SURFACE RECONSTRUCTION BY DYNAMIC INTEGRATION OF FOCUS, CAMERA VERGENCE, AND STEREO

A. LYNN ABBOTT AND NARENDRA AHUJA

Coordinated Science Laboratory
University of Illinois
Urbana, IL 61801 USA

ABSTRACT

This paper concerns estimation of surface maps for real scenes having a wide field of view and a wide range of depths. Much research has emphasized stereo disparity as a source of depth information. To a lesser extent, camera focus and camera vergence have also been investigated for their utility in depth recovery. We argue that these sources of visual information have mutually complementary strengths and weaknesses, and to obtain surface maps for real scenes these processes must be integrated. Such integration requires active control of camera orientations and imaging parameters to dynamically and cooperatively interleave image acquisition with surface estimation. Accordingly, a global surface map of the visual field is synthesized by systematically scanning the scene, and combining estimates of adjacent, local surface patches, each acquired by an intermediate camera configuration and having a small depth range. We present an algorithm to perform this integration, and describe its implementation on a dynamic stereo-camera imaging system. Experimental results are presented to demonstrate the superior performance of the integrated system over that of each of its components.

1. INTRODUCTION

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 suitable camera orientations, and, of course, with the area of interest in proper focus. In general, a three-dimensional (3D) scene point projects onto different relative locations in the two stereo images, and when the imaging geometry is known, the disparity between these two locations provides an estimate of the corresponding 3D position. For most past work in computational stereo vision, stereo disparity has served as the *only* source of 3D information.

Researchers have also considered other visual cues, usually as independent sources of depth information. For example, the role of camera focus as a means of depth estimation has been studied [Horn68, Krot86b, Subb87]. Some researchers, however, have considered mutual cooperation among these and other cues. Marr and Poggio [Marr79, Marr82] point out the role of eye movements in providing large relative image shifts for matching stereo images having large disparities. Sperling [Sper70] presents a model for the interactions of vergence, accommodation (focus), and binocular fusion. Geiger and Yuille [Geig87] describe a framework for using small vergence changes to help disambiguate stereo correspondences. Erkelens and Collewijn [Erke85] discuss interactions between vergence and stereo for biological systems. However, only limited use has been made of these sources in developing computational approaches [Krot86a], especially in a mutually cooperative mode such as discussed in [Sper70] and [Krot87].

This paper is concerned with the integrated use of focus, camera vergence and stereo disparity information for surface estimation. The motivation for such integration comes from the following observations: At any given time during imaging, sharp images can be acquired only for narrow parts of the visual field, capturing a limited depth range. The specific part of the visual field and the depth range imaged are determined by the camera vergence and focus used. However, real scenes are wide and deep; therefore the camera vergence and focus must be controlled to scan the scene to acquire complete data. The global surface map of the scene must be synthesized from local, partial maps obtained using the local data. To accomplish this, the acquired images can be analyzed for stereo correspondences, and a surface map obtained from the associated disparities. This partial surface map may then be used to direct movement of the cameras to new, unmapped portions of the scene. Thus a cooperation between camera motion and image analysis, or equivalently between image acquisition and 3D surface extraction, is necessary.

This paper presents a computational approach to accomplish the integration mentioned above. We first briefly discuss the stereo, vergence and focus processes individually as sources of 3D information (Section 2). We then argue that these sources complement each other and should be used in an integrated mode (Section 3). In Section 4, we describe some models of integration for biological vision, and present a computational model for dynamic surface reconstruction that we have developed and tested. In Section 5, we describe an algorithm that uses the integration approach. Section 6 describes an implementation of the above algorithm, and presents experimental results. Section 7 presents a summary.

2. STEREO DISPARITY, CAMERA VERGENCE, AND FOCUS AS DEPTH CUES

The binocular cues of stereo disparity and camera vergence and the monocular cue of focus have long been recognized as important sources of 3D information. Since all three play important roles in our approach, we will first briefly describe each of these cues individually, and review past work on using these cues to estimate surfaces.

2.1. Stereo Disparity

Many algorithms have been developed for estimation of surfaces from stereo images of a scene. Typically, the images are assumed to have been acquired from suitable viewpoints, and with knowledge of the imaging geometry provided. The paradigm used by these algorithms is: (1) detect suitable features in each image, (2) match corresponding features to determine their depths, and (3) interpolate to obtain a complete depth map. The features used are either edge-based or area-based. Edge-based algorithms use intensity edges as features and attempt to match individual edge points [Arno78, Bake81, Barn80, Grim81, Hend79, Kim86, Marr79, Ohta85], or linear edge segments which consist of chains of aligned edge points [Ayac85, Lim1987, Medi85]. These algorithms complete the matching process before surface interpolation is performed. depth map. Uniqueness of matching is only enforced by conditions that involve simple local relationships among disparity values and not the properties of the resulting surface.

We have developed an approach [Hoff87, Hoff85] that uses a piecewise surface smoothness constraint to obtain a surface map from two stereo images taken using a fixed, known camera configuration. 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. Thus, the surface characteristics are used to resolve matching ambiguities, and matching decisions are made so that the resulting surfaces are piecewise smooth.

Most algorithms, including ours, require an externally specified, coarse, initial disparity/surface estimate which is refined using stereo analysis to obtain a more accurate surface description. Without such an estimate, exhaustive search for correspondences would be required.

2.2. Camera Vergence

Camera vergence is important in surface estimation for two principal reasons: the 3D location of a point can be computed from knowledge of the vergence angle; secondly, camera vergence rotations reduce binocular disparity to a range suitable for stereopsis. In this section we describe the imaging geometry for vergence and the process of fixation.

2.2.1. Imaging geometry. Consider a stereo pair of cameras having their optical axes in the horizontal plane, so that each can rotate about its vertical axis. For simplicity, assume that the two cameras are at the same height, and the vertical rotation axis for each camera passes through its optic point. The center of each image is the projection of a scene point which lies on the optical axis of that camera. Referring to Figure 1, angle v is known as the *vergence angle*, and the *baseline* is the distance separating the optic points of the two cameras. The intersection of the optical axes of the cameras in front of the cameras is known as the *point of vergence*. From knowledge of the baseline distance and the rotation angles, it is possible to determine the vergence angle and the 3D location of the vergence point.

2.2.2. Vergence and fixation. When the point of vergence is known to lie on some surface in the scene, this is known as the *point of fixation*, and it is possible to use vergence information to estimate the location of the surface. It is a fundamental

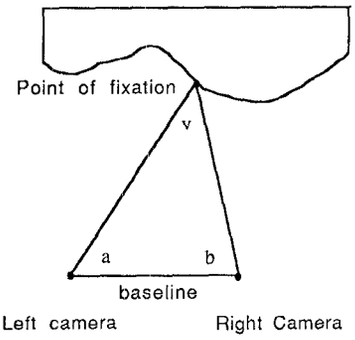

Figure 1: Top view of verged cameras. The vergence angle v can be calculated from knowledge of the baseline and the angles a and b.

problem, however, to verify that both cameras are in fact aimed at the same 3D scene location. A common approach to this problem is to vary the vergence angles so that the two images match, or are in registration near their centers.

Registration is a basic problem in image processing and cartography. For translational registration, one image array is shifted with respect to the other in search of the optimum value of a similarity criterion function. The most common methods of translational registration are minimum-distance and cross-correlation. The minimum-distance approach attempts to minimize the p-distance metric d, defined as

$$d_p(s,t) = \iint |I_l(x,y) - I_r(x+s, y+t)|^p \, dxdy$$

where I_l and I_r are the left and right images, and $p \geq 1$. For the normalized cross-correlation approach, it is necessary to maximize this similarity measure:

$$d^2(s,t) = \frac{\left[\iint I_l(x,y)\, I_r(x+s, y+t)\, dxdy\right]^2}{\left[\iint I_l^2(x,y)\, dxdy\right]\left[\iint I_r^2(x+s, y+t)\, dxdy\right]}$$

While it is typically assumed that the search image does indeed contain an appropriate subimage, no guarantee exists that *any* registration method will produce a correct or unique result [Barn72]. For close-range applications, these methods can fail when the surface gradient is sufficiently large relative to the image planes. Other problems, such as image periodicities and insufficient detail, can also lead to incorrect registration.

2.2.3. Occlusion. In general, it is not possible to fixate every scene point because nearby objects may occlude surfaces that are more distant. For example, see Figure 2. The left focal axis is aimed at point P (Figure2a). If the right camera is also aimed at this point (Figure 2b), point Q obstructs the view. Point P therefore cannot be the point of fixation. For vergence information to be useful, the two cameras can either aim at a nearby point Q (Figure 2c) or try to fixate another distant point R (Figure 2d).

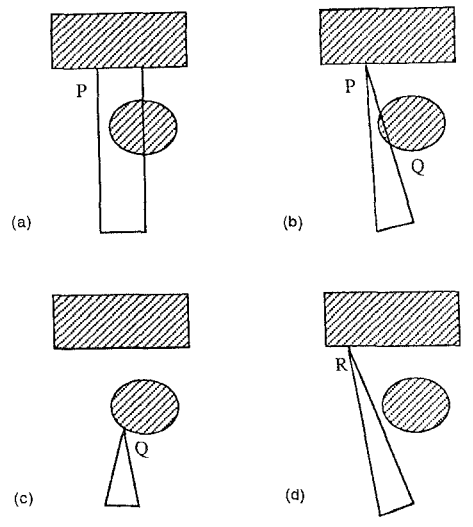

Figure 2: Demonstration of occlusion, using top views of verging cameras. Initially, scene point P projects onto the center of the left image (**a**).An attempt to fixate P brings the image of point Q on the circular object into the center of the right image (**b**). Because of this occlusion, the system could now attempt to fixate either the near point Q, as shown in (**c**), or could now try to fixate a point R on the distant surface (**d**).

Very little research has addressed the problem of occlusion within the context of fixation. We refer to the active process of detecting and avoiding occlusions during fixation as *exploratory fixation*. This will be addressed again in Section 5.

2.3. Focus

The degree of image blur for a particular scene object is directly related to the focus setting of the camera lens. Traditional approaches seek to vary the focus setting of the lens until image sharpness is maximized. The focus setting can be mapped to a depth measurement either mathematically or with predetermined look-up tables.

Mathematically, a scene point is in focus when the lens law is satisfied. Applying the thin lens model for a uniform medium, this law is formulated as

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f}$$

where u is the distance from the lens center to the scene point, v is the distance from lens center to the focal plane, and f is the focal length of the lens. If we know that a point is in focus, and if v and f are known, then from the lens equation we can in principle 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 point-spread function. This acts as a low-pass filter, and results in a loss of high spatial frequency components in the image. Measures of high-frequency content can therefore be used to develop a criterion function which assumes its optimum value when the image blur is minimized and the image is in sharpest focus.

Several autofocus methods have been proposed in the past. Horn [Horn68] describes a Fourier-transform method, in which the normalized high-frequency energy from a one-dimensional FFT is used as the criterion function. A survey and comparison of other criterion functions for focus is presented in [Ligt82]. The criterion functions described therein make use of such measures as signal power, gray level standard deviation, thresholded pixel counts, and summation of squared gradient in one dimension.

The accuracy of the range estimates from focus depends on the available depth of field, which in turn depends on the lens parameters of focal length and aperture. Narrower depths of field are implied for larger focal lengths and larger apertures, as long as imaging elements are not driven into saturation from lighting conditions. For given imaging parameters, overall accuracy of the estimated depth depends on the depth of the scene point of interest.

3. THE NEED FOR INTEGRATION

Each of the above processes may give an estimate of scene depth independently. Here we discuss the benefits of cooperation among these processes, and the need for camera movements in surface reconstruction for real scenes. First we discuss some salient characteristics of these depth cues.

- Focus provides an estimate of depth whose accuracy decreases with increasing object depth as well as with increasing depth of field.
- Vergence provides a depth estimate of the point of intersection of the optical axes. This estimate is quite accurate for large vergence angles but accuracy decreases for objects far away. However, to use the vergence cue, it must be ensured that the two cameras actually are aimed at a single scene point. This may not be the case if the view of one of the cameras is obstructed (Figure 2b).
- Stereo disparity provides an accurate depth estimate only if an accurate coarse estimate is available initially. Thus, if the scene has a large depth range, it may not suffice to give, for example, a constant-depth surface as a coarse surface estimate. Further, since the stereo analysis depends on locations of features (eg., edges) and their correspondences, it is necessary that the features appear sharp and well localized. This requires that the scene be well in focus when imaged.

Real scenes are often wide, and have large depth ranges. To map real surfaces, therefore, none of the above cues is sufficient by itself. Each has its own limitations and strengths. However, it is interesting to note that their strengths and weaknesses are complementary. For example, while stereo disparity can provide accurate surface reconstruction, it requires a coarse initial surface estimate which can be readily provided by focus and/or vergence. However, 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. For relatively close objects, fixation can be verified by ensuring that the depth estimates for the

image centers provided by the focus process are for the same 3D point.

For close objects, the focus and fixation processes may also be used to detect object boundaries, thus simplifying the occlusion problem during stereo analysis. The stereo process can use this additional evidence for occlusion to identify the regions in each image which could not have correspondences in the other image. Such regions can then be excluded from consideration in the surface estimation process.

Another problem that arises when stereo alone is used is that in any given configuration a surface estimate can be obtained for only a limited part of the complete visual field of interest. The part of the visual field which 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 by changing the orientations of the two cameras so that their optical axes intersect in different parts of the visual field. 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 surface may not give disparity values in a workable range; for example, the parts of the surface much closer than the point of fixation may give disparity values on the order of image dimensions, whereas those parts much farther may give disparities that are too small (less than a pixel) for depth recovery. This problem may be remedied by obtaining depth estimates of small parts of the surfaces at a time, having small depth ranges. 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.

4. A COMPUTATIONAL MODEL OF INTEGRATION

To obtain surface maps for real scenes, camera directions and other imaging parameters must be changed to image different parts of the scene. Like human eyes, the cameras must pan and tilt, converge and diverge, focus on near and far objects to acquire stereo images for different fixation points. This results in a dynamic data acquisition system in which surface estimation is integrated with image acquisition, both being performed over small portions of the scene repeatedly.

There are two major consequences of this. First, surface estimation is performed over the scene in a piecewise fashion and the local estimates must be combined to build a global surface description. Secondly, the next camera configuration for image acquisition is determined by the current configuration and the current state of the cumulative surface map. As the scan of the scene continues, depth maps generated for visual subfields around different fixation points must be merged to generate the composite depth map of the entire visual field, possibly having a much larger global depth range than the individual local maps.

The cameras must be under "active" control, i.e., the imaging parameters must be determined by the current state of the system. The term "active" has been used by some researchers simply to indicate that information from several images is to be integrated, although new camera parameters may not be based on information extracted from those images. Also, active vision is not to be confused with active *sensing* (eg., laser ranging). The essential elements of the active surface reconstruction paradigm, for a static scene,[1] can be represented by the following repeating pair of operations:

1) Visual target selection
2) Surface estimation in the target area

Here, a visual target means a potential point of fixation. An autonomous system must be able to select visual targets based on the current global surface map. After selection of a target, the system aims the cameras at the target and performs local surface estimation 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. These steps are discussed below separately.

4.1. Visual target selection

This step determines how the scene is traversed as a surface map is accumulated. The basic question is this: What criteria should be used to select a point in the unmapped part of the visual field where surface information should be acquired next? This is reminiscent of 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, it will be useful to review some facts concerning human eye movements.

4.1.1. Psychological 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 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 many studies of eye movements. The purpose of these studies is typically to infer properties of higher-level cognitive activity which govern the movements. Very few studies deal with 3D domains, and these are usually concerned with ergonomics or vehicle operation. Here we list findings from various 2D psychological studies. in which subjects exhibited the following strong tendencies during the selection of new fixation points:

(a) Sequences of visual targets are selected in a centrifugal order, beginning at the departure point (initial fixation point). This implies that proximity to the original point of fixation is an important criterion [Lévy81].
(b) Upward eye movement is preferred over downward movement [Lévy81].
(c) Eye rotation either to left or right is preferred, depending on the person [Lévy81].
(d) For several potential targets in the visual field, those lying closer to the fovea are more likely to be selected for fixation [Find81]. This effect may depend partially on the change in resolution from fovea to periphery.
(e) When scanning random 2D polygonal forms, eye fixations tend to concentrate near vertices [Božk82].
(f) During examination of pictures, saccades are directed to

[1] Our domain of research is presently limited to the case of stationary scenes. For a moving visual field, a *smooth pursuit* step should be added to this sequence. This would complete the biological pattern known as "optokinetic nystagmus."

peripheral areas of "informative detail," [Mack67] which involves higher-level recognition of image objects (eg., features of human faces).
(g) When symmetry is present in 2D displays, subjects tend to concentrate fixations along the axes of symmetry [Loch87].
(h) When peripheral stimuli are presented suddenly, the resulting strong temporal cue often leads to a saccadic eye movement toward the target [Find81, Find83].

These findings are not necessarily true for all situations, but can provide a basis for general criteria to guide visual target selection.

4.1.2. A simple computational model for target selection. The psychological results summarized above suggest that the following factors are important in the selection of the next point for fixation: 1) *absolute distance and direction* [eg., (a-d)], 2) *2D image characteristics* [eg., (e-g)], and 3) *temporal changes* [eg., (h)]. We identify the following additional criteria 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) *object boundaries:* when the selected point lies across an object boundary, any known qualitative relative distance information should be used to select a new object for fixation; 6) *occlusion regions:* the system should not attempt to fixate the parts of the visual field which 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 objects to farther objects; 7) *compactness:* successive fixation points should be selected so as to grow a surface outwards from initial fixation point, since most objects yield compact regions in the images; and 8) *complexity:* the total number of fixation points should be minimized.

Our current model is a simple one, and incorporates only those criteria that involve surface geometry. We have not taken into account any criteria that require an analysis of image gray-level structure or that involve any temporal changes. Our model incorporates criteria (1, 4, 6-8), but excludes the more complex criteria (2, 3, 5) which will be included in a later version. Thus our current model stresses proximity: angular proximity to the original fixation point P_0 and the current fixation point P_i, and distance of the target from the cameras. The target P is chosen so that the following weighted average is minimized:

$$f(P) = k_1 R(P) + k_2 A(P, P_i) + k_2 A(P, P_0)$$

The function R gives the estimated distance from P to the camera platform. The value $A(Q_1, Q_2)$ represents the angular separation between two 3D points, relative to the current camera location. Candidate targets P must not be contained in the composite surface map, and must lie within camera travel limits.

The first term favors scene points that lie near the imaging apparatus (criterion 6). Since the range to unmapped scene points is not known, the value $R(P)$ must be estimated from depth and gradient information in the composite surface map (criterion 4). The second term biases the choice of target to scene points which 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 7). Criterion 8 is met by choosing the next fixation point so as to uncover as much as possible of the currently unknown part of the visual field. This also helps to minimize the total number of camera movements.

4.2. Surface estimation

This step concerns the process of orienting cameras so as to bring the projection of the desired point of fixation to the center of each image and estimating the surface in the vicinity of the point of fixation. Once again, it will be useful to review some results about biological vision before we present our computational model.

4.2.1. Physiological vergence. Most work in biological vision of interest here has been concerned with modeling eye vergence movements in response to changes in the visual field. Biological vergence movements are traditionally decomposed into four components [Scho83]: *disparity vergence*, which is affected directly by binocular disparities, is probably the dominant component; *accommodative vergence*, which takes into account the observed image blur to determine vergence movements; *tonic vergence*, due to the effects of muscular tonus; and *proximal vergence* which is due to psychological expectation. From a computational viewpoint, only the first two components are of interest here. Tonic vergence reflects the tendency of an alert individual, in the absence of visual stimulation, to move the eyes to a convergent resting state. The fourth component, proximal vergence, is based on the fact that humans typically expect a familiar object to be of a particular size. When a similar object is recognized, distance is inferred from the apparent size of the object and involuntary vergence movements tend to verge the eyes to that distance.

The objective of the disparity-vergence system is to attain fixation by reducing disparities near the image center to zero, thereby allowing binocular fusion to take place. Krishnan and Stark [Kris77] present a system model of the disparity-vergence component, which accepts a disparity signal as input, and produces vergence control signals as output. Their goal is to model the dynamics of biological vergence. Computer simulations demonstrate the agreement of the model with empirical physiological data. No interaction with accommodation is used in this model, and they do not discuss the means by which the disparity signal might be derived from a stereo pair of images.

Hung and Semmlow [Hung80] describe an analytical model which integrates accommodation and disparity vergence subsystems. Their purpose is again to develop a model which agrees with experimental physiological data. Image blur and disparity each serve as stimuli which drive the accommodation and vergence control signals. They also do not discuss the means by which the blur and disparity signals might be derived from a stereo pair of images.

Sperling [Sper70] presents a model for fixation, based on the interactions of vergence, accommodation, and binocular fusion. His is an "energy" model, in which each of these three visual information sources contributes a separate component based directly on the visual input. Both vergence components, disparity vergence and accommodative vergence, are incorporated into the model.

This model differs from others in that it considers fusion as a separate visual cue. Binocular fusion is a cortical phenomenon which depends on disparity. Physiologically, fusion is possible only when disparities are sufficiently small so that stereopsis can take place.

The model is formulated such that three independent variables representing vergence (v), accommodation (a), and fusion (u) are varied so that the following criterion function is minimized:

$$g(v, a, u) + \iint w_v \, |I_l - I_r| \, dxdy + \iint w_u \, |I_l - I_r| \, dxdy + \iint w_a \left[|\nabla^2 I_l|^2 - |\nabla^2 I_r|^2 \right] dxdy$$

The multipliers $w_i(x, y)$ represent appropriate weighting functions, and are negative where appropriate.

The first term encourages agreement among depth estimates obtained individually from the vergence, accommodation, and stereopsis processes. This is accomplished by a convex-upward ("bowl-shaped") energy function which assumes a minimum value when all three estimates are identical, and increases in value otherwise. The second term provides a measure of the goodness of vergence, and is discussed in Section 2.2. This term should be a minimum when both image centers are in registration. The function $w_v(x, y)$ gives greater weight to the image centers than to the periphery. Sperling assumes symmetric vergence, without loss of generality.

Sperling defines the third, or fusion, term identically to that of vergence, with a single difference: the summation is performed only over foveal areas for visual targets within Panum's fusional area. He proposes that the summation be performed only over the images of particular objects having sufficiently small disparities. The function w_u is intended to incorporate such target masking. The last term, for accommodation, is based on the degree of image blur. This is accomplished by using derivatives to estimate the high-frequency content in the images, as was discussed in Section 2.3. The multiplier $w_a(x, y)$ weights this term relative to the others, and typically weights the image center higher than the periphery. This term should be at a minimum when both images are in sharpest focus.

The variables v, a, and u can be taken to represent the state of the fixation system. While the intent is that these state variables are varied smoothly until a minimum is found in the above equation, Sperling specifies system dynamics based on a gravimetric analog so that the system can also come to rest at a *local* minimum. This means that the same external stimulus can cause different system states, depending on previous visual stimuli. This agrees with physiological phenomena; examples of multistability abound in everyday life. This can occur, for example, whenever we see a partially reflective surface, look through a wire grid, or encounter the "wallpaper illusion."

4.2.2. A computational model for surface estimation. The biological models discussed above are clearly not intended for surface estimation. Indeed, these models consider only point operations (such as disparity computation), rather than emerging surface characteristics. (This is similar to most work on computational stereo, wherein point disparities alone are used without consideration of surface characteristics, as discussed in Section 2.1.) 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.

Nevertheless, the modeling of interaction among different cues is extremely pertinent for the integration of information derived from these cues. The model of Sperling is perhaps the most comprehensive in this regard and could serve well as a basis for the specification of a computational system. Our model is quite similar to the Sperling model; the major difference is that surface estimates are retained and treated as a separate state variable. The goal is to minimize the following criterion function of vergence (v), left and right focus (a_L and a_R, respectively), and the surface s:

$$g(v, a_L, a_R, s) + m(I_l, I_r) + \iint w_s C(s) \, dxdy + \iint w_a \left[|\nabla I_l|^2 - |\nabla I_r|^2 \right] dxdy$$

The first term is again intended to ensure agreement among the different depth estimates at the point of fixation, and is defined as

$$g(v, a_L, a_R) = w_{LR} \, |R_a(a_L) - R_a(a_R)| + w_{LV} \, |R_a(a_L) - R_v(v)| + w_{LS} \, |R_a(a_L) - R_s(s)|$$

where R_i maps its argument to a depth value.

The second term measures similarity of the two image centers. This corresponds to Sperling's term for vergence, but we suggest the normalized cross-correlation measure for registration. The third term computes the goodness of stereo fusion in terms of the smoothness of the computed surface, rather than in terms of intensity differences among corresponding pixels related by a fixed disparity value. The operator C computes coarseness of the surface s. The fourth term, for focus, is similar to Sperling's accommodation term, except that we use the gradient norm instead of the Laplacian to measure image sharpness.

When the above criterion function is minimized for a newly selected target point, the result is final selection of a point of fixation, extraction of sharp images, and an estimation of a local surface patch derived from stereo analysis.

The independent variables may again be seen as representing the state of the surface-estimation system. Rather than specify system dynamics analytically, as is the case for the Sperling model, we propose the algorithmic selection of new states for the variables, as will be described in the next section. The motivation for this is occlusion avoidance. The

Sperling model, for example, specifies that the independent variables vary smoothly as the system seeks equilibrium. It is possible that a satisfactory equilibrium point does not exist, according to this formulation, when occluding objects are present in the visual field.

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.

5. INTEGRATION ALGORITHM

In this section we describe an algorithm for partially achieving the integration described in the previous two sections. Both components of the active surface-reconstruction paradigm, target selection and surface estimation, are described. Its implementation (Section 6) demonstrates dramatic improvements in estimated surfaces over those possible without such interaction.

Before proceeding further, we note some salient differences between the surface-estimation model proposed in Section 4.2.2 and the algorithm presented. First, the algorithm does not seek to minimize the proposed criterion function explicitly through slow, smooth changes of the independent variables, as was the case in the Sperling formulation. For our model, dynamics are specified implicitly in the surface-reconstruction algorithm so that the independent variables are systematically chosen, leading to a (possibly local) minimum of the criterion function in a stepwise fashion.

Also, while the algorithm integrates vergence and accommodation cues directly, it does not integrate the surface smoothness. For the given criterion function of surface-estimation (Section 4.2.2), this amounts to the omission of the variable s from the first term, and leaving out the third term entirely. Equivalently, the algorithm divides the step of integrated surface estimation into two separate steps: fixation and surface estimation (see algorithm below). These will be described separately below. In the future, we plan to extend this algorithm to integrate the fixation and surface-estimation steps to more properly represent the integration paradigm.

We now present the dynamic surface-reconstruction algorithm at a high level. The following paragraphs elaborate on the individual steps of the algorithm.

Step 1: Visual target selection.

The goal of this step is the selection of a new candidate scene point at which to attempt fixation, based on the current surface map and imaging parameters. As described earlier (Section 4.1.2), this algorithm utilizes proximity measures to choose a target which is optimum with respect to a criterion function.

In addition, this step examines the depth map for occlusion information, so that the system does not attempt to fixate a scene area which is known to be occluded. Before attempting to find the next point of fixation, scene areas of occlusion are found by projecting currently accumulated surface estimates onto the image planes of both cameras (Figure 3). If the projections of two surface patches onto the right (left) image are contiguous, while the projection of the same patches onto the left (right) image plane are not contiguous, then the intervening region of the left (right) image is assumed to be occluded. Points in these occluded scene areas are not selected as targets.

SURFACE-RECONSTRUCTION ALGORITHM

Repeat until entire scene has been mapped
1. Select target based on current global scene description
2. Fixate
    - 2.1 Aim both cameras at target
    - 2.2 Repeat until both image centers show the same scene area in focus
        - 2.2.1 Obtain range estimates using focus for both cameras
        - 2.2.2 If range estimates from focus and vergence do not agree, aim cameras at scene point indicated by one focus estimate
    - 2.3 Vary vergence to register image centers
3. Perform local surface estimation
    - 3.1 Derive a set of depth estimates for stereo
    - 3.2 Invoke stereo with initial depth estimates
    - 3.3 Merge the resulting surface patch with an evolving composite surface map

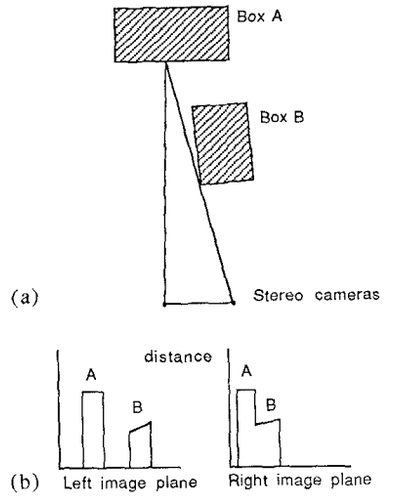

Figure 3: Occlusion detection from 3D surface estimates. The top view of a stereo imaging arrangement is shown in **(a)**. Because of the location and orientation of box B, its left surface is not visible to the right camera, and part of box A is hidden. To locate occlusion regions, the 3D surface description projected onto the left and right image planes. Projections of A and B onto the left image are disjoint **(b)**, whereas the projections of A and B onto the right image are contiguous. Such a condition indicates the existence of an occlusion.

Rather than evaluate the criterion function for potential targets directly, the system calculates the value of the function for points on the edge of the surface map since depth information is available only for points on the surface map. After a criterion minimum is found, the system selects a view orientation past the edge of the surface map, estimates depth for this scene area, and aims the cameras at that point. Typically, the resulting new camera orientations will extract new parts of surfaces that overlap partially with the current global surface map.

Step 2: Exploratory fixation.

The fixation process uses depth estimates from both focus and vergence. Each camera in turn uses focus to obtain a depth estimate for the scene point visible at the image center. From the knowledge of the current imaging geometry, the locations of the corresponding 3D points in the scene are determined for each camera. The method for obtaining depth estimates from focus was described in Section 2.3. If the depth estimates from focus and vergence differ substantially, the system assumes that an occlusion is present. The system must then take steps to rotate both cameras to one side of the edge discontinuity.

After all three range estimates are in agreement, a registration process is performed. The vergence angle is varied slightly in search of a maximum in a normalized cross-correlation criterion function (Section 2.2.2). Since this step results in a small camera motion which does not typically affect the image sharpness, the new camera configuration minimizes all terms of the criterion function except the third, and achieves fixation.

Step 3: Local surface estimation.

After obtaining an estimate of the distance to the point of fixation, images are extracted for use by stereo. The system then varies focus again to obtain a coarse grid of depth estimates to be used by the stereo process. These estimates need only to be accurate enough for stereo to receive appropriate search windows along the epipolar lines of the images.

The stereopsis process accepts the initial surface estimates and produces a surface patch about the point of fixation as described in Section 2.1. We used a modified version of the stereo algorithm reported in [Hoff87]. The derived surface patch will be for a part of the scene which is common to the visual fields of both the left and the right cameras.

The resulting surface patch is now merged with the cumulative surface map. When the current imaging geometry is known, this involves simple rigid rotations and translations of 3D surface patches to a home coordinate system. Since the newly obtained surface patches typically have partial overlap with previously mapped scene areas, the surface should smoothly extend beyond previously mapped parts.

6. IMPLEMENTATION AND RESULTS

First we describe briefly the UI imaging system, which permits automated stereo image acquisition under computer control. Next we present details and results of our algorithm as implemented on this system.

6.1. The UI Imaging System

Two high-resolution CCD video cameras, atop a stereo platform, are used to obtain image pairs. The orientation of the platform and the vergence angle of the cameras are controlled through the use of high-precision stepper-motor positioners. Four independent rotational units are used for tilt, pan, and vergence movements, yielding four degrees of freedom for camera orientation. Motorized zoom lenses are utilized, having focal lengths which range from 17.5 to 105 mm, yielding a 6× zoom ratio. The lens settings of zoom, focus, and aperture are driven by a DC-motor controller.

Camera orientations and lens settings are determined by the host workstation, a Sun Microsystems 3/160, which also controls image acquisition and performs most image processing. Images are digitized to 512×512 pixels of 8 bits each. Although the cameras can verge independently, we utilized only symmetric vergence, so that the point of fixation was always directly in front of the camera platform. This was done for ease of algorithm implementation, and should not cause loss of generality of the methods employed.

Currently, extensive calibration procedures have not been completed. To derive a mapping from focus settings to depth, a test object was placed at a sequence of known distances from one camera. The maximum of the focus criterion function was found for each distance, and least squares methods were used to fit a model based on the lens equation to these data. A few control points were used for initializing pan, tilt, and vergence angles. The stepper motors and position controller have proved to yield highly repeatable results. Stepper motor resolution ranges from 0.01° / step for vergence control to 0.001° / step for pan control.

6.2. Implementation Details

The present implementation comprises several independent software modules, which correspond to Steps 1 through 3 of the surface-reconstruction algorithm given in the previous section. The system can run autonomously, or the user can invoke individual modules directly (as was done for the results presented here).

In our current implementation, the stereo module requires parallel focal axes. After fixation, therefore, the camera vergence angle is reduced to zero before extracting final stereo images. (The current stereo module is being modified to accept nonzero vergence angles, so this will not be necessary in the future.) We shall continue to refer to the prior point of fixation as such, even though the cameras are now actually fixated at infinity.

6.3. Experimental Results

We now describe a single run of the algorithm, using the scene depicted in Figure 4. The vertical axes of the cameras are assumed to be aligned, so that their optical axes are always coplanar. The focal axes are initially parallel.

Figure 5 shows the initial stereo image pair of this scene, consisting of a vertical barrel next to a rectangular box, both resting on a flat table top and in front of a rear wall. The two images overlap only to a small extent because of the large baseline used (0.28 m). Notice that the scene point which projects onto the center of the left image is occluded from the view of the right camera. This starting point was chosen for

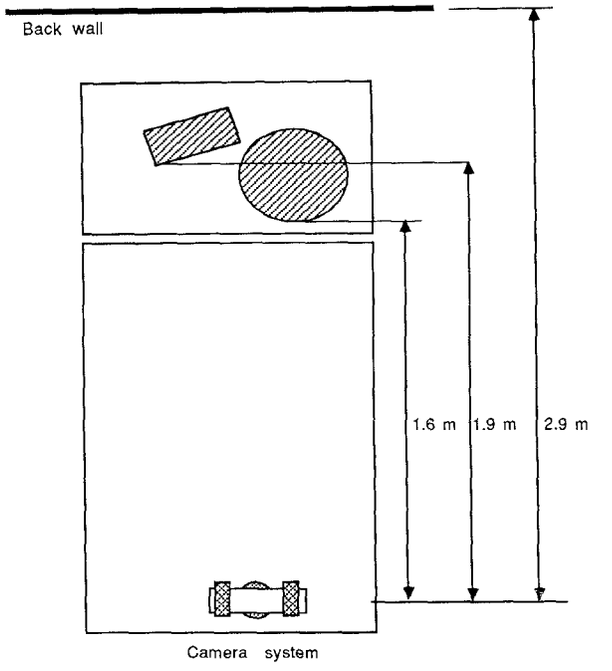

Figure 4: Top view of imaging environment (not to scale). The stereo camera platform appears at the lower portion of the figure. Scene objects and a wall are near the top of the figure. The dimensions are shown for a world-coordinate system with origin at the center of the platform.

illustrative purposes, so that a description of the fixation process about an occluding surface could be described initially.

After system initialization, the fixation process begins. Both lenses are zoomed to the maximum focal length, and the left camera computes a depth estimate along its optical axis. Among the possible focus settings, a search is made for the setting which causes a maximum in the criterion function for a small (48×48) window in the image center. Several levels of search are used, beginning with a very coarse search in which a few focus settings are equally spaced over the entire focus range, and narrowing the search space until acceptable depth accuracy is reached.

After the depth estimate is obtained, the system attempts to fixate that scene point by panning the camera platform and causing the cameras to verge (rotate) inward. This leads to the right camera's view of the desired scene point being obstructed by the barrel; by using focus changes to estimate the range along the optical axis of the right camera, the system detects this situation. The 3D location of the occluded point is estimated from the left camera's depth estimate and from knowledge of the imaging geometry, and is stored for future use.

The system reacts by attempting to fixate the nearer point, i.e., the scene point corresponding to the center of the right image. 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. Because the depth estimates are very close, and agree with the depth calculated from the known vergence angle, the system performs a registration process to increase the accuracy of the vergence angle. As the vergence angle is changed by small amounts, small windows at the image centers are compared using a normalized cross-correlation of those windows. The system then selects the best vergence setting, ant the resulting fixated images, still at maximum zoom, are shown in Figure 6.

The system then causes the lenses to return to their shortest focal lengths, with the resulting images shown in Figure 7. The system is now ready to present these images to the stereo module. But since our current stereo module expects parallel focal axes, the camera vergence angle is reduced to zero before extracting final stereo images.

To get an initial surface estimate for the stereo algorithm, a 10×10 grid of coarse depth estimates is derived from focus for both the left and right images, as described in algorithm Step 3. After the grid is derived, the algorithm continues:

- Find the grid location corresponding to the point of fixation.
- Find the largest rectangular set of grid points which contain this fixation point, and which have the same depth estimate.
- Mask out all grid points outside this window, to serve as an indication that no stereo correspondences are to be considered for this portion of the image.
- Replace all grid points within this rectangle with the more accurate depth estimate obtained during the process of fixation.

The stereo module is then invoked, with these depth points as initial estimates for corresponding scene regions. The resulting depth map, referenced to the coordinate system of the left camera, is shown in Figure 8.[2]

The next step is to select a new fixation point and appropriate camera orientations, so that the scene description can be extended. As described earlier, this module examines the border of the surface map, and selects an edge point which is optimum with respect to target-selection criterion function. For this case the optimum edge point lies along the left edge of the surface map, corresponding to the left edge of the barrel. The system then makes the assumption that this surface extends smoothly to the left. It selects a pan angle designed to permit overlap of new surface patches with the current surface map (approximately 1°), and extrapolates a depth value for that viewing direction based on nearby points in the depth map. This pan angle and depth value determine the new 3D target, which the system will now attempt to fixate.

The cycle now proceeds as described in Section 5. The system attempts to fixate the new scene point, based on depth estimates from focus and vergence. As before, the left camera is aimed at a scene point not visible at the right camera. But the system compares this scene location with its internal surface map, and detects the occlusion before focusing the

[2] Although the stereopsis algorithm produces actual 3D depth points as output, for display purposes we have shown a map of the *disparities*, which are reciprocally related to depth. This is to demonstrate that curved surfaces have actually been extracted; when the actual depth values were plotted, the resulting surface appeared flat because the total difference in depth for the barrel (approximately 7 cm) was a very small percentage of the depth range of the scene.

right camera. The system rotates again to the left, and no occlusions are detected this time. A scene point located on the rear wall is fixated, and stereo images are extracted. The rough grid of initial depth estimates this time provides a valid estimate only for a small region about the fixation point. The resulting incremental depth map is smaller, and is shown merged with the previous depth map in Figure 9.

The algorithm proceeds in this fashion, incrementally building up the global surface map from small patches derived by the stereo module. After six separate fixations, the resulting surface appears as shown in Figure 10. The predicted depths agree favorably with known object distances.

Finally, we highlight the importance of integration as advocated in this paper by comparing the results of this algorithm with output from the same stereo module when no effort is made to provide accurate initial depth estimates. Instead of providing depth estimates based on other cues, a frontal surface was supplied as the initial estimate to the stereo algorithm for the same scene in Figure 7. The initial depth estimate was a frontal surface located at 2.9 m from the cameras. As can be seen from the resulting surface map (Figure 11), reliable disparity estimates are produced only for scene locations near this depth.

7. SUMMARY

We have argued that to obtain surface maps of large scenes having large depth ranges, individual depth cues such as stereo disparity, camera vergence, and focus are not sufficient by themselves; rather they should be used in a tightly integrated mode where they complement each other's strengths to define a more powerful and complete mechanism for surface estimation. We have also discussed the importance of dynamic selection of camera orientations based on the evolving scene description. This amounts to integrating image *acquisition* with image *analysis* for surface estimation. We have demonstrated through experimental results that integration leads to significant improvements in the quality of surface maps over those obtained from individual cues.

8. ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation under grant IRI-86-05400. Many thanks to Subhodev Das for improving and tailoring the stereo module used with the UI imaging system.

REFERENCES

[1] D. Arnold, "Local context in matching edges for stereo vision," in *Proceedings: DARPA Image Understanding Workshop*, Cambridge, pp. 65-72, May 1978.

[2] N. Ayache and B. Paverjon, "Fast Stereo Matching of Edge Segments using Prediction and Verification of Hypotheses," *Proceedings: Computer Vision and Pattern Recognition*, pp. 662-664, June, 1985.

[3] H. H. Baker and T. O. Binford, "Depth from Edges and Intensity Based Stereo," in *Proceedings: International Joint Conference on Artificial Intelligence*, Vancouver, pp. 631-636, Aug. 1981.

[4] S. T. Barnard and W. B. Thompson, "Disparity Analysis of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-2, no. 2, pp. 333-340, July, 1980.

[5] D. I. Barnea and H. F. Silverman, "A Class of Algorithms for Fast Digital Image Registration," *IEEE Transactions on Computers*, vol. C-21, no. 2, pp. 179-186, February, 1972.

[6] V. Božkov, Z. Bohdanecký, and T. Radil-Weiss, "Perception, Exploration and Eye Displacements," in *Cognition and Eye Movements*, ed., P. Fraisse. North-Holland, pp. 24-33, 1982.

[7] C. J. Erkelens and H. Collewijn, "Eye Movements and Stereopsis during Dichoptic Viewing of Moving Random-Dot Stereograms," *Vision Research*, vol. 25, no. 11, pp. 1689-1700, 1985.

[8] J. M. Findlay, "Local and Global Influences on Saccadic Eye Movements," in *Eye Movements: Cognition and Visual Perception*, ed., J. W. Senders. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 171-179, 1981.

[9] J. M. Findlay, "Visual Processing for Saccadic Eye Movements," in *Spatially Oriented Behavior*, ed., M. Jeannerod. New York: Springer-Verlag, pp. 281-303, 1983.

[10] D. Geiger and A. Yuille, "Stereopsis and Eye-Movement," *Proceedings: First International Conference on Computer Vision*, pp. 306-314, June, 1987.

[11] W. E. L. Grimson, "A computational theory of visual surface interpolation," Report No. 613, Artificial Intelligence Lab, MIT, June 1981.

[12] R. L. Henderson, W. J. Miller, and C. B. Grosch, "Automatic stereo reconstruction of man-made targets," *Soc. P.I.E.*, vol. 186, no. 6, pp. 240-248, 1979.

[13] W. Hoff and N. Ahuja, "Surfaces from Stereo," *Proceedings: DARPA Image Understanding Workshop*, pp. 98-106, Dec. 1985.

[14] W. Hoff and N. Ahuja, "Extracting Surfaces from Stereo Images: An Integrated Approach," *Proceedings: First International Conference on Computer Vision*, pp. 284-294, June 1987.

[15] B. K. P. Horn, "Focusing," Report No. 160, Artificial Intelligence Lab, MIT, 1968.

[16] G. K. Hung and J. L. Semmlow, "Static Behavior of Accommodation and Vergence: Computer Simulation of an Interactive Dual-Feedback System," *IEEE Transactions on Biomedical Engineering*, vol. BME-27, no. 8, Aug., 1980.

[17] N. H. Kim and A. C. Bovik, "A Solution to the Stereo Correspondence Problem Using Disparity Smoothness Constraint," in *Proceedings: IEEE Conference of Systems, Man, and Cybernetics*, Atlanta, GA, Oct. 1986.

[18] V. V. Krishnan and L. Stark, "A Heuristic Model for the Human Vergence Movement System," *IEEE Transactions on Biomedical Engineering*, vol. BME-24, no. 1, January, 1977.

[19] E. P. Krotkov, J. Summers, and F. Fuma, "Computing Range with an Active Camera System," *Proceedings: Eighth International Conference on Pattern Recognition*, pp. 1156-1158, Oct. 1986.

[20] E. P. Krotkov, "Focusing," Report No. MS-CIS-86-22, GRASP Laboratory, University of Pennsylvania, 1986.

[21] E. P. Krotkov, "Exploratory Visual Sensing for Determing Spatial Layout with an Agile Stereo Camera System," Report No. MS-CIS-87-29, GRASP Laboratory, University of Pennsylvania, 1987.

[22] A. Lévy-Schoen, "Flexible and/or Rigid Control of Oculomotor Scanning Behavior," in *Eye Movements: Cognition and Visual Perception*, ed., J. W. Senders. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 299-314, 1981.

[23] G. Ligthart and F. C. A. Groen, "A Comparison of Different Autofocus Algorithms," *Proceedings of the Sixth International Conference on Pattern Recognition*, pp. 597-600, Oct. 1982.

[24] H. S. Lim and T. O. Binford, "Stereo Correspondence: a Hierarchical Approach," *Proceedings: DARPA Image Understanding Workshop*, 1987.

[25] P. J. Locher and C. F. Nodine, "Symmetry Catches the Eye," in *Eye Movements: From Physiology to Cognition*, ed., A. Lévy-Schoen. North-Holland, pp. 353-361, 1987.

[26] N. H. Mackworth and A. J. Morandi, "The Gaze Selects Informative Details within Picture," *Perception and Psychophysics*, vol. 2, pp. 547-552, 1967.

[27] D. Marr and T. Poggio, "A Computational Theory of Human Stereo Vision," *Proceedings of the Royal Society of London*, vol. 204, pp. 301-328, 1979.

[28] D. Marr, *Vision*. Freeman, 1982.

[29] G. Medioni and R. Nevatia, "Segment-Based Stereo Matching," *Computer Vision, Graphics, and Image Processing*, vol. 31, pp. 2-18, July 1985.

[30] Y. Ohta and T. Kanade, "Stereo by Intra- and Inter-Scanline Search Using Dynamic Programming," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-7, pp. 139-154, March 1985.

[31] C. M. Schor and L. B. Ciuffreda, *Vergence Eye Movements: Basic and Clinical Aspects*. Boston: Butterworths, 1983.

[32] G. Sperling, "Binocular Vision: A Physical and a Neural Theory," *American Journal of Psychology*, vol. 83, pp. 461-534, 1970.

[33] M. Subbarao, "Direct Recovery of Depth-map II: A New Robust Approach," Report No. 87-03, Department of Electrical Engineering, State University of New York, Stony Brook, 1987.

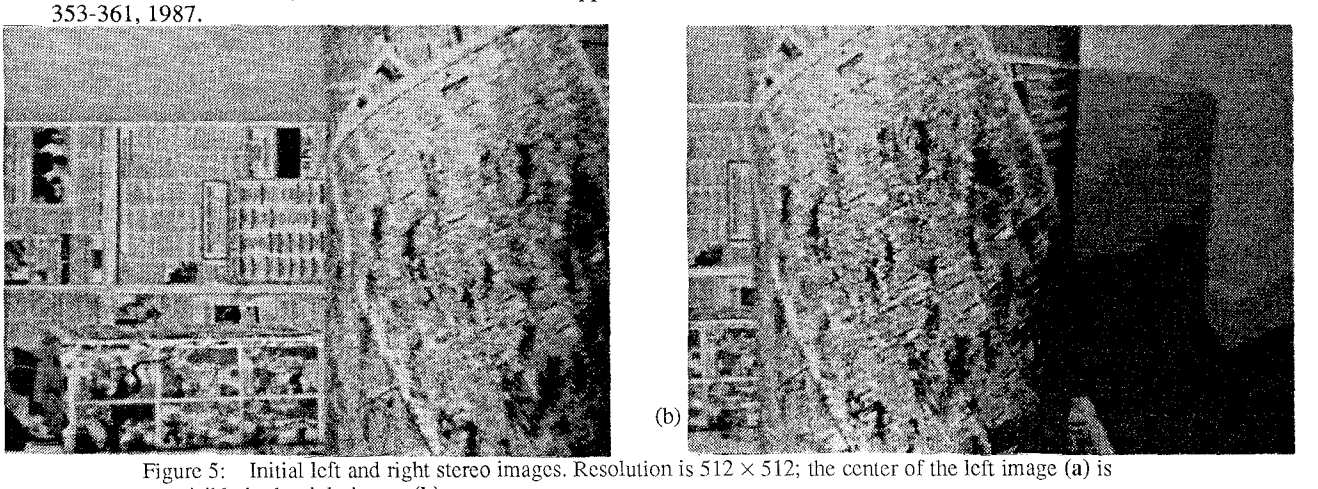

Figure 5: Initial left and right stereo images. Resolution is 512×512; the center of the left image (**a**) is not visible in the right image (**b**).

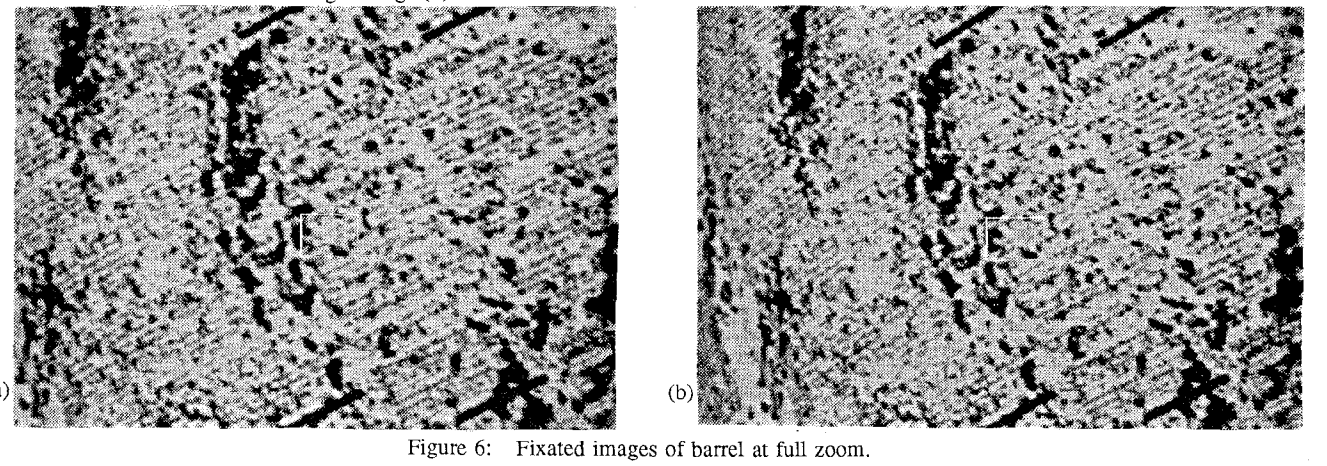

Figure 6: Fixated images of barrel at full zoom.

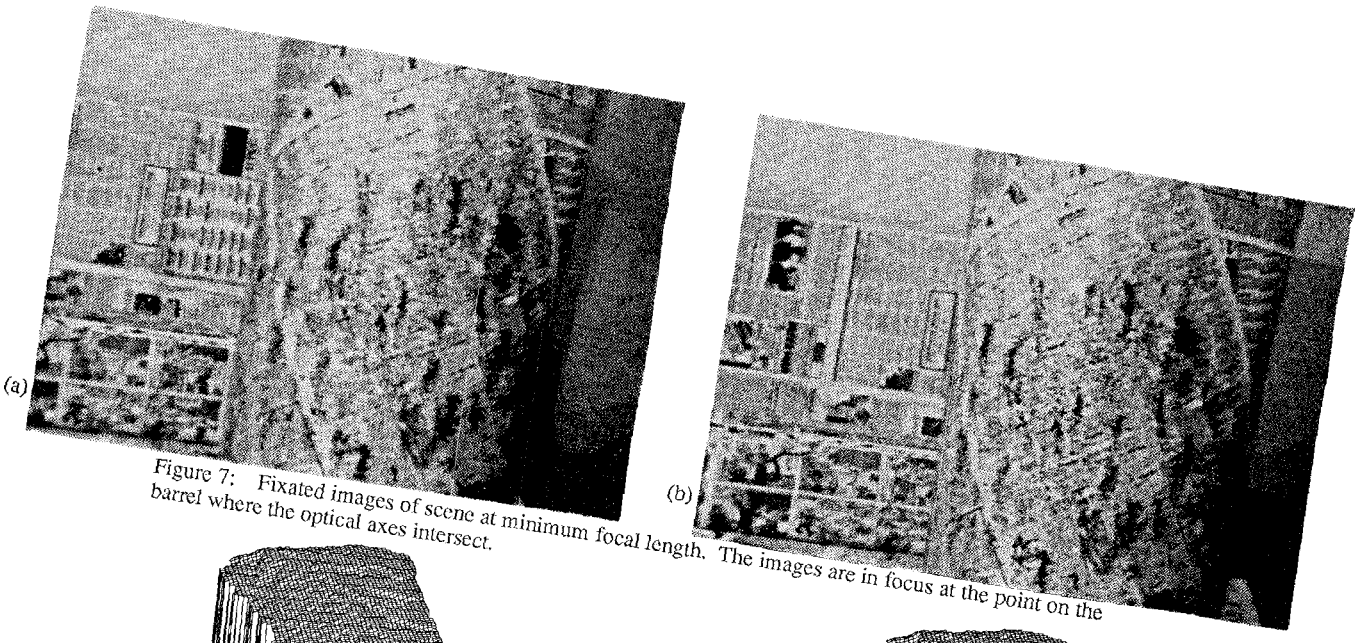

Figure 7: Fixated images of scene at minimum focal length. The images are in focus at the point on the barrel where the optical axes intersect.

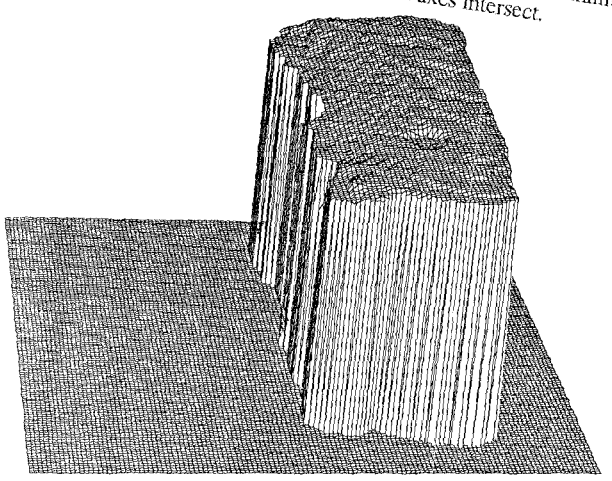

Figure 8: Initial surface estimated by the stereo algorithm for this scene.

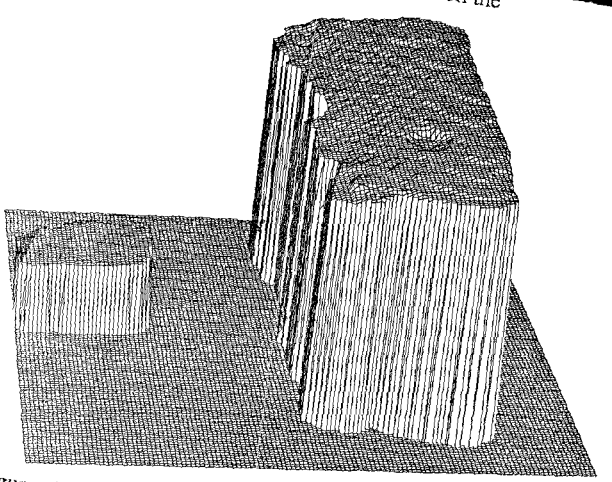

Figure 9: Intermediate surface description. As new surface estimates are obtained, they are transformed a home coordinate system and then merged. The result is an evolving, global scene description.

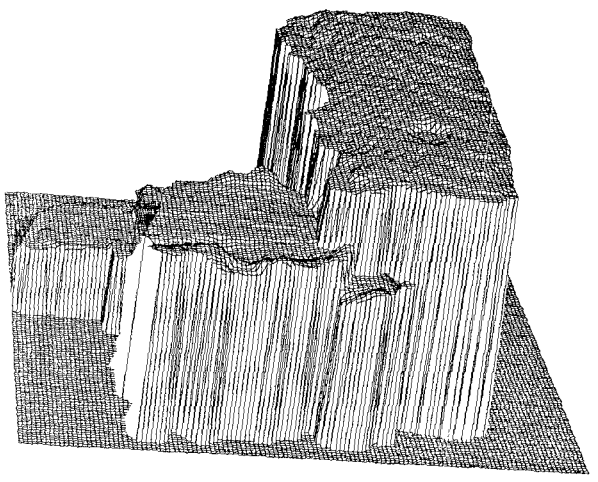

Figure 10: Intermediate surface description for extended visual field. This is the depth map after 6 fixations.

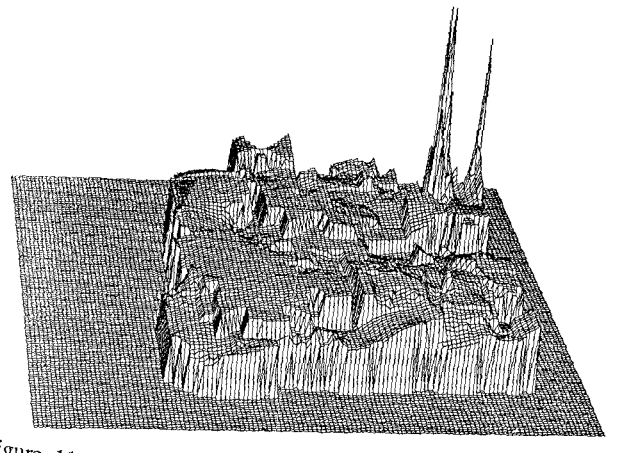

Figure 11: Surface estimates produced by stereo without valid initial estimate. This scene description was produced for the scene of Figure 7, but with an initial estimate of a frontal surface at a constant of 2.9 m over the entire scene. Only surfaces at approximately that depth were accurately estimated.