properly focused and implies that such

[1] In fact, changes in the focus setting result in a magnification change in the image, and therefore, correspondences need to be determined. This magnification effect may be neglected if sharpness is measured within a small window near the image center.

stereo methods benefit from larger depths of field since this will cause sharp image features over larger ranges of depth. In contrast, range estimates from focus are more accurate when the depth of field is shallow. Further, because stereo and vergence both derive depth information from binocular correspondences, they can be misled by spatial periodicities in the two images. Since focus is monocular, no problem exists in this regard.

2) Capability and Complexity: Camera vergence movements serve as a gross initial attempt to reduce the stereo disparity, thus reducing the size of the search space for finding correspondences. To achieve suitable camera movements requires an initial estimate of the location of the scene point that is to be the center of both stereo images. Focus methods can also benefit from an initial surface estimate since, given the estimate, the search for the final focus setting could be made efficient. Further, neither stereo nor vergence can derive depth estimates for such parts of a surface that are not visible to both cameras. However, monocular cues such as focus can yield the depth information whenever the surface is visible from at least one viewpoint. Both focus and vergence are "line-of-sight" methods and provide depth estimates for single points in the scene. Stereo processing, on the other hand, yields depth estimates for many image points simultaneously.

3) Accuracy: Focus methods are fundamentally limited by the depth of focus for the lens, which typically depends on spatial sensor-array quantization. Stereo accuracy also depends on the sensor quantization. The accuracy for stereo and vergence greatly depends on accurate knowledge of imaging parameters, particularly relative camera position, but can provide depth data with subpixel accuracy. The theoretical accuracy for each cue decreases with increasing object distance but at different rates.

B. The Role of Integration

From the above discussion, we see that each of the different cues may yield an estimate of scene depth independently. However, they have different requirements and performances. In this section, we will compare these cues with respect to the above three characteristics. We will argue that the strengths and weaknesses of the cues are indeed complementary and that it is possible to use them together such that in any given situation, the most useful cues are automatically selected and used to provide only such scene information as they can most reliably extract. Thus, we will discuss the benefits of cooperation among the surface estimation processes associated with different cues and the need for camera movements during surface reconstruction for real surfaces. To highlight their interrelationships, we first summarize the input/output characteristics of the cues, which follow from their description in the previous section.

1) Stereo: Surface estimation from stereo requires the following:

i) Visibility of the surface part of interest from both cameras, or alternatively, limited range of stereo disparity
ii) knowledge of the camera positions and orientations
iii) in-focus images
iv) initial, coarse surface estimate for efficiency.

If the surface has a large depth range, it may not suffice to have, for example, a constant-depth surface as a coarse surface estimate. Stereo delivers a surface estimate over a large region, whose accuracy depends on, among other factors, the sharpness of image features.

2) Focus: Surface estimation from focus requires i) visibility, ii) a (coarse) estimate of depth for computational efficiency, and iii) a narrow depth of field for accuracy. Focus delivers i) sharp images and ii) a depth estimate at the point of interest. The accuracy of the depth estimate decreases with increasing object depth.

3) Vergence: The capability to verge requires i) an estimate of the location of the point of vergence and ii) visibility of the point of vergence from both cameras. Vergence delivers i) a depth estimate for the point of vergence and ii) a reduced range of stereo disparity in the images. The depth estimate provided is quite accurate for large vergence angles, but the accuracy decreases as object depth increases.

To reconstruct surfaces for real scenes, none of the above cues is sufficient by itself since either its requirements are not met, or it does not give sufficiently accurate reconstruction. However, a closer look reveals that the requirements and deliverables of the different cues have interesting correspondences. The input of one is often the same as the output of another (see Fig. 2). Thus, if the cues can be bootstrapped as a system, accurate surface estimation is possible due to their combined strength. This is the central idea of the proposed integration approach. To illustrate, observe that stereo disparity can provide accurate surface reconstruction, but it requires a coarse initial surface estimate; such an estimate can be provided by focus and/or vergence although they may themselves not be as accurate as stereo (e.g., for large distances). Further, depth estimation of a scene point from vergence is valid only if it is ensured that both cameras are actually fixated at that point. This is not a serious problem for distant objects since occlusion then becomes insignificant. However, for relatively close objects, fixation needs to be ensured or verified (e.g., by ensuring that the depth estimates for the image centers are for the same 3-D point.) On a more practical note, the vergence process is more difficult for nearer objects because of the greater likelihood for occlusion and because of increased perspective distortion in the images. However, the near field is precisely where focus methods are most accurate. Therefore, when used cooperatively, focus can be used to guide vergence movements.

Another example of the mutual interdependence of the cues results from the visibility requirement that all cues have. Thus, an estimate can be obtained for only a limited part of the complete surface of interest. The part that can be imaged and analyzed is limited both along the lateral dimensions and along the depth dimension. The former limitation occurs because of the limited field of view of the cameras and may be remedied through vergence control by changing the orientations of the two cameras so that their optical axes intersect at different lateral locations on the surface. The latter limitation arises for two reasons. First, the entire surface may not be in focus simultaneously over its large depth range. Second, the entire

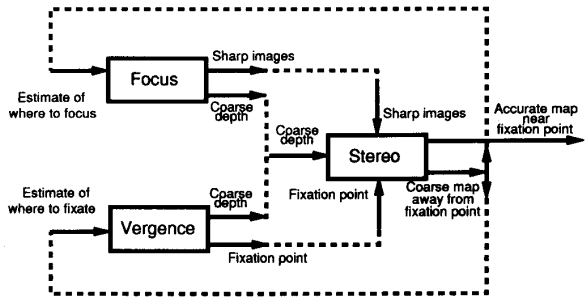


Fig. 2. Complementary nature of the focus, vergence, and stereo cues. Here, the prerequisites for a cue are shown by incoming arrows, and the estimates the cue can produce are shown by outgoing arrows. With integration such as that described in this paper, the cues cooperate to meet each other's prerequisites, as shown by the interconnecting (dotted) lines. Integration thus combines the strength of the cues.

surface may not give disparity values in a workable range; for example, the parts of the surface that are much closer than the point of fixation may give disparity values on the order of image dimensions, whereas those parts that are much farther may give disparities that are too small (less than a pixel) for depth recovery. This problem may be remedied by fixating at different parts of the scene through vergence control and obtaining depth estimates of the surfaces a small depth range at a time. These local surface patches can be imaged by changing the vergence angles of the cameras so that the point of fixation moves along the depth dimension while simultaneously adjusting focus to obtain sharp images.

Another aspect of the complementary nature of the cues involves the optimal imaging conditions for each cue. For example, stereo provides more information when the images represent wide fields of view; in contrast, focus and vergence work best with narrow fields of view so that these line-of-sight mechanisms are less likely to encounter large depth ranges. When zoom lenses are available, one solution is first to invoke the focus and vergence mechanisms, using the narrow field of view of a lens at full zoom, and then to obtain stereo-based depth separately from images taken at a smaller focal length. This method has added advantages in terms of depth of field; the accuracy of focus is best when the depth of field is low, and this is true when larger focal lengths are used; stereo ranging is most accurate with large depths of field, which is the case for smaller focal lengths.

The above argument suggests a surface-reconstruction scheme in which camera vergence, focus, zoom, and stereo are controlled in a coherent and integrated way, and the surface reconstruction takes place incrementally (small parts at a time). The focus and vergence processes serve the stereo cue in the sense that their goal is to fulfill the requirements of stereo, which then yields the final surface map. Stereo is assumed to be the most accurate cue (although it is assisted with initial estimates from the other cues). Stereo-based depth estimates supercede estimates of depth from other cues. The surface map that results can be used to guide focus and vergence control for subsequent fixations. Further, the surface map can be used to infer occluded parts, and the focus/vergence processes can be directed to avoid these parts.

The discussion of this section (and this paper) assumes the presence of a continuous-depth surface in the scene with no sharp depth discontinuities. When the visual field contains multiple objects or surfaces, the integration paradigm assumes greater complexity. For example, at the stage when the entire surface of a fixated object has been scanned, and thus, the acquired surface map does not smoothly extend, surface reconstruction must be resumed by fixating on a new object. However, 3-D information about a new object is unavailable by definition. (Otherwise, why would we want to fixate?) An approach to treat this general case is introduced in [22] and [6].

C. A Control Structure for Integration

The analysis of the previous section leads to a paradigm for active surface reconstruction, which can be represented by the following repeating pair of operations: *visual target selection* and *surface estimation* in the target area. This paradigm is the basic theme of the of the work presented in this paper. In the first step, a 3-D scene target is chosen to attempt fixation. The target is chosen based on the current global surface map. After selection of a target, the cameras are aimed at the target, and local surface estimation is performed in the vicinity of the fixation point. The newly obtained surface is added to the accumulating composite surface map before iterating back to the first step.

IV. Background—Visual Target Selection

Every active-vision mechanism must be able to decide *where to look*. This is the problem of *target selection*, which is addressed in this section. After choosing areas of interest in the scene, gaze directions can be directed toward the chosen target. The target selection criteria determine how the scene is scanned as a surface map is accumulated. These criteria are used to select a point in the visual field where surface information should be acquired next. This is relevant to eye movements in human vision for which slow eye movements interspersed by frequent jumps (saccades) characterize the continuous search for target points. Before we devise computational criteria for this purpose (Section VI), it will be useful to review some facts concerning human eye movements and consider other computational research in this domain.

A. Psychological and Physiological Studies

Human eye movements occur so that the image of a scene area of interest falls on the fovea, where retinal resolution is highest. The selection of a point for fixation is a complex and highly *goal*-dependent process, which usually takes place below the level of conscious thought. The psychological literature contains a large number of eye-movement studies. The purpose of these studies is typically to infer properties of higher level cognitive activity that govern the movements. Very few studies deal with 3-D domains, and these are often concerned with ergonomics or vehicle operation. Several relevant studies are reported in [23]–[25].

Our interest in the psychological studies is to learn about target selection criteria used by humans and to evaluate their relevance to a computational formulation. The following are

some significant findings from various 2-D psychological studies in which subjects exhibited strong tendencies during the selection of new fixation points:

a) Sequences of visual targets are often selected in a centrifugal order, beginning at the departure point (initial fixation point) [26]. This implies that proximity to the original point of fixation is an important criterion.
b) Upward eye movement is preferred over downward movement [26].
c) Eye rotation either to the left or the right is preferred, depending on the person [26]. This tendency was seen in children under of 5 years of age, implying that these preferences are not acquired through reading habits.
d) For several potential targets in the visual field, those lying closer to the fovea are more likely to be selected for fixation [27]. This effect may depend partially on the change in resolution from fovea to periphery.
e) When scanning random 2-D polygonal forms, eye fixations tend to concentrate near vertices [28].
f) During examination of pictures, saccades are directed to peripheral areas of "informative detail," [29] which involves higher level recognition of image objects (e.g., features of human faces).
g) When symmetry is present in 2-D displays, subjects tend to concentrate fixations along the axes of symmetry [30].
h) When peripheral stimuli are presented suddenly, the resulting strong temporal cue often leads to a saccadic eye movement toward the target [27].

As we will see in Section VI, the computational criteria we have used in our work are consistent with several of the above biological criteria, although our adaptation of them is motivated by the purely computational advantages they offer.

B. Computational Studies

We now summarize some computational studies that involve target selection. These studies incorporate a range of target selection criteria. Our goal in the approach presented in this paper is to concentrate only on surface reconstruction without any higher level objectives such as recognition.

Koch and Ullman describe a mechanism for the selection of visual targets based on an abstract measure of *saliency* [31]. Clark and Ferrier describe an implementation of a two-level control system for the control of tilt, pan, and coupled vergence movements [32]. Krotkov describes the use of a computer-controlled camera system that extracts a sparse set of line segments as stereo features [4]. This system aims the cameras at locations predicted by candidate line matches. Ballard and Ozcandarli discuss the use of eye movements within the context of object recognition [5]. Coombs and Brown discuss the control issues of gaze stabilization for an active camera system [33]. Burt has considered hierarchical approaches to the selection of targets for attention [34]. Shmuel and Werman have proposed a mathematical model that uses iterative Kalman-filtering techniques to predict a new camera pose for optimal reduction of uncertainty of an evolving depth map [35].

V. Background—Cooperative Surface Estimation

The previous section discussed identification of a location of interest in the scene, followed by direction of gaze towards the location. After this *fixation* process, the point of interest will project onto the centers of the stereo images. The presented approach is called "active" because of such scanning of the scene. Once fixation is achieved, the resulting stereo pair of images is used for surface reconstruction in the vicinity of the point of fixation. As in the previous section, it will be useful to first review some aspects of biological vision before presenting (in Section VI) the computational model for integrated surface reconstruction used in our work. Some of the physiological studies have significant computational ramifications.

A. Physiological Studies

Much work in biological vision has been concerned with modeling eye vergence movements in response to changes in the visual field. Other physiological research concerns the interactions that exist among different visual cues. These latter interactions are of particular relevance for the computational approach presented in this paper.

Biological vergence movements are traditionally decomposed into four components according to the cause or goal of the movements [36]. The following two components are of interest in this paper. The first is *disparity* (or *fusional*) vergence, which tends to reduce binocular disparities at the center of the retina. This is probably the dominant component of vergence, without which stereoscopic fusion cannot take place. *Accommodative* vergence is the second component and is assumed to result from image blur.[2] This represents a link between the accommodation and vergence mechanisms. Analogs of these components have been incorporated into the computational model described in Section VI. Foley has considered vergence as an independent depth cue [37]. In the absence of other depth cues, his findings indicate that perceived depth increases linearly with vergence angle, which yields reliable perception of relative depth, but the perception of absolute depth is inaccurate. Krishnan and Stark present a system model of the disparity-vergence component, which accepts a disparity signal as input and produces vergence control signals as output [38]. Similarly, Hung and Semmlow describe an analytical model that integrates the accommodation and disparity-vergence subsystems [39]. Another quantitative model is presented by Schor [40]. In this model, image blur and disparity both serve as stimuli that drive the accommodation and vergence control signals. The goal in each case is to model the physical dynamics for physiological vergence. The means by which the control signals might be derived from a stereo pair of images is not discussed.

Sperling presents a model that characterizes the interactions of accommodation, vergence, and binocular fusion for a biological system [3]. His is an "energy" model in which each of these three visual information sources contributes a separate energy component based directly on the visual input. Two vergence components (disparity vergence and accommodative vergence) are incorporated into the model.

[2] *Accommodation* is the physiological term for changes in focus for the eye.

This model differs from others in that it discusses methods for deriving the control signals from images and considers binocular fusion as a separate visual cue. It is formulated such that three independent state variables representing accommodation a, vergence v, and fusion u are permitted to vary so that an objective function is minimized. The overall objective is to minimize three separate energy measures e_a, e_v, and e_u, which are defined as follows:

$$\begin{aligned} e_a &= g_a(a-v) - \int\int w_a\left[|\nabla^2 I_L|^2 + |\nabla^2 I_R|^2\right] dxdy \\ e_v &= g_v(v-a) + \int\int w_v |I_L - I_R| dxdy \\ e_u &= g_u(u-v) + \int\int w_u |I_L - I_R| dxdy. \end{aligned} \tag{7}$$

Sperling refers to the first term g_i in each of these three equations as an "internal energy" component and to the second term h_i as an "external energy" component. Each function g_i represents a penalty for disagreement between two different distance cues and is formulated as a convex-upward function of its argument. The functions h_i in the equations represent the effects of the retinal images to induce change in the state variables. The units of a and v are given as diopters and degrees, respectively. The functions $w_i(x, y)$ are weights that emphasize particular retinal locations. The first energy component e_a balances the need to reduce image blur with the need for a and v to be in agreement. The second energy component e_v provides a measure of the spatial similarity of the two images, which is similar to (2), and reflects the "quality" of the vergence angle. The external energy term is expected to be a minimum when both image centers are in registration, which should occur when the eyes are aimed at the same scene location. Sperling defines the last, or fusion, component e_u identically to that of vergence, with a single difference: The summation is performed only over foveal areas having disparities small enough that fusion can take place. These disparities correspond to objects lying within Panum's fusional area. The variables a, v, and u can be taken to represent the state of the fixation system. The intent is that these state variables are permitted to vary smoothly until a minimum energy state is found.

Most of the physiological models discussed above are clearly not intended for surface estimation. Indeed, these models consider only point or area operations rather than emerging surface characteristics. As a result, many issues relevant to surface reconstruction are unaddressed or unresolved. For example, the question of occlusion from one eye is never raised. The model of Sperling is the most comprehensive known to us since it incorporates the stereo fusion phenomenon and discusses the derivation of control signals directly from the retinal images. Although the model is presented in analytic form, it is not intended as a computational paradigm. Nonetheless, the modeling of interaction among different cues is extremely pertinent for the integration of information derived from these cues. The model of Sperling has partly motivated the computational model reported in this paper.

B. Computational Studies

Tenenbaum was an early advocate of a computational active-vision approach [18]. He presents methods for a vision system to "accommodate" the environment by changing imaging parameters, based on an analysis of information obtained from the sensors themselves. Bajcsy has strongly advocated the approach of active vision; she argues that feedback from partial visual processing should be utilized to guide the selection of new imaging parameters [1], [2]. Krotkov describes the use of a computer-controlled camera system that integrates focus, vergence, and stereo for range estimation [4]. The approach is cooperative in that focus and vergence are used initially to obtain estimates for stereo matching and then to verify a sparse set of matches. In their analysis of surface reconstruction from stereo images, Marr and Poggio point out the role of eye movements in providing large relative image shifts for matching stereo images having large disparities, thus implying the need for active data acquisition [41], [42]. Ballard and Ozcandarli argue that the incorporation of eye movements radically changes (simplifies) many vision computations; for example, the computation of depth near the point of fixation becomes much easier using object-centered reference frames [5]. Aloimonos *et al.* consider the analytical implications of active vision methods [43]. They show that mathematically ill-posed, nonlinear, or unstable problems for a passive observer can become well-posed, linear, or stable under active observation. Bandopadhay *et al.* consider tracking by a moving observer in a static environment as a means to reduce the complexity of computing the motion parameters [44]. Geiger and Yuille describe a method for using small vergence changes to help disambiguate stereo correspondences [45].

An important consequence of integration of depth cues via active vision is the need to fuse depth maps obtained using different viewpoints and imaging configurations. This also requires camera calibration. Some such algorithms for fusion of depth maps and camera calibration are now mentioned.

Ferrie and Levine describe a hierarchical approach to feature matching across images obtained from several viewpoints [46]. Kamgar-Parsi *et al.* present a method for registration of overlapping range images within the context of terrain mapping [47]. Ayache and Faugeras describe a method for registering and fusing depth maps obtained using passive stereo [48]. As stereo image pairs are extracted for new camera poses, depth and uncertainty information is refined recursively through sequential optimization using extended Kalman filtering. Takahashi and Tomita present a method for self-calibration of stereo camera orientations [49]. The method assumes that the cameras are initially calibrated, but with time, the orientations of the cameras may differ from the calibrated angles. The difference between the observed feature locations and those expected from calibration is used to guide the calibration of the relative camera orientations.

The assumption that the locations of measured depth points (or feature locations in the images) are perturbed with Gaussian noise is made fairly often for mathematical tractability. It facilitates recursive approaches, such as Kalman filtering, since data acquisition is assumed to be serial. The utility assumption

of Gaussian noise is useful only when it is known that the correspondences are correct. However, the most serious errors in stereo surface reconstruction arise from feature mismatches across stereo images. Therefore, in the work reported in this paper, the individual 3-D points obtained for each viewpoint are replaced by a surface through them. The surfaces for different viewpoints are then used for registration, calibration, and fusion across viewpoints. When two estimated surfaces overlap, for example, they are assumed to belong to the same object. An error term is computed from the differences in depths between the estimated surfaces, and this is used to guide corrective mechanisms. In addition, since the method described here deals with dense depth maps that contain large numbers of surface points, feature points are not retained across separate fixations because of the high computational cost that would result.

VI. A Computational Approach to Integrated Active Surface Reconstruction

At the end of Section III, a two-step control structure for surface reconstruction that meets the need for integration, as presented in that section, was described. This section describes the computational formulation for each of the two steps. Section VI-A is concerned with the first step, target selection. Section VI-B discusses the representations and notation needed in Section VI-C to describe the formulation for integrated surface reconstruction.

A. A Computational Model for Visual Target Selection

The psychological results summarized in Section IV-A (and referenced in brackets) suggest that the following factors are important in the selection of the next point for fixation:

1) *Absolute distance and direction* [a)–d) above]
2) *2-D image characteristics* [e)-g)]
3) *Temporal change* [h)]. Additional criteria may be identified based on purely computational considerations.
4) *Surface smoothness:* The selected point should smoothly extend the known surface unless the point lies beyond an object boundary.
5) *Occlusion regions:* The system should not attempt to fixate the parts of the visual field that are not visible from both cameras; thus, to maximize the rate of growth of the image area analyzed and the likelihood of correctly predicting occlusion regions, the scan should proceed from near to far.
6) *Compactness:* Successive fixation points should be selected to grow a surface outwards from an initial fixation point since most objects yield compact regions in the images.
7) *Complexity:* The total number of fixation points should be minimized. This minimizes the total camera movement, which in turn minimizes the time taken to map the entire scene. This is a significant factor since camera movement is a mechanical (and therefore slow) process.

The model for the approach presented in this paper incorporates only those criteria that involve surface geometry and does not take into account any criteria that require an analysis of image gray-level structure or involve any temporal changes. This model incorporates criteria 1), 4)–7), but excludes the more complex criteria 2) and 3), which are topics for further research. Thus, this current model stresses proximity: angular proximity of a potential target to the original fixation point $\boldsymbol{p}_0$ and to the current fixation point $\boldsymbol{p}_{POF}$ and distance of the target from the camera position $\boldsymbol{p}_{CAM}$. The target $\boldsymbol{p}$ is chosen so that the following weighted average is minimized:

$$E = a_1\|\boldsymbol{p} - \boldsymbol{p}_{CAM}\| + a_2 A(\boldsymbol{p}, \boldsymbol{p}_{POF}) + a_3 A(\boldsymbol{P}, \boldsymbol{p}_0). \quad (8)$$

The weights a_i balance the three terms of E. The value $A(\boldsymbol{p}_i, \boldsymbol{p}_j)$ represents the angular separation between two 3-D points relative to the point of projection for the current camera location. Candidate targets $\boldsymbol{p}$ are constrained to lie on the border of the composite surface map and must be within camera travel limits.

The first term of E favors scene points that lie near the imaging apparatus (criterion 5)). The second term biases the choice of target to scene points that lie near the current fixation point (criterion 1)). This tends to minimize short-term large camera movements. The third term ensures that an evolving surface description will tend to develop outward ("centrifugally") from the point of departure (criterion 6)). Criterion 7) is met, in a simple way, by choosing the next fixation point on the border of the current surface map that uncovers as much as possible of the currently unknown part of the visual field, subject to the condition that fixation is not attempted outside of the current map. This latter condition ensures that a selected target point exists on the object surface. Further, meeting this condition guarantees overlap between the current surface map and the surface patch to be reconstructed next, which is required for self-calibration, as will be explained in Section VI-C. More sophisticated algorithms for choosing the next fixation point could be used, which could further reduce the overlap without sacrificing surface quality. It should be noted here that when a scene contains multiple objects, additional criteria to select the next object for fixation will need to be developed. However, this significantly more complex case is beyond the scope of this paper. (See [22] and [6] for an approach to this problem.)

B. Representation for Active Control of Imaging Parameters

The goal of this section is to present concepts and terminology for our formulation of integrated surface reconstruction described in Section VI-C. Our formulation views active surface reconstruction as the output of an active system whose dynamics capture the characteristics of camera movements. In general, a dynamic system may be defined by the specification of the following quantities:

1) *A description of the input to the system:* If some of the input values depend on the output of the system, the system is said to employ feedback. Other forms of input may include disturbances and external control signals.
2) *Prior knowledge:* This includes system parameters and fundamental assumptions.
3) *A system state specification:* The state of a system is a function of inputs to the system and of internal

system parameters. The specification often includes a specification of system dynamics, characterizing changes that occur in the system state as a function of the above quantities.

4) *A description of the output of the system:* This is specified as a function of the items listed above.
5) *A statement of the goal of the system:* This is often given as an error function to be minimized.

The integrated surface reconstruction in our formulation may be viewed as the output of a dynamic system that is a particular instance of the general model given above. The components of such a system are described in the following:

1) The primary input is *stereo images* I_L and I_R and a *shaft-encoder vector* $\hat{\boldsymbol{q}}$. The vector of shaft-encoder readings $\hat{\boldsymbol{q}} = [\hat{q}_0, \hat{q}_1, \ldots, \hat{q}_{M-1}]^T$ reflects the current state of physical effectors that control the imaging characteristics of the system. For example, an encoder, when attached to a dc motor that controls the focus setting for a lens, can provide the system with knowledge of the physical setting of the focus ring.
2) Prior knowledge for the system is represented by the vector of *initial system parameters* $\hat{\boldsymbol{\beta}}$. This represents prior knowledge provided to this system and includes calibrated system constants. For example, the mapping from a particular actuator setting to a physical angle may depend on several constants. These are contained in $\hat{\boldsymbol{\beta}}$.
3) The state of the system is contained in several system vectors and matrices. The most fundamental of these are the *actuator vector* $\boldsymbol{q}$, the *system parameters* $\boldsymbol{\beta}$, and the *surface maps* S and S_c. Each is now described separately.

   The actuator vector $\boldsymbol{q} = [q_0, q_1, \ldots, q_{M-1}]^T$ represents M degrees of freedom by which the processor controls the physical characteristics of the imaging system. Each element q_i is a particular "axis setting" to which a motor is driven corresponding to the input $\hat{q}_i$ described above.[3] The motors control both camera position and lens parameters. Limitations of the physical system determine constraints on the values that $\boldsymbol{q}$ can attain.

   The vector $\boldsymbol{\beta}$ represents current values of system parameters. These are updated values of the externally supplied constant vector $\hat{\boldsymbol{\beta}}$ and are used in the local optimization of camera-calibration parameters. Together, the vectors $\boldsymbol{\beta}$ and $\boldsymbol{q}$ are used to define the imaging model. Functions of these quantities map to such system parameters as focal length, vergence angle, and camera position. In particular, the camera projection and transformation matrices can be represented as functions of $\boldsymbol{q}$ and $\boldsymbol{\beta}$. Details of the mapping used in this work and of the initial calibration procedure are described in [50].

   The functions S and S_c are the local (for one viewpoint) and composite surface maps constructed by the system, respectively. Each is a map $z = f(x, y)$ of estimated surface points with respect to a reference coordinate system.
4) The outputs of this system are the surface maps S and S_C, and the actuator control vector $\boldsymbol{q}$ (described above). Clearly, the output of ultimate interest is S_C, with S and $\boldsymbol{q}$ being outputs of intermediate interest. $\boldsymbol{q}$ represents the result of Step 1, namely, target selection.
5) Finally, the goal of the system is to extremize each of a set of objective functions through the manipulation of the variables $\boldsymbol{q}, \boldsymbol{\beta}$, and S. In terms of the active system representation, the two steps of fixation and surface reconstruction can be described as follows.

To represent the sequential fixation process, assume that the system acquires a sequence of stereo image pairs $\{(I_L(t_i), I_R(t_i)) \mid i = 0, 1, 2, \cdots\}$. Each image depends on the state of the physical imaging system, which is given by system control values $\boldsymbol{q}(t_i)$ at time t_i. The values $\boldsymbol{q}(t_i)$ are used with system parameters $\boldsymbol{\beta}$ to derive a set of transformation matrices that can be used to estimate 3-D locations from stereo correspondences in the images $I_L(t_i)$ and $I_R(t_i)$. The system uses these images and transformation matrices to create a surface map $S(t_i)$, which is then merged with the previously obtained composite surface map $S_c(t_i - 1)$ to form the new composite map $S_c(t_i)$. The system then selects new actuator control values $\boldsymbol{q}(t_{i+1})$ based on properties of the images and the composite map, and the sequence continues.

To represent surface information obtained from different fixations, the map $S_c(t_i)$ incorporates the information obtained for fixations 0 through i. In the model for surface estimation developed here, it is assumed that local surface maps $S(t_i)$ are discarded after each fixation. Information about individual image features is not retained after they have been used to derive the composite surface representation $S_c(t_i)$. This means that $S_c(t_i)$ is constructed only from the maps $S(t_i)$ and $S_c(t_{i-1})$. This is motivated by computational and biological considerations. Computationally, since the stereo-reconstruction method described here utilizes a large number of scene features and since the number of fixations can grow without bound, the cost involved in maintaining these features separately for all fixations is considered to be too large. This approach may correspond to the physiological case for which it is assumed that transsaccadic fusion does not occur.

C. Integrated Surface Reconstruction

This section develops an analytical formulation that guides the surface estimation process. As is the central theme of the desired approach, this model allows the integration of information from several visual sources. The integration model is formulated in terms of an objective function to be minimized. The parameters varied to perform this minimization are of three types: actuator settings, which are represented by the vector $\boldsymbol{q}$, calibrated system parameters, which are represented by the vector $\boldsymbol{\beta}$, and the derived local surface S. The objective function is defined as a linear summation of several criterion terms or components described below. The role of each component is clear from its name, and the motivation for its use follows from the discussion in the previous sections.

[3] Because this work is not concerned with modeling the dynamics of the physical system, we assume that $\boldsymbol{q}$ always reflects the state of the physical actuators. It is implicit that lower-level control operates to bring the input shaft-encoder readings $\hat{\boldsymbol{q}}$ to the same values as the control settings $\boldsymbol{q}$.

1) Normalize Image Contrast: Clearly, the use of the different visual cues requires that sufficient image detail be present in the images. This corresponds to an adequate degree of image contrast, which is possible only when the lens aperture is set appropriately, matching the level of image irradiance to the sensitivities of the sensors. This can be done by controlling the aperture setting of each camera to bring the average image brightness to a predetermined level. Such consistency of the image contrast can be achieved by independently controlling the aperture of each camera so that the following criterion is minimized:

$$E_c = \mid E_{c0} - \int\int_{R^2} w_c I_L dxdy \mid + \mid E_{c0} - \int\int_{R^2} w_c I_R dxdy \mid . \quad (9)$$

The constant E_{c0} is the desired sum of brightness values for the image region, and the function $w_c(x, y)$ emphasizes the centers of the two images. Because the aperture setting for one camera does not affect the image of the opposite camera, the error terms can be combined into the single function E_c.

2) Minimize Image Blur: An error term E_f is defined; it tends to minimize the amount of blur in the image:

$$E_f = -\int\int_{R^2} w_f \left[\| \nabla I_L \|^2 + \| \nabla I_R \|^2 \right] dxdy. \quad (10)$$

When a camera is in sharp focus, the energy of the intensity gradient summed over the image will tend to be a maximum, and E_f will be at a minimum. The weight function $w_f(x, y)$ is used to emphasize the central region of each image. With this definition, the minimum of E_f is attained only when both cameras are in sharp focus.

3) Minimize Disparity at Image Centers: Cross-correlation measures may be used to quantify the degree of similarity between two image regions and may therefore be used to obtain a disparity measure for the image centers. When the following function E_v is minimized, the vergence angle should be optimal when both cameras are aimed at the same scene point. The term E_v uses a normalized cross-correlation measure similar to that of (2):

$$E_v = -\frac{\left[\int\int_{R^2} w_v I_L I_R dxdy\right]^2}{\left[\int\int_{R^2} w_v I_L^2 dxdy\right]\left[\int\int_{R^2} w_v I_R^2 dxdy\right]}. \quad (11)$$

The weight function w_v emphasizes the centers of the images.

4) Maximize Surface Smoothness: Scenes are assumed to be piecewise smooth. The following criterion function E_s uses the square of the quadratic variation in depth to measure surface nonsmoothness. Thus, minimization of E_s maximizes surface smoothness.

$$E_s = \int\int_{\Gamma} w_s \left[\left(\frac{\partial^2 S}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 S}{\partial xy}\right)^2 + \left(\frac{\partial^2 S}{\partial y^2}\right)^2 \right] dxdy. \quad (12)$$

The integration is performed over the domain Γ of S. The weight function w_s favors the center of the image over peripheral regions.

5) Minimize Differences in Depth Estimates Among Individual Depth Cues: The cues of focus, vergence and stereo provide four different estimates of the location of the point of fixation. Typically, when the system is properly fixated, the depth estimates for all visual cues should agree. The degree to which these depth estimates differ is added as a penalty E_p within the overall objective function.

Some metric is needed so that any two of these depth estimates can be compared. Although Euclidean distance might seem most reasonable, $f(z_1, z_2) = \mid z_1 - z_2 \mid$, further consideration suggests a different method of comparison. Since the accuracy of each cue decreases with distance, a better metric is one offering a greater penalty for near distances than for far distances for a given value of $(z_1 - z_2)$. The method chosen is a comparison of the reciprocals of depth using the criterion function $f(z_1, z_2) = \mid \frac{1}{z_1} - \frac{1}{z_2} \mid$. This is similar to the model of Sperling, in which depths are measured in diopters and by vergence angle, each of which is inversely related to depth. Each of these measures therefore tends to zero as distances increase without bound.

Notationally, assume that the function $\delta(\boldsymbol{p}_1, \boldsymbol{p}_2)$ maps the two 3-D points to the absolute difference of the reciprocal of the depths in a common coordinate frame. If the four-point estimates are denoted $\boldsymbol{p}_{Lf}, \boldsymbol{p}_{Rf}, \boldsymbol{p}_v$, and $\boldsymbol{p}_s$, for left focus, right focus, vergence, and stereo, then the agreement among them can be maximized by minimizing the following function:

$$E_p = \delta(\boldsymbol{p}_{Lf}, \boldsymbol{p}_{Rf}) + \delta(\boldsymbol{p}_{Lf}, \boldsymbol{p}_v) + \delta(\boldsymbol{p}_{Rf}, \boldsymbol{p}_v) + \delta(\boldsymbol{p}_s, \boldsymbol{p}_v). \quad (13)$$

Of the six comparisons possible for these four depth estimates, only four are used in the formulation of E_p. The first component is instrumental in detecting the presence of occlusions. The second two components are used to verify that the point of vergence agrees with the estimates from focus. The final term is used to compare depth from the stereo process with the point of vergence.

6) Maximize Agreement of Surface Estimates Across Fixations: The surface S obtained for the vicinity of the current fixation point should agree with any overlapping portions of the previously obtained composite surface map S_c. This agreement can be based on the differences of depths between surfaces as well as on the differences of surface-normal directions between the two surfaces.

The mechanical nature of the camera movement system and the associated lack of precise knowledge about the imaging parameters may result in poor agreement between S and S_c. In particular, there are several reasons to expect such lack of agreement:

1) The actual configuration of cameras is represented by a mathematical model, and no such imaging model will perfectly represent the imaging system.
2) A physical imaging system must undergo an initial calibration procedure, during which imaging parameters (represented by $\hat{\boldsymbol{\beta}}$) are determined. This initial system calibration is typically optimal in the least-squares sense, i.e., it will be most accurate, for the limited set of actuator-variable settings $\boldsymbol{q}$, which were used to perform the calibration but may not be correct for other settings.

3) The values of the imaging parameters $\hat{\boldsymbol{\beta}}$ indicated by the shaft encoders are unreliable due to such factors as component wear, component replacement, and varying lighting conditions. Thus, there may be disagreement between S and S_c. The system should be able to cope with such errors without having to perform a new, initial system calibration.

Not all of the above factors are equally serious. Typically, the worst errors will tend to result from incorrect values of system parameters. The other two factors yield much smaller errors than those caused by stereo mismatches, which is a major source of errors. In regions where stereo correspondences are assumed to be correct, the discrepancies in the depth values for the surfaces can be attributed to loss of calibration (errors in $\boldsymbol{\beta}$), which may differ by a small amount from the originally calibrated values $\hat{\boldsymbol{\beta}}$. These observed discrepancies may used to find improved estimates of system parameters $\boldsymbol{\beta}$ to yield the best possible fit of the newly derived surface patch to the previously obtained composite map.

Thus, a method that could be used for improving agreement between S and S_c is to permit small changes in the imaging parameters $\boldsymbol{\beta}$ during the process of surface reconstruction. This changes the surface map S. Let Γ and Γ_c denote the regions over which S and S_c, respectively, are defined; then, $\Gamma \cap \Gamma_c$ denotes their *overlap region*. Let A denote the area of this overlap region.

Using the area as a normalizing factor, the following term can be used as a measure of mean-squared depth difference per unit area for S and S_c:

$$E_{a1} = \frac{1}{A} \int \int_{\Gamma \cap \Gamma_c} w_{a1} |\, S - S_c \,|^2 dxdy. \tag{14}$$

It is also possible to utilize differences in surface orientation as a measure of disagreement. Let $\boldsymbol{n}$ and $\boldsymbol{n}_c$ denote the surface normals for S and S_c at some point. These are specified by the equations $\boldsymbol{n} = [-\frac{\partial S}{\partial x}, -\frac{\partial S}{\partial y}, 1]^T$ and $\boldsymbol{n}_c = [-\frac{\partial S_c}{\partial x}, -\frac{\partial S_c}{\partial y}, 1]^T$. Using this notation, a measure of disagreement based on the angle between the surface normals is

$$E_{a2} = -\frac{1}{A} \int \int_{\Gamma \cap \Gamma_c} w_{a2} |\cos^{-1}(\boldsymbol{n} \cdot \boldsymbol{n}_c)|^2 dxdy. \tag{15}$$

Together, the above two error terms can be used to express the disagreement between S and S_c and, thus, to select new values for $\boldsymbol{\beta}$. However, the estimates $\boldsymbol{\beta}$ may have associated tolerances, i.e., the errors in $\boldsymbol{\beta}$ may not be arbitrarily large. Therefore, large deviations in $\boldsymbol{\beta}$ should be appropriately penalized. This penalty can be based on the Euclidean distance between the two parameter vectors given by the expression

$$\| \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \|^2 = \sum_i | \beta_i - \hat{\beta}_i |^2. \tag{16}$$

Since it may be desirable to penalize changes in individual parameters differently, the following penalty term is more appropriate:

$$E_{a3} = \sum_i w_{\beta i} | \beta_i - \hat{\beta}_i |^2. \tag{17}$$

Then, the final objective-function component for self-calibration is

$$E_a = \frac{1}{A} \int \int_{\Gamma \cap \Gamma_c} w_{a1} |S - S_c|^2 dxdy - \frac{1}{A} \int \int_{\Gamma \cap \Gamma_c} w_{a2} |\cos^{-1}(\boldsymbol{n} \cdot \boldsymbol{n}_c)|^2 dxdy + \sum_i w_{\beta i} | \beta_i - \hat{\beta}_i |^2. \tag{18}$$

The overall self-calibration term is therefore based on the differences of depths between surfaces, on the differences of surface-normal directions between the two surfaces, and on a term that penalizes large departures in system parameters from originally calibrated values.

The relative weights for the components of E_a(w_{a1}, w_{a2}, and $w_{\beta i}$) determine the degree to which each contributes in the overall optimization. In particular, the weights $w_{\beta i}$ in the third term determine the willingness of the system to permit changes in the calibrated values for the imaging parameters $\hat{\boldsymbol{\beta}}$. Large values of $w_{\beta i}$ relative to w_{a1} and w_{a_2} indicate confidence in the original calibration, meaning that little change is permitted in the parameters, and thus, limited optimization over $\boldsymbol{\beta}$ is performed. A very small weight indicates the system's willingness to permit large changes in $\boldsymbol{\beta}$. It is required that the weights sum to unity.

The values of the weights $w_{\beta i}$ relative to w_{a1} and w_{a2} are determined by the actuator settings $\boldsymbol{q}$. When $\boldsymbol{q}$ is near a setting that was used during the original calibration procedure, there is high confidence in $\boldsymbol{\beta}$, and hence, $w_{\beta i}$ are large relative to w_{a1} and w_{a2}. The weights are varied according to the function

$$e^{-\frac{|\boldsymbol{x}|}{k}} \tag{19}$$

where $\boldsymbol{x}$ is the distance to the nearest actuator variable used during the original calibration process. The value of this function lies between 0 and 1 and can be interpreted as a measure of *confidence* in the initial system calibration. The constant k determines the rapidity with which the confidence in the originally calibrated values $\hat{\boldsymbol{\beta}}$ diminishes with the distance.

At this stage, it may be observed that enforcing the agreement of surface estimates does not involve another depth cue, but rather, it amounts to the integration of camera calibration with surface estimation.

7) Composite Criterion Function: Having described the individual criteria that should be optimized, we are ready to specify the composite objective function E for optimization. As stated earlier, a first definition of E is a weighted sum ($E = \lambda_c E_c + \lambda_f E_f + \lambda_v E_v + \lambda_s E_s + \lambda_p E_p + \lambda_a E_a$), which incorporates all of the individual objective functions. This composite objective function must be minimized by suitably choosing the values of $\boldsymbol{q}, \boldsymbol{\beta}$, and S. Thus, the optimization problem becomes

$$\min_{\boldsymbol{q},\boldsymbol{\beta},S} (\lambda_c E_c + \lambda_f E_f + \lambda_v E_v + \lambda_s E_s + \lambda_p E_p + \lambda_a E_a) \tag{20}$$

where weights λ_i are constrained to satisfy

$$\lambda_i \geq 0 \quad \text{and} \quad \sum_i \lambda_i = 1. \tag{21}$$

It is desirable to decouple control of the lens apertures from the rest of the system, so that the aperture settings depend only on the contrast function E_c. Unless this is done, the system will attempt to manipulate image brightness through the control of other actuators, such as focus or camera rotation. Changes in any actuator setting can indeed lead to variations in average image brightness, whereas it should be solely the task of aperture control to compensate for brightness changes when they occur. This rationale leads to the decomposition of the optimization problem in (20) into two optimization subproblems:

$$\min_{q_1,q_2} E_c \tag{22a}$$

$$\min_{\boldsymbol{q}^*,\boldsymbol{\beta},S} (\lambda_f E_f + \lambda_v E_v + \lambda_s E_s + \lambda_p E_p + \lambda_a E_a). \tag{22b}$$

These two minimizations are to be performed independently. The variables q_1 and q_2 represent the lens aperture controls. The symbol $\boldsymbol{q}^*$ represents the remaining actuator variables that control focus and camera orientation.

In the above formulation, it is important that an appropriate set of weights $\{\lambda_i\}$ is chosen. The selected values should ensure reasonable tradeoffs between the separate criteria when conflicts occur. If any single value λ_i is too low or too high, the corresponding cue will not be appropriately represented in the total cost. One possibility is that the weights are chosen to normalize the components so that each has equal representation when the overall function is optimized [51]. Alternately, the weights themselves may be determined dynamically by the nature of the images. This latter approach is being pursued in our ongoing work.

When the objective function is minimized for a given scene target, the result is the final selection of a point of fixation, the extraction of focused images, and the estimation of a local surface map derived from stereo analysis. The system must then obtain an enhanced composite surface description S_c by merging the local surface description S with the composite map.

VII. Algorithm, Implementation and Experimental Results

A. An Integration Algorithm

An algorithm is now presented for achieving the integration described in the previous section. Both components of the active surface-estimation paradigm (target selection and surface estimation) are described. The implementation of this algorithm (Section VII-B) demonstrates significant improvements in estimated surfaces over those possible without such interaction.

First, we will point out a matter of practical difficulty with the use of the optimization problem in (22); there is a high computational cost associated with the computation, during minimization, of those terms involving E_s because of the high cost of the stereo reconstruction process. Therefore, it is desirable to optimize with respect to variables that are not incorporated into this reconstruction process. This is shown in the following equations:

$$\min_{q_1,q_2} E_c \tag{23a}$$

$$\min_{\boldsymbol{\beta},S|\boldsymbol{q}^*} \left(\min_{\boldsymbol{q}^*} (\lambda_f E_f + \lambda_v E_v + \lambda_s E_s + \lambda_p E_p + \lambda_a E_a) \right). \tag{23b}$$

The second operation has been separated into two minimizations, where the inner minimization is to be performed first. Further modularization of the inner optimization is possible. To see this, recall that the goal of the terms in (23) involving only focus and vergence cues is to ensure sharp and registered images, and these terms alone can achieve that goal. Stereo is performed on the resulting images, and it does not involve any tradeoff with the actuator controls for focus and vergence. Thus, the stereo term acts as a constant during optimization over $\boldsymbol{q}^*$. The agreement term involving E_a is obviously a constant during optimization over $\boldsymbol{q}^*$. Therefore, the optimization problem in (23b) can be further decomposed:

$$\min_{q_1,q_2} E_c \tag{24a}$$

$$\min_{\boldsymbol{\beta},S|\boldsymbol{q}^*} , \left(\lambda_s E_s + \lambda_a E_a + \left(\min_{\boldsymbol{q}^*} (\lambda_f E_f + \lambda_v E_v + \lambda_p E_p) \right) \right). \tag{24b}$$

This is also an important decomposition because the optimization over $\{\boldsymbol{\beta}, S\}$ can be very expensive computationally. Indeed, this decomposition is responsible for separation of target fixation (not selection) from surface reconstruction around a target, thus greatly reducing the computational complexity.

The second component of this model (see (24b)) exhibits several similarities to the Sperling model. The term $\lambda_f E_f$, for focus, is similar to Sperling's accommodation measure, except that the gradient norm is used instead of the Laplacian as the criterion to minimize image blur. The term $\lambda_v E_v$ measures similarity of the two image centers and corresponds to Sperling's calculation for vergence, except that the normalized cross-correlation measure for registration is used here rather than Euclidean distance. The term $\lambda_s E_s$ computes the quality of stereo fusion in terms of the smoothness of the resulting surface, rather than in terms of intensity differences among corresponding pixels related by a fixed disparity value. The term $\lambda_p E_p$ penalizes disagreement among different depth cues and is analogous to the separate internal-energy components used by Sperling.

A difference from the Sperling model is the term $\lambda_a E_a$, which reflects the evolving nature of the composite map. After a local surface patch has been estimated, this is integrated with a composite surface description. The evolving, global surface map is a product of the aggregation of many such local patches.

The steps in the algorithm are outlined in the statement of the algorithm (which is given below) and is followed by a discussion of each of the steps.

Surface-Reconstruction Algorithm

Repeat until entire scene has been mapped

1. Select target

   1.1 Locate previous fixation point on composite map

1.2 Beginning at fixation point, search map for a nearby point on boundary of known depth region
1.3 Traverse this entire boundary, evaluating target-selection criterion at each location
1.4 Select boundary point having minimum criterion value as new target

2. Fixate
   2.1 Set lenses to full zoom
   2.2 Repeat until both image centers show the same scene area in focus
      2.2.1 Aim both cameras at target
      2.2.2 Adjust lens apertures to normalize image contrast
      2.2.3 Obtain range estimates using focus for both cameras
      2.2.4 If the difference in range values is too great, as compared with the depth of field, assume an occlusion is present, and redefine the target to be the nearest point detected with focus
   2.3 Adjust the vergence angle to register image centers
   2.4 Set lenses to intermediate zoom setting
   2.5 Adjust lens apertures to normalize image contrast
   2.6 Obtain stereo images in focus at the image centers; also obtain defocused images for later segmentation

3. Obtain integrated estimate of local surface
   3.1 Obtain depth estimates using focus information
      3.1.1 Detect the in-focus region about the image center for the focused images by comparing these images with the defocused images
      3.1.2 Within the entire in-focus region, obtain focus-based range estimates
   3.2 Detect image features to be matched
      3.2.1 Apply Laplacian-of-Gaussian filters to the focused stereo image pair at several levels of resolution
      3.2.2 Detect zero-crossings in the filtered images
      3.2.3 Mask out any features that are not within the in-focus regions of the original focused images
   3.3 Invoke a modified Hoff–Ahuja stereo algorithm to detect stereo correspondences (not surface map)
      3.3.1 Use the range estimates from Step 3.1.2 to locate candidate matches
      3.3.2 Cluster resulting 3-D points
      3.3.3 Reject false matches and disambiguate multiple matches based on emerging surface characteristics and on values in the composite surface map
   3.4 To self-calibrate, repeat until optimum fit is reached
      3.4.1 Adjust system-calibration parameters for best fit of 3-D points to composite map
      3.4.2 Fit quadratic patches to the resulting 3-D locations
   3.5 Interpolate quadratic patches to produce a dense, local depth map

4. Merge local map with composite map
   4.1 For areas of the local map that do not overlap the composite map, range values are simply added to the composite map
   4.2 For areas that overlap, replace the current depth value in the composite map by the mean of overlapping values

Discussion: Step 1 of this algorithm selects a visual target to be fixated, as discussed in Section IV. Targets are constrained to lie on the border of S_c. Typically, the resulting new camera orientations will extract new parts of surfaces that overlap partially with the current global surface map. The objective function to be minimized is given in (8).

Step 2 implements the exploratory fixation process. At the end of this step, focused stereo images are obtained. In addition, defocused images are extracted to be used in a segmentation process. These images are obtained by focusing the lens sequentially at different depths so that successive depths of field are adjacent in depth. For example, if a lens is to be focused at three different depths u_0, u_{0n}, and u_{0f} in sequence, these depths can be chosen so that the associated depths of field border one another. Since the extremes for these adjacent depths of field coincide, u_{0n} and u_{0f} may be written in terms of u_0 as follows:

$$u_{0n} = \frac{u_0 f(A+C)}{Af - Cf + 2u_0 C} \text{ and}$$
$$u_{0f} = \frac{u_0 f(A-C)}{Af + Cf - 2u_0 C}. \tag{25}$$

These adjacent depth intervals are useful for segmentation. When the depth of field is not large, images may be obtained for these three focus settings that correspond to object depths of u_0, u_{0n}, and u_{0f}. The first of these images is called the "in-focus" image; the goal is to remove from this image all portions that are not within the depth of field for this focus setting. The focus-based objective function is evaluated for windows at corresponding locations in each of the three images. If the in-focus image yields the maximum response to this criterion, this location is retained within the resulting segmented image. Otherwise, the location is rejected.

For Step 3, the stereopsis process accepts initial surface estimates and produces a surface patch about the point of fixation as described in Section II-A. The derived surface patch will be for a part of the scene that is common to the visual fields of both the left and the right cameras. The surface is made to agree with overlapping areas of previously acquired composite-map surfaces (see (18)). To achieve this agreement,

small changes in the camera parameters β are permitted. Step 3, therefore, corresponds to the performance of the outer minimization of (24b).

The merging of local surface patches with the cumulative surface map (Step 4) is integral to this process. Since the newly obtained surface patches typically have partial overlap with previously mapped scene areas, the surface should smoothly extend beyond previously mapped parts. The method used in this work is simply the assumption of the mean range values for areas of overlap when these maps are merged. The next section describes an implementation of the above algorithm.

B. Implementation Details

The algorithm was implemented on the University of Illinois Active Vision System [50] (Fig. 3(b)). This system acquires high-resolution stereo images from cameras that can tilt, pan, and verge under computer control. The lenses are motorized, and the processor can control settings of focus, aperture, and zoom. All computations are performed by a Sun 3/160 workstation. The vergence movements of the cameras can be driven independently, but the implementation utilizes only symmetric vergence so that both cameras verge by the same amount to aim at the point of fixation. This was done for ease of algorithm implementation and should not represent a loss of generality of the method. Two different focal lengths (determined by the zoom settings) are used by this system during surface reconstruction: approximately 105 mm for focus ranging, and approximately 47 mm for stereo surface estimation. The former setting is used to reduce the depth of field and the field of view during the fixation process. The latter setting widens the field of view for stereo processing.

The implementation consists of several independent software modules that correspond to the individual steps of the surface-reconstruction algorithm given in the previous section. The system runs autonomously, but it can be stopped and later restarted where it left off. This is useful because of the long time required by the surface-reconstruction modules. In developing the algorithm/implementation reported in this paper, we have neglected issues of computational speed. The current implementation on a Sun 3/160 takes 2–4 hr per fixation. Clearly, this is far too long for any practical utility of the algorithm. Several ways in which this speed can be increased significantly include using newer and faster machines, parallelizing the computation, and using a faster passive stereo algorithm. We have begun work on addressing all of these methods. For one version of the passive stereo algorithm, a speedup of 15 to 20 has been obtained (relative to the Sun 3) as a result of a parallel implementation on an eight-processor Intel HYPERCUBE [14].

C. Experimental Results

Experiments were conducted with two scenes. In the first scene, the surface to be reconstructed is that of a barrel oriented vertically and resting on a table. The second scene contains a chair for which the surface is to be reconstructed.

1) Scene 1, A Barrel: A diagram of the imaging environment is shown in Fig. 3, and an overview, as seen from the right camera, is shown in Fig. 4. The barrel is conical (approximately cylindrical). Textured surfaces were used so that the images contain a high degree of visual detail.

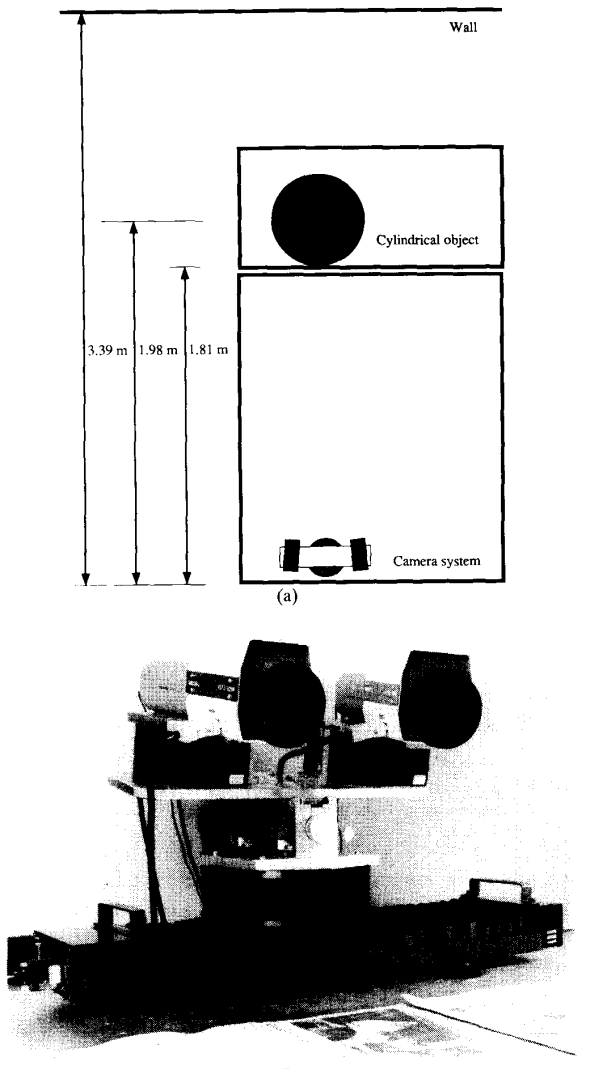


Fig. 3. (a) Top view of imaging environment for scene 1 (not to scale); (b) the University of Illinois Active Vision System.

2) Scene 1, First Fixation: The system accepts an externally specified initial target for fixation. The system then calculates actuator settings and aims both cameras at this 3-D location. The focal lengths are set to 47 mm, and the lens apertures are adjusted to the scene illumination. The stereo image pair acquired at this stage are shown in Fig. 5. The initial target was deliberately chosen so that when both cameras are aimed at it, the line of sight for the left camera intersects the wall, and the line of sight for the right camera intersects the barrel (which is similar to Fig. 1(a)). The fixation process now begins. Both lenses are set to a focal length of 107 mm, and the system manipulates the focus setting of the left camera to obtain a depth estimate for the line of sight of this camera. The

Fig. 4. Overview of scene 1.

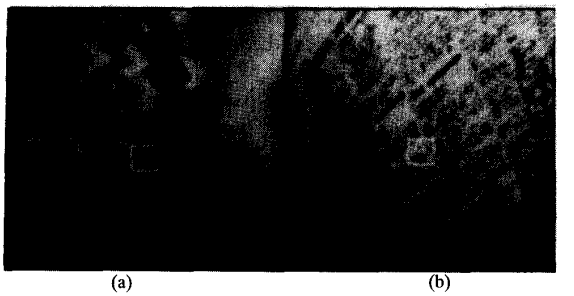
(a) (b)

Fig. 5. Initial left and right stereo images. The scene point that appears at the center of the (a) left image is not visible in the (b) right image.

procedure is to minimize the objective function E_f for a small (49 × 49) window at the image center. To reduce image noise, ten images are acquired at each focus setting and temporally averaged. After the optimal focus setting is found, the system uses a precalibrated equation to map that setting to a distance estimate. From this estimate, a 3-D point in world coordinates is obtained, and the system attempts to fixate that scene point by panning the camera platform and causing the cameras to verge (rotate) inward. This causes the right camera's view of the desired scene point to be obstructed by the barrel; by using focus ranging to estimate the distance along the optical axis of the right camera, the system detects this situation (Fig. 1(b)). The resulting focused images are shown in Fig. 6.

The system reacts to the occlusion condition by attempting to fixate the nearer scene point corresponding to the center of the right image (Fig. 1(c)). The platform and cameras are rotated so that both cameras aim at this point, and a depth estimate from focus is obtained with the left camera to verify the distance. The resulting images are shown in Fig. 7. In this case, the two depth estimates from focus agree closely with one another, relative to the estimated depth of field, and the corresponding 3-D locations are in agreement with the calculated point of vergence. The system therefore assumes that no occlusions are present.

The system now improves the accuracy of the vergence angle through a registration procedure. Using a 49 × 49 window at the center of the left image as a reference, the



(a) (b)

Fig. 6. Initial attempt at fixation. Both cameras are aimed at a point on the distant wall. (a) This point is visible in the left image but (b) is occluded by the barrel in the right image.

(a) (b)

Fig. 7. Focused images of barrel: (a) In the left image, the center is enclosed by a rectangle; (b) two windows are highlighted in the right image.

(a) (b)

Fig. 8. Fixated images of scene for stereopsis. Calibrated focal lengths for the left and right cameras are (a) 47.7 and (b) 47.2 mm, respectively.

system computes the normalized cross-correlation function E_v for 32 windows on each side of the center of the right image and selects the one yielding the highest cross-correlation measure. The location of the best match in the right image, corresponding to a horizontal disparity of 8 pixels, is highlighted in Fig. 7(b). (The vergence process controls horizontal disparity. A vertical disparity is also seen in the figure because the cameras are not perfectly aligned. The calibration process accounts for this small misalignment.) From the location of this window, a correction to the vergence angle is calculated, and the cameras are rotated appropriately. This process repeats, where each time there are half as many correlation windows in the right image, until the horizontal disparity at the image centers is, at most, one pixel.

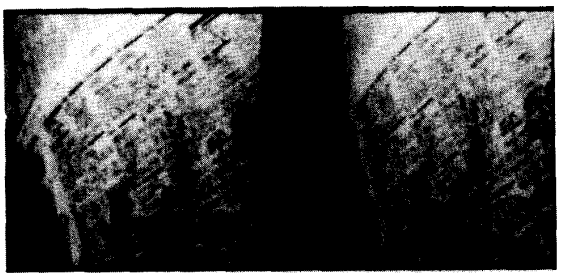


Fig. 9. Segmented 256 × 256 images for first fixation.

The system then causes the lenses to return to the shorter focal length (47 mm), acquires stereo images (Fig. 8), and determines the image/scene region in the vicinity of the point of fixation over which a stereo estimate is to be obtained. To segment out such a region, the two lenses are defocused relative to the fixated surface. Ordinarily (Section VII-A), defocused images corresponding to *nearer* and *farther* distances relative to the object would be obtained. However, because of the near proximity of the scene object to the cameras, only "far-focused" images are obtained in this instance. This is done by focusing the two cameras to points more distant than the fixation point such that the new depth of field is just beyond the current depth of field. Using the resulting defocused images, a segmentation process compares the focused and defocused images separately for the left and right cameras. For each location in the images from one camera, the minimum value of E_f determines the image for which the corresponding scene point is most in focus. The resulting segmentation for the focused images is shown in Fig. 9.

The estimated depth of the point of fixation is used as an initial estimate over the segmented parts of the images. The stereo module is then invoked with these depth points as initial estimates for corresponding scene regions. Depth maps are obtained using the modified Hoff and Ahuja algorithm [8]. The resulting surface map for resolution level 256 × 256 is shown in Fig. 10. Depth values corresponding to a single row of this surface map are plotted in Fig. 11.

3) Scene 1, Second Fixation: The next goal is to select a new target to extend the scene description, which at this stage consists solely of the initial patch. As described earlier, the system examines the border of the current surface map and selects an edge point that is optimum with respect to the target-selection criterion function. For the surface map at hand, the optimum edge point lies along the lower edge of the surface map (Fig. 10) and is located approximately on the surface of the barrel. The system now attempts to fixate a surface location in the vicinity of this target.

From this stage on, the cycle repeats. The system attempts to fixate the new scene point, based on depth estimates from focus and vergence. The system aims the cameras at the new target and estimates a distance along the line of sight of the left camera. The estimated distance does not precisely correspond to the target location, and the aim of the right camera is corrected for this. The right camera then obtains a focus-based range estimate, and no occlusions are detected in this instance. The registration process causes a fine adjustment to the vergence angle. Final stereo images are shown in Fig. 12.

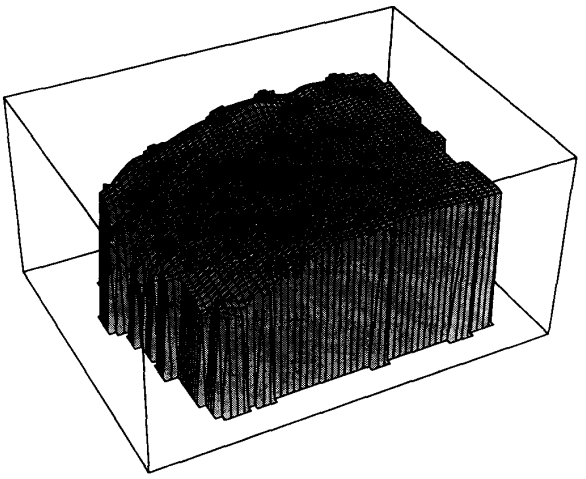


Fig. 10. Surface map for first fixation.

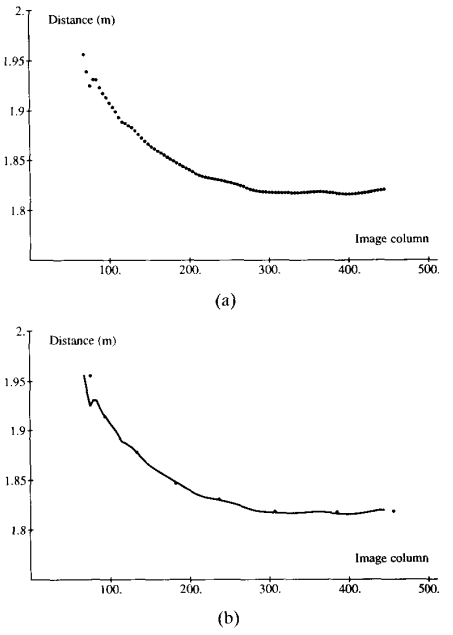


Fig. 11. Computed distances for center row of left image: (a) Locations of individual scene points, as derived by the stereo process, are plotted for this row; (b) several feature correspondences for the same image row were determined by hand for comparison.

The goal is now to use these images to build a local depth map (a second surface patch) that can be merged with the map from the previous fixation (Fig. 10). Furthermore, depth information from the previous fixation is to be used where possible to assist in the construction of the new local map. Recall that this method does not attempt to match features in the current image pair to *features* from previous fixations. Instead, the system accesses only the previously obtained surface map. Equation (18) presented a criterion for enforcing agreement of two surface maps. The optimization

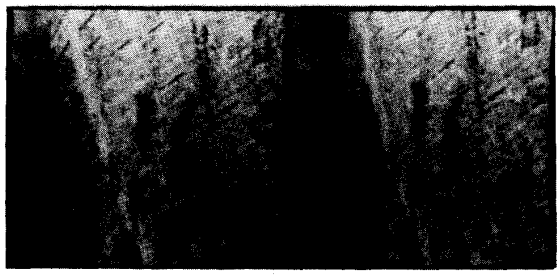


Fig. 12. Final stereo image pair for second fixation of barrel.

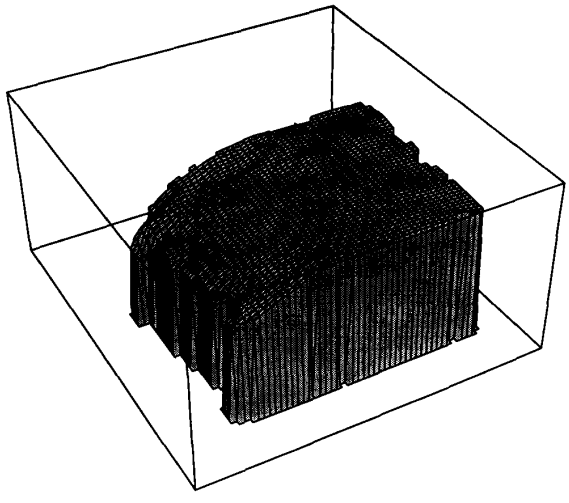
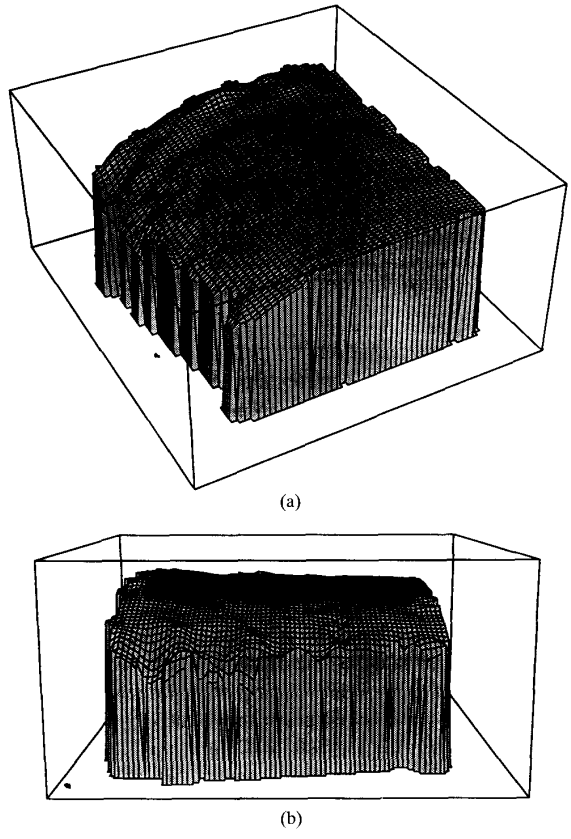


Fig. 13. Optimized local surface map for second fixation (level 256). The local surface has been constructed using the squared difference in absolute depth as an error measure.



(a)

(b)

Fig. 14. Composite surface map after second fixation: (a) View from above; (b) view from the left side.

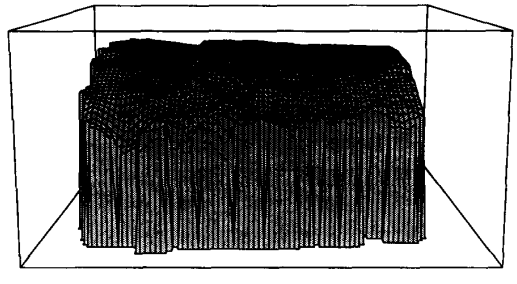


Fig. 15. Merge of the two surface maps without optimization.

procedure implemented here does not use the second term in this equation, effectively setting its relative weight w_{a2} to zero. The vector β is initialized to $\hat{\beta}$ and is then used along with the surface-smoothness constraint to identify stereo correspondences (and thus 3-D point estimates). Using these correspondences, optimization is performed so that the resulting surface will yield the most agreement with the previously obtained composite map. The quantity to be minimized is the sum of squared distances between the stereo 3-D point estimates and overlapping points in the composite map S_c. The optimization variable is the vector β. A full-scale nonlinear search is performed, using a modified Levenberg-Marquardt algorithm. The result of this procedure is a locally optimum set of calibration parameters (β). Using these parameters, a local surface patch S is calculated (Fig. 13), which must be merged with S_c.

Before merging, the system compares the overlapping regions and obtains a measurement of the differences between the depth values. The RMS error is computed per composite-map range element using all difference values and has the value 5.292 mm for these two depth maps. Fig. 14 shows the composite map that results by merging the local maps for the first two fixations. For areas where the maps overlap, the merge process normally retains the mean of any two overlapping z values, as shown here. Without optimization involving calibration parameters, the quality of fit for the two surfaces degrades (Fig. 15). The RMS error here is 8.426 mm, which is a degradation of approximately 59% over the error for the optimized case.

To illustrate the effectiveness of the optimization process, the method was tested in the presence of a deliberate error. The right camera was rotated outward by a small amount, effectively introducing a global horizontal disparity increment in the image. This method was chosen since depth estimates are most sensitive to changes in the vergence angle. The amount of rotation was six motor steps, corresponding to 0.06°, and introduced a depth error of approximately 1 cm.

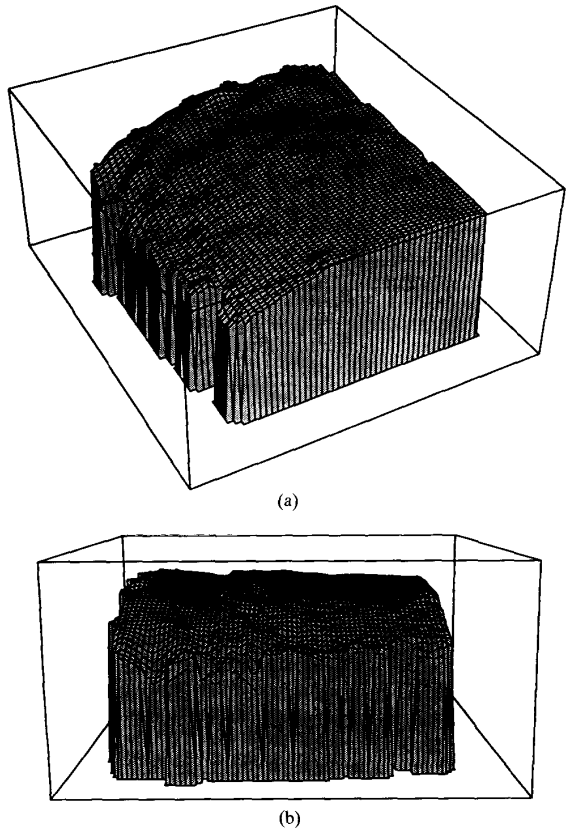


Fig. 16. Composite surface map with deliberate vergence error in second fixation: (a) View from above; (b) view from the left side.

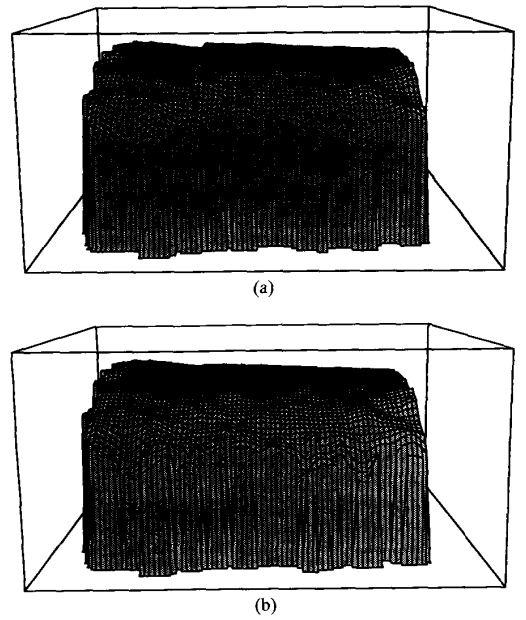


Fig. 17. Composite surface map with deliberate vergence error with optimization of calibrated imaging parameters: (a) Confidence measure of 0.85; (b) confidence measure of 0.5.

The resulting composite map, without optimization, is shown in Fig. 16. The RMS error for the overlapping regions is now 19.0 mm, or approximately 3.59 times the error of the original, optimized case. If the optimization algorithm is now applied, the second surface is "drawn" toward the first as shown in Fig. 17. Two different rows for these surface maps are shown in Fig. 18. The error for these overlapping regions drops to 11.358 and 4.123 mm as the coefficients $w_{\beta i}$ are decreased relative to w_{a1} and w_{a2} in (18).

4) Scene 1, Subsequent Fixations: The algorithm now continues, automatically selecting scene targets, fixating in the vicinity of the targets, building a local surface map, and merging it with the evolving composite surface map. Fig. 19 shows the state of the composite map after each of several fixations.

As discussed in Section III, a major motivation for this work comes from the observation that for scenes with large depth ranges, a single imaging configuration cannot yield an accurate and complete surface map. One reason is that no single, constant-depth initial estimate may suffice. To demonstrate this, the algorithm was invoked several times with the same image pair (Fig. 8) but with initial estimates of decreasing accuracy. Four resulting depth maps are shown in Fig. 20. The depth maps "collapse" after the estimates reach a certain threshold. This is because the search windows for correspondences are derived directly from the initial estimates. When these windows no longer contain the correct matches, any candidate correspondences tend to be random and will support a surface only by chance.

5) Scene 2, A Chair: The barrel in Scene 1 is now replaced by a sofa chair, and the system performs a sequence of fixations on this object. Fig. 21 is an overview of this scene with the camera system shown in the foreground. The operation of the system is similar to that for the previous scene. The system first fixates a location on the backrest of the chair. The resulting surface map is shown in Fig. 22(a). The remainder of Fig. 22 shows the state of the composite map at several instants during the scanning process, illustrating the evolution of the map for this object. After the final (36th) fixation, approximately one third of the visible surface area of the chair has been mapped.

D. Performance Evaluation

In order to assess the quality of reconstructed surfaces obtained with this method, we compared the depth values obtained for the barrel scene (Scene 1) with a model of a cone assumed to approximate the barrel. For the entire composite map, resulting from eight fixations, there are 28 782 range points; the mean absolute error over all range points is 3.156 mm, and the RMS error is 5.006 mm.

These numbers improve if the border regions of the composite map are ignored in the error analysis. This is true because stereo reconstruction is least reliable near occluding boundaries, particularly if the object curves away from the cameras. For a rectangular window of 17 985 range points from the center of the composite map, the mean absolute error is 2.137 mm, and the RMS error is 2.733 mm. Since the visible

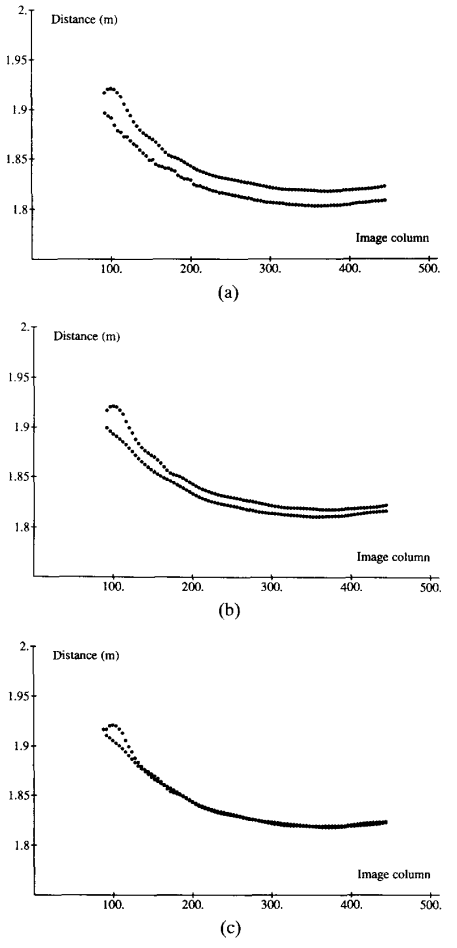


Fig. 18. Single rows of surface maps with deliberate vergence error: (a) Confidence measure of 1.0 (no optimization); (b) confidence measure of 0.85; (c) confidence measure of 0.5.

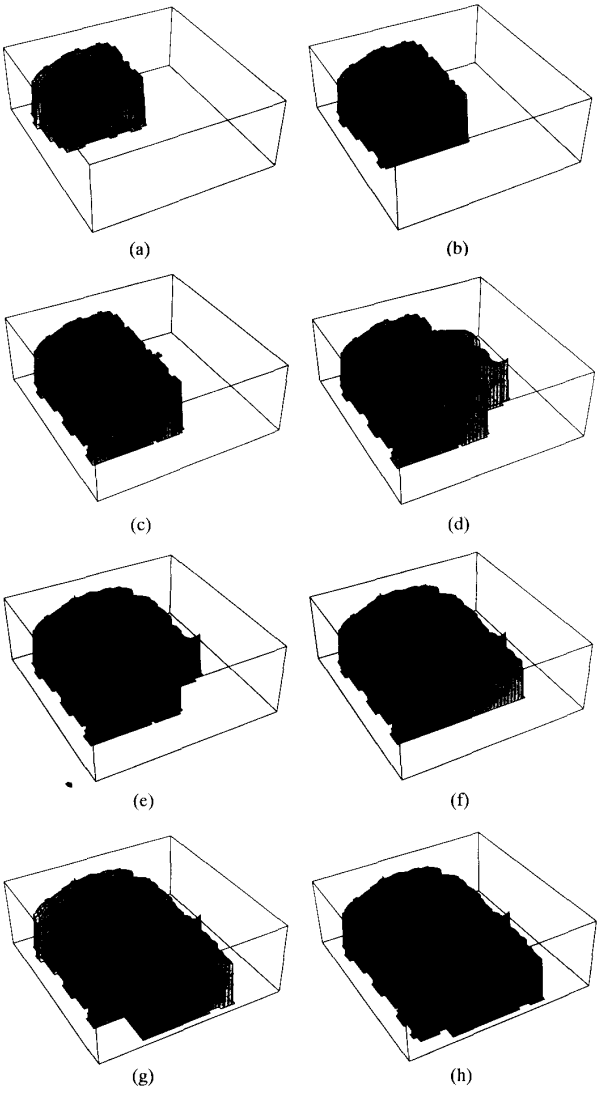


Fig. 19. Evolving scene description: (a) Surface map after first fixation; (b) resulting composite surface map after two fixations; (c) after three fixations; (d) after four fixations; (e) after five fixations; (f) after six fixations; (g) after seven fixations; (h) after eight fixations.

portion of the barrel lies in the approximate depth range of 1.8 to 2.0 m, these average error measurements correspond to approximately 0.15% of the depth range. This figure includes the error due to incorrect knowledge of the ground truth as discussed next.

During this evaluation of performance analysis, a 2-D array of error values was generated. If this is displayed as an image, visual analysis confirms a good fit of the composite map with the model of the barrel. There are, however, several notable factors contributing to the observed errors. A small, global offset is apparent, which implies that the barrel does not possess a perfectly vertical axis. Since our model assumes a right circular cone with a vertical axis, this small mismatch contributes to the overall error. There are also small, higher frequency components visible in the error image, resulting in several local maxima and minima. These small ripples appear to result from the fact that the barrel is not perfectly conical in shape. (This was illustrated in Fig. 11.) The effect is enhanced by paper that has been used to wrap the barrel. Close examination of the visible surface of the paper clearly shows small ripples in the shape. Finally, artifacts that result from the method used to merge local depth maps into the composite map are visible. After the optimization process is complete for one fixation, simple averaging is used to merge overlapping portions of these two maps into an updated composite map. Additional accuracy could be gained if a more sophisticated method were used.

Overall, the resulting surface map (except for a few locations near the border of the map) is sufficiently accurate that we could not detect any errors by hand measurement, such as by using a ruler.

VIII. Summary

We have argued that individual depth cues such as stereo disparity, camera vergence, and focus are not sufficient by

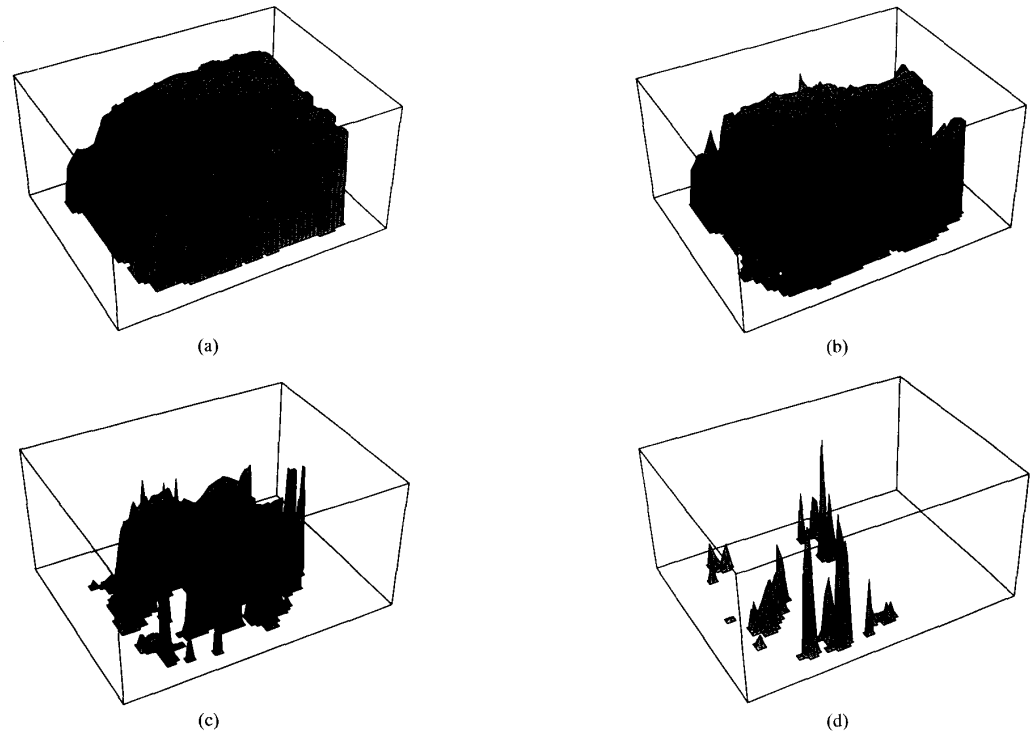


Fig. 20. Impact of inaccurate initial depth estimates: (a) Initial estimate is 1.9 m, which is 10 cm more distant than the closest portion of the object; (b) initial estimate incorrect by 20 cm; (c) initial estimate incorrect by 30 cm; (d) initial estimate incorrect by 40 cm.

themselves for surface reconstruction of large scenes having large depth ranges. Rather, these cues must be used in a tightly integrated mode to exploit their complementary strengths while eliminating their weaknesses. Doing so yields a more powerful and complete mechanism for surface estimation than provided by any of the individual cues. The method of integration described in this paper depends on the active selection of imaging parameters (vergence, focus, aperture, and zoom) based on the evolving scene description. This makes the system *active* in the sense that image *acquisition* is tightly coupled with image *analysis* for the purpose of surface estimation.

There are two distinct phases in the approach described here: *visual target selection* and *surface reconstruction*. Each phase is formulated as an optimization problem by devising suitable objective functions to be minimized. Each objective function incorporates and balances several individual criteria. As a part of the surface-reconstruction process, this method implements a phase of *exploratory fixation*, during which occlusions are detected and avoided using focus and vergence information.

The approach does not require rigid values of calibrated imaging parameters. Instead, these parameters are considered to be somewhat flexible when local surface maps are constructed. This "elasticity" permits small adjustments in these values; these adjustments are used to to minimize the discrepancies in depth between overlapping portions of surface maps obtained from different fixations. Such self-calibration represents integration of camera calibration with surface reconstruction and avoids a constant need for user-supplied calibration targets or distinctive features (such as ridge boundaries) that are often required by camera calibration algorithms.

Fig. 21. Overview of scene 2: a stuffed chair. The chair is in the background and is to be scanned by the camera system, which is shown in the foreground.

This method is intended for scenes which contain a single continuous surface which is free of depth discontinuities. Dense composite surface maps are automatically constructed by exploiting the benefits of integration, as described above.

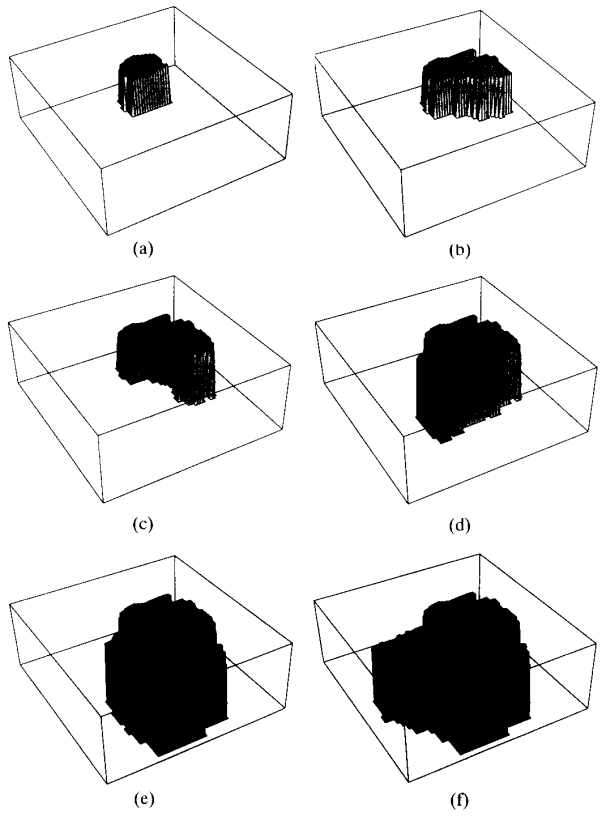


Fig. 22. Composite surface map of scene 2. The initial maps represent a portion of the backrest of the chair: (a) Surface map after first fixation; (b) after three fixations; (c) surface map of chair after six fixations; (d) after 10 fixations; (e) after 25 fixations; (f) after 36 fixations.

An implementation of this method has been tested and has proven to be highly accurate for surfaces which are smooth and which have adequate visual detail. A performance analysis has shown an average error of less than 0.15% at a distance of approximately 2 m. Active surface reconstruction for scenes containing depth discontinuities involves significant additional complexity, and is the subject of another paper [6].

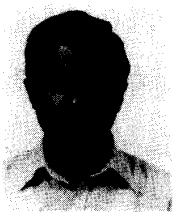
REFERENCES

[1] R. Bajcsy, "Active perception vs. passive perception," in *Proc. Workshop Comput. Vision*, Oct. 1985, pp. 55–59.
[2] ——, "Perception with feedback," in *Proc. DARPA Image Understanding Workshop*, Apr. 1988, pp. 279–288.
[3] G. Sperling, "Binocular vision: A physical and a neural theory," *Amer. J. Psychol.*, vol. 83, pp. 461–534, 1970.
[4] E. P. Krotkov, *Active Computer Vision by Cooperative Focus and Stereo*. Springer-Verlag, 1989.
[5] D. H. Ballard and A. Ozcandarli, "Eye fixation and early vision: Kinetic depth," in *Proc. Second Int. Conf. Comput. Vision*, Dec. 1988, pp. 524–531.
[6] S. Das and N. Ahuja, "Multiresolution image acquisition and surface reconstruction," in *Proc. Third Int. Conf. Comput. Vision*, Dec. 1990.
[7] W. Hoff and N. Ahuja, "Surfaces from stereo," in *Proc. DARPA Image Understanding Workshop*, Dec. 1985, pp. 98–106.
[8] ——, "Surfaces from stereo: Integrating feature matching, disparity estimation and contour detection," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. PAMI-11, pp. 121–136, Feb. 1989.
[9] R. D. Eastman and A. M. Waxman, "Disparity functionals and stereo vision," in *Proc. DARPA Image Understanding Workshop*, Dec. 1985, pp. 245–254.
[10] T. E. Boult and L. -H. Chen, "Synergistic smooth surface stereo," in *Proc. Second Int. Conf. Comput. Vision*, Dec. 1988, pp. 118–122.
[11] A. L. Abbott and N. Ahuja, "Surface reconstruction by dynamic integration of focus, camera vergence, and stereo," in *Proc. Second Int. Conf. Comput. Vision*, Dec. 1988, pp. 532-543.
[12] E. Altman and N. Ahuja, "A dynamical systems approach to integration in stereo," in *Proc. DARPA Image Understanding Workshop*, Sept. 1990, pp. 423–427.
[13] W. Hoff and N. Ahuja, "Extracting surfaces from stereo images: An integrated approach," in *Proc. First Int. Conf. Comput. Vision*, June 1987, pp. 284–294.
[14] A. N. Choudhary, S. Das, N. Ahuja, and J. H. Patel, "Surface reconstruction from stereo images: An implementation on a hypercube multiprocessor," in *Proc. Fourth Conf. Hypercube Concurrent Comput. Applications*, Mar. 1989.
[15] W. E. L. Grimson, *From Images to Surfaces: A Computational Study of the Human Early Visual System*. Cambridge, MA: MIT Press, 1981.
[16] T. J. Olson and R. D. Potter, "Real-time vergence control," in *Proc. IEEE Conf. Comput. Vision Patt. Recogn.*, 1989, pp. 404–409.
[17] B. K. P. Horn, "Focusing," Rep. No. 160, MIT Artificial Intell. Lab, 1968.
[18] J. M. Tenenbaum, "Accommodation in computer vision," Ph.D. Dissertation, Stanford Univ., 1971.
[19] E. P. Krotkov, "Focusing," Rep. No. MS-CIS-86-22, GRASP Lab., Univ. of Pennsylvania, Apr. 1986.
[20] R. A. Jarvis, "Focus optimisation criteria for computer image processing," *Microscope*, vol. 24, pp. 163–180, 1976.
[21] G. Ligthart and F. C. A. Groen, "A comparison of different autofocus algorithms," in *Proc. Sixth Int. Conf. Patt. Recogn.*, Oct. 1982, pp. 597–600.
[22] S. Das and N. Ahuja, "Integrating multiresolution image acquisition and coarse-to-fine surface reconstruction from stereo," in *Proc. IEEE Workshop Interpretation 3-D Scenes*, Nov. 1989, pp. 9–15.
[23] D. Noton and L. Stark, "Scanpaths in saccadic eye movements while viewing and recognizing patterns," *Vision Res.*, vol. 11, pp. 929–942, 1971.
[24] J. K. O'Regan and A. Lévy-Schoen, "Integrating visual information from successive fixations: Does trans-saccadic fusion exist?," *Vision Res.*, vol. 23, no. 8, pp. 765–768, 1983.
[25] R. Groner, G. W. McConkie, and C. Menz, *Eye Movements and Human Information Processing*. Amsterdam: North-Holland, 1985.
[26] A. Lévy-Schoen, "Flexible and/or rigid control of oculomotor scanning behavior," in *Eye Movements: Cognition and Visual Perception* (D. F. Disher, R. A. Monty and J. W. Senders, Eds.). Hillsdale, NJ: Lawrence Erlbaum, 1981, pp. 299–314.
[27] J. M. Findlay, "Local and global influences on saccadic eye movements," in *Eye Movements: Cognition and Visual Perception* (D. F. Disher, R. A. Monty and J. W. Senders, Eds.). Hillsdale. NJ: Lawrence Erlbaum, 1981, pp. 171–179.
[28] V. Božkov, Z. Bohdanecký, and T. Radil-Weiss, "Perception, Exploration and eye displacements," in *Cognition and Eye Movements* (R. Groner and P. Fraisse, Eds.). Amsterdam: North-Holland, 1982, pp. 24–33.
[29] N. H. Mackworth and A. J. Morandi, "The gaze selects informative details within picture," *Perception Psychophys.*, vol. 2, pp. 547–552, 1967.
[30] P. J. Locher and C. F. Nodine, "Symmetry catches the eye," in *Eye Movements: From Physiology to Cognition* (J. K. O'Regan and A. Lévy-Schoen, Eds.). Amsterdam: North-Holland, 1987, pp. 353–361.
[31] C. Koch and S. Ullman, "Selecting one among the many: A simple network implementing shifts in selective visual attention," MIT AI Memo 770, 1984.
[32] J. J. Clark and N. J. Ferrier, "Modal control of an attentive vision system," in *Proc. Second Int. Conf. Comput. Vision*,, Dec. 1988, pp. 514–513.
[33] D. J. Coombs and C. M. Brown, "Intelligent gaze control in binocular vision," in *Proc. Fifth IEEE Int. Symp. Intell. Contr.*, Sept. 1990.
[34] P. J. Burt, "Algorithms and architectures for smart sensing," in *Proc. DARPA Image Understanding Workshop*, Apr. 1988, pp. 139–153.
[35] A. Shmuel and M. Werman, "Active vision: 3D from an image sequence," in *Proc. 10th Int. Conf. Patt. Recogn.*, June 1990, pp. 48–54.
[36] C. M. Schor and L. B. Ciuffreda, *Vergence Eye Movements: Basic and Clinical Aspects*. Boston: Butterworths, 1983.
[37] J. M. Foley, "Primary distance perception," in *Handbook of Sensory Physiology*. Berlin: Springer-Verlag, 1978.
[38] V. V. Krishnan and L. Stark, "A heuristic model for the human vergence movement system," *IEEE Trans. Biomed. Eng.*, vol. BME-24, no. 1, Jan. 1977.

[39] G. K. Hung and J. L. Semmlow, "Static behavior of accommodation and vergence: Computer simulation of an interactive dual-feedback system," *IEEE Trans. Biomed. Eng.*, vol. BME-27, no. 8, Aug. 1980.
[40] C. M. Schor, "The relationship between fusional vergence eye movements and fixation disparity," *Vision Res.*, vol. 19, no. 12, pp. 1359–1367, 1979.
[41] D. Marr and T. Poggio, "A computational theory of human stereo vision," in *Proc. Royal Soc. London*, vol. B, no. 204, pp. 301–328, 1979.
[42] D. Marr, *Vision*. San Francisco: Freeman, 1982.
[43] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active vision," in *Proc. First Int. Conf. Comput. Vision*, June 1987, pp. 35–54.
[44] A. Bandopadhay, B. Chandra, and D. H. Ballard, "Egomotion using active vision," in *Proc. IEEE Conf. Comput. Vision Patt. Recogn.*, June 1986, pp. 498-503.
[45] D. Geiger and A. Yuille, "Stereopsis and eye-movement," in *Proc. First Int. Conf. Comput. Vision*, June 1987, pp. 306–314.
[46] F. P. Ferrie and M. D. Levine, "Integrating descriptions from multiple views," in *Proc. Workshop Comput. Vision*, Dec. 1987.
[47] B. Kamgar-Parsi, J. L. Jones, and A. Rosenfeld, "Registration of multiple overlapping range images: Scenes without distinctive features," in *Proc. IEEE Conf. Comput. Vision Patt. Recogn.*, 1989, pp. 282–290.
[48] N. Ayache and O. D. Faugeras, "Building, registrating and fusing noisy visual maps," *Int. J. Robotics Res.*, vol. 7, no. 6, Dec. 1988.
[49] H. Takahashi and F. Tomita, "Self-calibration of stereo cameras," in *Proc. Second Int. Conf. Comput. Vision*, Dec. 1988, pp. 123–128.
[50] A. L. Abbott, "Dynamic integration of depth cues for surface reconstruction from stereo images," Ph.D. Dissertation, Univ. of Illinois, 1990.
[51] M. A. Gennert and A. L. Yuille, "Determining the optimal weights in multiple objective function optimization," in *Proc. Second Int. Conf. Comput. Vision*, Dec. 1988, pp. 87–89.

A. Lynn Abbott (M'80) received the B.S. degree from Rutgers University, New Brunswick, NJ, in 1980, the M.S. degree from Stanford niversity, Stanford, CA, in 1981, and the Ph.D. degree from the University of Illinois, Urbana-Champaign, in 1990, all in electrical engineering.

From 1980 to 1985, he was a Member of the Technical Staff at AT&T Bell Laboratories, Holmdel, NJ. He is currently an Assistant Professor in the Bradley Department of Electrical Engineering, Virginia Polytechnic Institute and State University, Blacksburg. His research interests include computer vision, robotics, artificial intelligence, and computer architecture.



Narendra Ahuja (F'92) received the B.E. degree with honors in electronics engineering from the Birla Institute of Technology and Science, Pilani, India, in 1972, the M.E. degree with distinction in electrical communication engineering from the Indian Institute of Science, Bangalore, India, in 1974, and the Ph.D. degree in computer science from the University of Maryland, College Park, in 1979.

From 1974 to 1975, he was Scientific Officer in the Department of Electronics, Government of India, New Delhi. From 1975 to 1979, he was at the Computer Vision Laboratory, University of Maryland, College Park. Since 1979, he has been with the University of Illinois at Urbana-Champaign, where (since 1988) he is currently a Professor in the Department of Electrical and Computer Engineering, the Coordinated Science Laboratory, and the Beckman Institute. His interests are in computer vision, robotics, image processing, image synthesis, and parallel algorithms. He has been involved in teaching, research, consulting, and organizing conferences in these areas. His current research emphasizes integrated use of multiple image sources of scene information to construct 3-D descriptions of scenes, the use of integrated image analysis for realistic image synthesis, the use of the acquired 3-D information for navigation, and multiprocessor architectures for computer vision.

Dr. Ahuja was selected as a Beckman Associate in the University of Illinois Center for Advanced Study for 1990–1991. He received University Scholar Award (1985), Presidential Young Investigator Award (1984), National Scholarship (1967–1972), and President's Merit Award (1966). He has coauthored the books *Pattern Models* (New York: Wiley, 1983) with B. Schachter, and *Motion and Structure from Image Sequences* (New York: Springer-Verlag, 1992) with J. Weng and T. Huang. He is on the editorial boards of the journals IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE; *Computer Vision, Graphics, and Image Processing*; *Journal of Mathematical Imaging and Vision*; and *Journal of Information Science and Technology*. He is a fellow of the American Association for Artificial Intelligence and a member of the Association for Computing Machinery, the Society of Photo-Optical Instrumentation Engineers, and the Optical Society of America.