Integrating Multiresolution Image Acquisition and Coarse-to-fine Surface Reconstruction from Stereo

Subhodev Das Narendra Ahuja

Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
1101 W. Springfield Ave.
Urbana, Illinois 61801

Abstract

This paper is concerned with the problem of surface reconstruction from stereo images for large scenes having large depth ranges, where it is necessary to aim cameras in different directions and to fixate at different objects. This paper presents an approach to acquiring coarse structural information about the scene in the vicinity of the next fixation point during the current fixation, and utilizing this information for surface reconstruction in the vicinity of the next fixation point. The approach involves processing of peripheral, low resolution parts of the current images away from the image center, in addition to accurate surface estimation from the central, high resolution parts containing the fixated object. The processing of the low resolution parts yields coarse surface estimates to be refined after the cameras have refixated, and the parts of the scene around the new fixation point (currently at low resolution) are imaged more sharply. The coarse estimates are obtained from both stereo and focus. The choice as to which estimate is actually used depends on which one is determined to be more accurate in the given situation. Thus, the approach presented also involves dynamic integration of the use of stereo and focus as sources of depth information, in addition to integrating multiresolution image acquisition and their coarse-to-fine processing.

1 INTRODUCTION

This paper is concerned with the problem of surface reconstruction from stereo images for large scenes having large depth ranges. At any given time during imaging of such scenes, sharp images can be acquired only for narrow parts of the visual field, capturing a limited depth range. Abbott and Ahuja have argued [1] that to reconstruct surfaces of such scenes, it is necessary to integrate the use of camera focus, camera vergence, and stereo disparity. Surface estimation must be performed over a scene in a piecewise fashion, and the local surface estimates must be combined to build a global description. In [1], an algorithm was outlined to achieve such integration through iteration of the following three steps: visual target selection, fixation, and stereo reconstruction. The scope of that algorithm is limited to the reconstruction of the surface of a single object, i.e., the successive fixations scan the surface of the same object. When the entire surface of the fixated object has been scanned, the acquired surface map does not smoothly extend, and therefore surface reconstruction must be resumed by fixating on a new object. Since surface reconstruction from stereo requires coarse initial estimates, such estimates must be obtained for the new object before reconstruction can continue.

The support of the National Science Foundation under grant ECS 8352408 and Rockwell International is gratefully acknowledged.

We present an approach to acquiring coarse structural information about the scene in the vicinity of the next fixation point during the current fixation, and utilizing this information for surface reconstruction in the vicinity of the next fixation point from stereo images. The approach involves processing of peripheral, low resolution parts of the current images away from the image center, in addition to accurate surface estimation from the central, high resolution parts containing the fixated object. The processing of the low resolution parts yields coarse surface estimates to be refined after the cameras have refixated, and the parts of the scene around the new fixation point (currently at low resolution) are imaged more sharply. The coarse estimates are obtained from both stereo and focus. The choice as to which estimate is actually used depends on which one is determined to be more accurate in the given situation. Thus, the approach presented also involves dynamic integration of the use of stereo and focus as sources of depth information, in addition to integrating multiresolution image acquisition and their coarse-to-fine processing.

Section 2 describes in greater detail the background and motivation behind the work reported in this paper. Section 3 discusses stereo and focus as independent sources of depth information, and the relative accuracies of the depth estimates derived from them; knowing their relative performance is necessary for selecting one over the other. Section 4 presents an algorithm that performs the desired integration of coarse-to-fine image acquisition and stereo analysis, wherein the initial surface estimates required for stereo reconstruction of the newly fixated object are obtained dynamically from either focus or stereo, and for the parts of the scene not occluded from either viewpoint. Section 5 gives details of implementation and the experimental results, and Section 6 presents concluding remarks.

2 BACKGROUND AND MOTIVATION

In this section we present the past research related to the work reported in this paper, and the motivations that lead to the development of the approach described in the following sections. The various steps in our algorithm (Section 4) are directly related to specific motivations discussed below (Section 2.2).

2.1 Background

This paper pursues the basic theme of active, intelligent data acquisition which is well described by Bajcsy [3,4]. In their analysis of surface reconstruction from stereo images, Marr and Poggio also point out the role of eye movements in providing large relative image shifts for matching stereo images having large dis-

CH2813-4/89/0000/0009/$01.00 1989 IEEE

parities, thus implying the need for active data acquisition. Ballard and Ozcandarli [5,6] point out 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. Aloimonos et al [2] show that active control of imaging parameters leads to simpler formulations of many vision problems that are not well behaved in passive vision. Geiger and Yuille [11] describe a framework for using small vergence changes to help disambiguate stereo correspondences. Other efforts have concentrated on modeling biological mechanisms of interactions among vergence, accommodation and stereopsis. Erkelens and Collewijn [10] discuss interactions between vergence and stereo for biological systems. Sperling [17] presents a model for the interaction of vergence, accommodation (focus), and binocular fusion in human vision. However, there has been only limited use made of the different cues in developing detailed computational approaches and implementations for surface estimation from stereo images [14], especially in a mutually cooperative mode such as discussed in [3,13,17].

2.2 Motivation

Consider the initial state in which one of the objects in a scene is fixated on and the surface reconstruction of the scene begins. The cameras successively fixate on different parts of the same object. The surface patches obtained during the different fixations are combined to obtain a composite surface. During each one of these fixations the stereo images are acquired using a focal length that yields the sharpest image of the object in the vicinity of the fixation point. Any parts of the scene that may lie within the visual field of a camera but are outside the depth of field appear out of focus, with the degree of blur determined by the distance from the fixation point. Therefore, different parts of the scene adjacent to the object boundary, possibly imaged during different fixations, may appear blurred to different degrees depending on the extent of the depth discontinuity. The degree of defocus of these parts can be used to yield surface reconstructions which would not be as accurate as obtained for those parts within the depth of field, but which could serve as coarse estimates of the surfaces in these parts. Similarly, the out of focus images could be stereo analyzed to derive surface estimates which would also be inaccurate due to poor localization of features. The availability of such coarse maps would make it possible to select a new fixation point on a new object, thus bringing into sharp focus the new object currently out of focus. Further, the coarse surface map available for the new object could serve as the initial estimate for more accurate surface estimation from stereo pairs of higher resolution images. Once the cameras are fixated at the newly selected object, the resolution of the rest of the objects lying in the direction of the selected object also improves. Therefore, as the finest stereo reconstruction is achieved for the selected object, the accuracy of the surface information available for those other objects which are now closer to the fixation point also improves.

A stereo pair acquired at any time in general contains parts having different resolutions. The object under fixation has the maximum resolution, and will be said to be in the *central visual field*; the other objects have their image resolutions graded by the magnitudes of their depths relative to the point of fixation (and not by their locations in the image as in human vision), and will be said to be in the *peripheral visual field*. Thus, the acquisition of successive images result in a temporal interleaving of coarse-to-fine resolution sequences of the different objects, in which the peripheral objects gradually move in to occupy the central visual field. If the objects in the central field can be segmented out in each image, then they can be stereo analyzed using, as the initial surface estimate, the best available surface estimate from a previous fixation. Thus, our use of the terms "central" and "peripheral" is different from the use of the same terms in human vision where they refer to the presence of graded resolution - the highest resolution being in the foveal region, decreasing towards the periphery.

Interestingly, the availability of the coarse depth map for the unmapped parts of the scene has advantages other than the ability to select a new fixation point. First, the computational blurring operation is replaced by instantaneous optical blurring. In addition to the speed advantage, this may lead to more realistic coarse images than obtained by artificial blurring used in the coarse-to-fine stereo algorithms. Second, the image blurring operation is integrated with the (mechanical) reconfiguration of the cameras, and thus with image acquisition. A range of images of increasing resolution can be acquired while the cameras verge and focus on the new fixation point. Third, similar to blurring, stereo analysis can be performed in parallel with camera reconfiguration. This enables the inherently serial, coarse-to-fine analysis of stereo pairs [12] to be performed in parallel with image acquisition, i.e., the stereo algorithm can be initiated on a coarse stereo pair while the imaging parameters are being changed to acquire the finer resolution images. The number of stereo pairs acquired before fixation is achieved would depend on the amount of camera reconfiguration required; the larger the amount of camera reconfiguration, the greater would be the opportunity to acquire intermediate resolution images.

While selecting an object for fixation in the peripheral field, one strategy may be to use a near-to-far scan, i.e., the closest object is fixated on and stereo analyzed first followed by the next closest object. There are two computational advantages of using such a scan. One, by reconstructing the surfaces of the near objects first, the occluded portions of the farther objects can be identified. Thus, knowing those parts of the scene which are occluded from at least one viewpoint would avoid selection of such points for fixation. The second advantage in starting with the near objects is that doing so maximally exploits the focus cue which is computationally simpler but more effective for short range objects. This helps simplify the earliest, no-information stage out of which the surface mapping process must bootstrap itself.

The accuracies of the stereo reconstructions for the different regions (having different resolutions) in the peripheral visual field would be different owing to the different degrees of localization error caused by the variable amount of blurring. The error in the stereo based estimates may be computed in terms of feature location errors. When these errors are larger than the focus based depth estimates, the latter estimates may be used in place of the stereo based estimates. This defines an automatic and dynamic means of chosing between stereo and focus as the sources of coarse surface information. In the next section, we will discuss stereo and focus as independent sources of surface information and their relative performance.

3 STEREO AND FOCUS AS SOURCES OF DEPTH

The binocular cue of stereo disparity and the monocular cue of focus have long been recognized as important sources of 3D

information. In this section we will be discussing how each of them is used and what its limitations are.

3.1 Stereo

Stereo vision concerns the recovery of three dimensional depth from two or more different viewpoints. In this section we will discuss the binocular stereo, assuming that the two images have distinguishing features visible to both viewpoints and the viewpoints are reasonably well separated.

Given an object point $P = (X, Y, Z)$, let (x_l, y_l) and (x_r, y_r) denote the coordinates of the perspective projection of P in the left and right images, respectively. Also, let f_l (f_r) be the distance of the left (right) image plane from the corresponding lens center O_l (O_r). If $f_l = f_r = f$, the left and the right image planes are coplanar and there is no vertical displacement between the lens centers, then the object distance $Z = Z_l = Z_r$ can be expressed as

$$Z = f\frac{b - (y_l - y_r)}{y_l - y_r} \tag{1}$$

where the quantity $y_l - y_r$ is the *disparity* in left and right image locations of P along the Y direction and b (called *baseline*) is the distance between the two cameras. To obtain depth requires selecting feature points (mostly detected intensity edges) from the two images, matching points belonging to one image with those from the other, and using the disparity in image locations of a matched pair of points to compute the corresponding 3D coordinates from (1). The search in the right image for the match of a given point in the left image can be restricted to a unique line called the *epipolar line*, determined by the camera geometry.

3.1.1 Accuracy of range estimation from stereo

The accuracy of the computed 3D positions is limited by the errors in camera geometry, detected feature locations and matching. We will now discuss the former two types of error which are relatively tractable.

The relationship between the errors in stereo camera geometry and the estimated range has been widely investigated [16,18]. Because of the discrete nature of the image plane coordinates, point projections can be expressed in terms of pixels only. This quantization affects the coordinate values by a maximum of $\pm 1/2$ pixel. The disparity, $d = y_l - y_r$, can therefore be in error by as much as ± 1 pixel. Let $\Delta d = \hat{d} - d$ be the error in disparity, where $\hat{d}$ is the quantized disparity. Substituting d for $y_l - y_r$ in (1) the error in estimated range can be expressed as

$$\begin{aligned}\Delta z &= \hat{z} - z = f\frac{b - \hat{d}}{\hat{d}} - z = f\frac{b - d - \Delta d}{d + \Delta d} - z \\ &= f\frac{b - \frac{fb}{z+f} - \Delta d}{\frac{fb}{z+f} + \Delta d} - z = -\frac{(z+f)^2\Delta d}{fb + (z+f)\Delta d}.\end{aligned} \tag{2}$$

The edges detected by most edge detectors are often a distorted version of the true edges. An edge detector, also used in our work, is the Laplacian of the Gaussian operator ($\nabla^2 G$) [15]. The features are the zero-crossings in the convolution of the image with this operator. It has been shown that the locations of the zero-crossings do not coincide with the true edges for non-linear illumination and non-ideal (non-step and finite length) edges [7]. When edges are matched, this location error leads to disparity error.

If Δd_{tot} denotes the total error due to quantization (Δd_{qnt}) and feature localization (Δd_{loc}) then the total normalized error in the estimated range can be written as

$$|\Delta z/z| = \frac{1}{z}\frac{(z+f)^2\Delta d_{tot}}{fb + (z+f)\Delta d_{tot}} \tag{3}$$

where $\Delta d_{tot} = \Delta d_{qnt} + \Delta d_{loc}$.

3.2 Focus

To estimate depth from focus, the distance between the lens center and the image plane of a camera system is varied to register a sharp image of an object point. The distance yielding the sharpest image depends on the depth of the object point and can be used to estimate the depth. Focus is an attractive source of depth because it does not require a solution to the feature correspondence problem.

The lens equation relating the distance u of an object point on the optic axis of a perfect lens to the distance v of the corresponding image point is

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f}, \tag{4}$$

where f is the focal length of the lens. For such a point, the object distance can be computed using the above equation if v and f are known. The distance v is changed by adjusting the focus setting (denoted by p). This directly affects the degree of image blur for objects in the field of view. The distance v is linearly related to p. Let p_{max} and p_{min} denote the focus settings that correspond to the maximum (infinity) and minimum (u_{min}) object distances, respectively, at which the lens can be focused. The lens equation can then be alternately written as

$$u = \frac{u_{min}(p_{max} - p_{min}) - fp}{(p_{max} - p_{min}) - p}. \tag{5}$$

The parameters p_{max} and p_{min} are functions of the focal length of the lens.

3.2.1 Accuracy of range estimation from focus

The optimum focus setting is usually identified by some criterion function that measures the high-frequency content and assumes maximum value when the image is in sharpest focus. It would be desirable that the peak of the measure function be sharp, repeatable under different imaging conditions and yield the true focus setting. The peak may be poorly localized for several reasons. A change in illumination level or sensor noise can cause a shift in the location of the peak. Changes in image magnification that accompany focus adjustment may lead to multiple peaks. Poor localization leads to inaccuracies in depth estimates. A lack of image detail can cause the peak to be nearly flat. Shorter focal lengths, smaller apertures, and greater object distances can also cause the peak to be flatter. The flatness of the peak is measured by the *depth of field* of the lens.

The defocused image of a point is a circle (called the *blur circle*). The extent of defocusing is proportional to the radius of this circle. Let u_0 and u be the depths of the object point at which the lens of aperture A is focused and of a distant point, respectively. Then the diameter of the blur circle of the distant point is

$$D = \frac{Af}{u_0 - f}(1 - \frac{u_0}{u}). \tag{6}$$

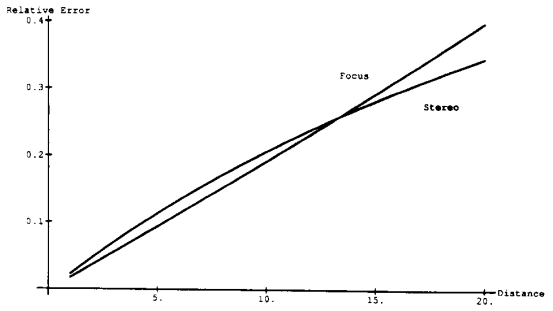


Figure 1: Range Estimation errors due to Stereo and Focus showing the crossover point. Distances are in meters.

Let us assume that for any image the differences in sharpness cannot be distinguished for blur circles with diameter smaller than C, referred to as the *circles of confusion*. For a given location of the image plane $s = v_0$, corresponding to object distance u_0, the projections of all objects located in the interval $[u_2, u_1]$ appear equally sharp. This interval is the depth of field. The uncertainty in the estimated object depth can be expressed as

$$\Delta z = u_2 - u_1 = \frac{2ACu_0f(u_0 - f)}{A^2f^2 - C^2(u_0 - f)^2}. \quad (7)$$

3.3 Relative Merits of Stereo and Focus as Depth Cues

Stereo disparity and focus both serve as important sources for range determination. From our earlier discussions we have seen that each of these depth cues has its own weaknesses. It is interesting to study the relative performance of these functionally dissimilar sources of depth under the same imaging conditions. To do this, let us assume that stereo is free from matching errors. Since estimates from both sources degrade with increasing distance, a basis for comparison may be the reliability of the depth estimates as a function of the distance. One can examine (7) and see that $\Delta z = u_2 - u_1 \to \infty$ as $z = u_0 \to \frac{Af}{C} + f$, whereas $|\Delta z| \to \infty$ as $z \to \infty$ from (3). Similarly, as $z \to 0$, the focus based error $\Delta z = u_2 - u_1 \to 0$ whereas the stereo based error $|\Delta z| \to \frac{f\Delta d_{tot}}{b + \Delta d_{tot}}$. Thus, the error in focus based range estimate is lower for close range objects whereas stereo gives lower error for objects at large distances. There must exist a value of object range z for which both focus and stereo are equally reliable estimators. The exact location of this *crossover point* depends upon various imaging parameters.

Using the values $C = 0.0342$ mm, $A = 33.9$ mm, $f_{stereo} = 61$ mm, $f_{focus} = 105$ mm, $b = 5$ cm, $\Delta d_{qnt} = 1$ pixel (1 pixel = 0.0115 mm), $\Delta d_{loc} = 6$ pixels, equations (3) and (7) are plotted in Figure 1. The plot shows the crossover point to the right of which the error due to the depth of field overwhelms the error in stereo.

4 ALGORITHM

In this section we describe an algorithm to achieve the desired integration of multiresolution image acquisition and their coarse-to-fine processing, and depth estimation from focus and stereo described in Sections 1 and 2. Having achieved fixation on an object, the algorithm consists of the following steps: (1) the coarse estimate of the depth of the current fixation point available from the surface map from the fixation on a previous object (except for the first fixation) is used to derive the best focus setting that yields a more accurate focus based depth estimate, (2) stereo images are acquired and segmented into focused and defocused regions, (3) it is determined which of stereo and focus based coarse surface estimates is more accurate, which is then used as the initial surface estimate for stereo analysis. Surface reconstruction for the in-focus and out-of-focus parts is performed separately. These steps of the algorithm are discussed in the following subsections.

4.1 Depth Estimation from Focus

The cameras fixate at different parts of a scene and extracts surfaces over a limited visual field and depth range during any single fixation [1]. The focus setting of the lens is used to estimate the depth of the fixation point. Focusing is attempted using the largest available focal length that ensures small depth of field, hence more accurate depth estimates. The focal axis setting (image plane distance) is varied and a criterion function, defined as the total squared gradient over a window (called the *window of fixation*) around the image center, is evaluated. The setting, p, corresponding to the maximum of the criterion function is used to estimate the depth at the image center from (5). The search for the best focus setting is slow because of the large range of settings that needs to be explored. In our algorithm, the selection of the next fixation point is always based on the available surface information. So, except for the very first fixation for which the search extends over the entire range of focus settings, an approximate depth estimate of the fixation point is available from previous fixations, limiting the search to the vicinity of the focus setting that corresponds to this depth estimate.

4.2 Image Acquisition and Segmentation

Stereo images are acquired with a focal length smaller than the one used for estimating depth from focus to increase the field of view. The fixation point is in focus in these images. Additionally, two images are obtained for each viewpoint in which the fixation point is out-of-focus. One of these images is focused nearer than the original scene point, and the other is focused farther such that the corresponding depths of field are adjacent to the depth of field of the fixation point. The parts of the scene that are in sharp focus (corresponding to objects that lie within the depth of field of the scene) are identified by comparing the sharpness of the "in-focus" stereo images to the "out-of-focus" images and segmenting the in-focus images [9]. The in-focus regions of the image constitute the central field of view and the defocused regions comprise the peripheral field. Two surfaces separated by a depth larger than the depth of field corresponding to the focus setting of the in-focus image cannot be included in the segmented image simultaneously. A depth discontinuity contour separating these surfaces becomes a part of the boundary of a segmented region.

4.3 Initial Estimate from Focus or Stereo

Focus estimates depth for the window of fixation only. The accuracy of this estimate is related to the depth of field. At every

fixation except for the very first one, an approximate depth estimate is also available in the vicinity of the fixation point from surface maps obtained during previous fixations. The estimate from focus is used as the initial estimate for stereo analysis as long as the depth is smaller than the crossover depth. In such situations, stereo reconstruction takes place only for the finest level of image resolution using a small value of σ (σ_{min}) for the $\nabla^2 G$ operator. This σ gives the best trade-off between localization and stability of the detected zero-crossings. When the estimated depth is greater than the crossover depth, there are two cases that need to be considered. If the current fixation point is within the depth of field of the previous fixation point, stereo estimate is accurate and is used as the initial estimate. When the current fixation point is outside the depth of field of the previous fixation, the stereo estimate is inaccurate due to defocusing of the local features; the accuracies of the focus and stereo based estimates need to be compared. The distance between the previous and current fixation points is used to determine the amount of defocusing due to optical blurring from (6). Assuming that the point spread function of the lens is modeled by a 2D Gaussian, the radius of the Gaussian kernel $\sigma_l = kD$ ($k > 0$ and is a characteristic of the lens system) expresses the extent of optical blurring of the current fixation window during previous fixation. If $\nabla^2 G$ with $\sigma = \sigma_f$ was used to detect features in the periphery, then the Gaussian expressing these two blurring effects has a kernel of $\sigma = \sigma_t = \sqrt{{\sigma_l}^2 + {\sigma_f}^2}$. This $\sigma = \sigma_t$ is now used to compute Δd_{loc} and hence the stereo inaccuracy, Δz, from (3). The coarse stereo error is compared with the focus based error. If stereo is found to be more accurate then it is used as the initial estimate and for determining σ_{max}, else focus is used.

In the absence of any focus based depth estimates for the peripheral regions, stereo reconstruction has to progress across multiple levels. The peripheral features undergo optical blurring. Matches for the features in the unmapped parts of the periphery are located by searching over large image regions. It is therefore desirable that the number of matchable features be fairly small. Since a focal length significantly smaller than the full zoom is being used in our current algorithm to acquire the stereo images, the optical blurring which is a function of the focal length (refer to (6)) cannot guarantee significant smoothing of the various parts of the periphery to meet the desired objective. Hence, large σ's are used for the feature detector to introduce additional smoothing. As the fixation point moves from near to far objects, a distant peripheral object in the direction of fixation becomes sharper. This reduces the σ_l associated with the defocusing of the object, and hence σ_t of the effective Gaussian. Finer images of the peripheral object are automatically obtained when the fixation point changes. Correspondingly, the surface map of this object is refined as the fixation point moves closer to the object. In other words, the generation of multiresolution surface maps for the periphery occurs in parallel with image acquisition.

Stereo reconstruction is attempted only for those parts of the scene that are not occluded from either viewpoint as estimated from the available surface estimates. Details of this step are omitted for brevity. Once the features have been detected in both the images, features from the central and peripheral fields are separated and matched to yield 3D points. These points are clustered and surface smoothness constraint is enforced to remove the false ones [9]. Quadratic surface patches are fit to the clusters of points. Finally, range values are interpolated from these smooth surfaces and used to update the range map. Surface reconstruction results in accurate (*fine*) surfaces for the central regions and relatively inaccurate (*coarse*) maps for the periphery. These maps are then merged with those from the previous fixations to build the composite surface map.

5 IMPLEMENTATION AND RESULTS

In this section we present details and results of implementing the range sensing algorithm on a dynamic imaging system. The system consists of two Cohu 4815 CCD cameras mounted on a stereo platform and equipped with Vicon V17.5-105M motorized zoom lenses. High-precision stepper-motor rotational units are used to control independent pan, tilt and vergence angles. The system can also translate horizontally with one degree of freedom.

5.1 Implementation Details

A focal length of 61 mm and a baseline of 5 cm are used to acquire the stereo images. A calibration process determined the distance of the left image plane from the corresponding lens center, f_l, to be 47.5 mm at this focal length. The optic axes are parallel when stereo images are registered. The calibrated parameters p_{max} and p_{min} of (5) for a focal length of 105 mm using the left camera are $p_{max} = 15378$ and $p_{min} = -379$. The same parameters for a focal length of 61 mm are 10847 and -142, respectively. The manufacturer's specification for u_{min} is 1.3 m. The circle of confusion is experimentally determined to be 6 pixels of the CCD imaging array or 102.6 μm. The relation between the diameter D of the blur circle and the spread parameter σ of the Gaussian associated with lens defocusing is experimentally found to be $\sigma = kD + \sigma_0$. The parameters k and σ_0 are calibrated to be 0.1 and 1.1 for $f = 61$ mm, and 0.2 and 0.9 for $f = 105$ mm, respectively.

5.2 Experimental Results

The dynamic camera system was made to scan an indoor 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. A focal length of 105 mm (full zoom) is used for the fixation process. The fixated point must be visible to both cameras. Figure 2 shows the system fixated at a point close to the left edge of the barrel. The range here from stereo maps of earlier fixations is $z = 1.56$ m (the measured distance is 1.54 m). All distances are relative to a reference point on the camera unless mentioned otherwise. The focus ranging process is required to scan the axis settings between $p = 1438$ and 2645 corresponding to this coarse stereo estimate, instead of the entire range between 0 and 15000, to locate the optimal setting. The setting which maximizes the criterion function within the window of fixation is found to be $p = 2023$. The distance of the fixation point and the near and far extremities of the depth of field from the image plane are 1.646, 1.59 and 1.707 m, respectively. The focal axis setting corresponding to this object distance is $p = 685$ at $f = 61$ mm which is used to acquire the stereo images of Figure 3. Only one defocused image is shown for each viewpoint (Figure 4). Figure 5 shows the partial map for the current left viewpoint that was obtained during previous fixations. Figure 6 has the in-focus regions of the left and the right images following segmentation. Most of these regions belong to the barrel as expected, except for few spurious patches. The box and the rear wall are the peripheral objects. The nearly accurate range from focus is used to run a single level of stereo reconstruction. The features within

the central field are extracted using a $\nabla^2 G$ of $\sigma = \sigma_{min} = 6$ and matched to yield the left range map (Figure 7) for the 512×512 level. (A similar map may be derived for the right viewpoint also.)

The coarse-to-fine stereo reconstruction in the peripheral region proceeds at multiple levels - 64, 128, 256 and 512. The depth in the periphery ranges from 1.707 m, the distant extremity of the depth of field, to our chosen maximum depth of 20 m. Features are extracted using a $\nabla^2 G$ of σ = 8, 9, 10, and 12 for the levels 64, 128, 256 and 512, respectively. The final left range map at the completion of the current fixation cycle is shown in Figure 8. During the combination of different maps, fine range values override coarse values.

Several fixations later, the cameras are aimed at the back wall. The depth at the fixation point from the coarse peripheral maps is 2.93 m. The range of focal axis setting that needs to be investigated is between p = 8332 and 9282. The criterion function for the window of fixation is maximized at a setting of p = 8570 which implies z = 2.81 m. The distance of the fixation point is 3.06 m, as measured. The stereo images of Figure 9 are registered with a focal axis setting of p = 6654. The stereo images are segmented into central and peripheral regions. The box and the barrel now belong to the near periphery which is reconstructed first. The updated range maps are used to detect occlusions in the central field comprised of the wall.

The range from focus is used as the initial estimate for the wall and one level of reconstruction is done. The left range map of the wall is shown in Figure 10. Since the coarse estimate from stereo was found to be closer to the measured distance than the focus estimate, as a comparison, multiresolution reconstruction using the coarse stereo as the initial estimate is done next. The zero-crossings are detected with $\nabla^2 G$ of $\sigma = \sigma_{min}$ so that the localization errors do not exceed the error due to the depth of field. The reconstructed maps at the completion of 512×512 level are shown in Figure 11.

6 SUMMARY

In this paper we have presented an approach to reconstruction of surfaces from stereo for large scenes having large depth ranges. We have argued that surface extraction and camera reconfiguration should be integrated to explore scenes and build global surface maps from local patches. We have also discussed the importance of using the depth cue of focus together with stereo and how their cooperative interaction can be exploited to improve the surface estimation process. In particular, our method offers a way to select focus or stereo dynamically to give coarse estimate of a surface that can be used to initiate coarse-to-fine stereo analysis. Accurate surface maps are obtained for those parts of the scene which are in perfect focus. Relatively inaccurate maps are derived for the defocused regions. These maps help the dynamic imaging system to select targets in a piecewise near-to-far order thereby discovering occlusion of parts of a distant surface by a nearer surface. Multiresolution surface reconstruction is achieved in parallel with image acquisition. Experimental results demonstrate that the approach presented may be well suited for active data acquisition and surface reconstruction by an autonomous stereo system.

References

[1] A. L. Abbott and N. Ahuja. Surface reconstruction by dynamic integration of focus, camera vergence, and stereo. In *Proc. Second Intl. Conf. on Computer Vision*, pages 532–543, Tarpon Springs, FL, December 1988.

[2] J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. In *Proc. First Intl. Conf. on Computer Vision*, pages 35–54, London, UK, June 1987.

[3] R. Bajcsy. Active perception vs. passive perception. In *Proc. Workshop on Computer Vision*, pages 55–59, Bellaire, MI, October 1985.

[4] R. Bajcsy. Perception with feedback. In *Proc. DARPA Image Understanding Workshop*, pages 279–288, Cambridge, MA, April 1988.

[5] D. H. Ballard. Eye movements and spatial cognition. Technical Report 218, University of Rochester, November 1987.

[6] D. H. Ballard and A. Ozcandarli. Eye fixation and early vision: Kinetic depth. In *Proc. Second Intl. Conf. on Computer Vision*, pages 524–531, Tarpon Springs, FL, December 1988.

[7] V. Berzins. Accuracy of laplacian edge detectors. *Computer Vision, Graphics, and Image Processing*, 27:195–210, 1984.

[8] S. Das, A. L. Abbott, and N. Ahuja. Surface reconstruction from focus and stereo. In *Proc. 5th Intl. Conf. on Image Analysis and Processing*, Positano, Italy, September 1989.

[9] C. J. Erkelens and H. Collewijn. Eye movements and stereopsis during dichoptic viewing of moving random-dot stereograms. *Vision Research*, 25:1689–1700, 1985.

[10] D. Geiger and A. Yuille. Stereopsis and eye-movement. In *Proc. First Intl. Conf. on Computer Vision*, pages 306–314, London, UK, June 1987.

[11] W. Hoff and N. Ahuja. Surfaces from stereo: Integrating feature matching, disparity estimation, and contour detection. *IEEE Trans. Pattern Anal. Machine Intell.*, PAMI-11:121–136, February 1989.

[12] E. P. Krotkov. Exploratory visual sensing for determining spatial layout with an agile stereo camera system. Ph.D. Thesis MS-CIS-87-29, GRASP Laboratory, University of Pennsylvania, Philadelphia, PA, 1987.

[13] E. P. Krotkov, J. Summers, and F. Fuma. Computing range with an active camera system. In *Proc. Eighth Intl. Conf. on Pattern Recognition*, pages 1156–1158, October 1986.

[14] D. Marr and E. Hildreth. Theory of edge detection. In *the Royal Soc. of London, vol. B 207*, pages 187–217, 1980.

[15] F. Solina. Errors in stereo due to quantization. Technical Report MS-CIS-85-34, Dept. of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, 1985.

[16] G. Sperling. Binocular vision: A physical and a neural theory. *American Journal of Psychology*, 83:461–534, 1970.

[17] A. Verri and V. Torre. Absolute depth estimate in stereopsis. *Journal Opt. Soc. America*, 3:297–299, March 1986.

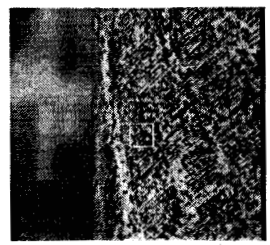
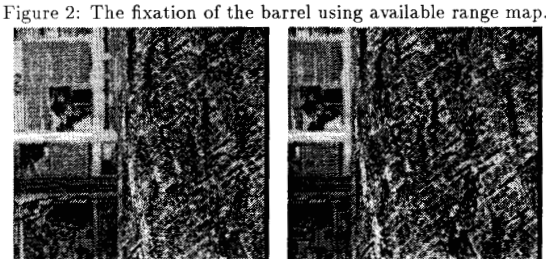
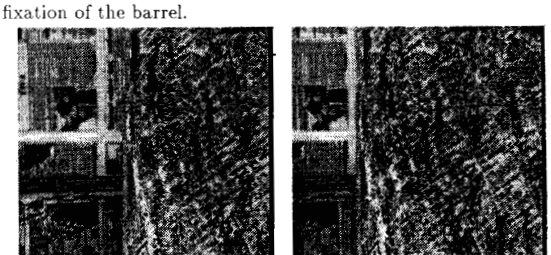


Figure 2: The fixation of the barrel using available range map.



(a) (b)

Figure 3: Stereo image pair: (a) left, (b) right, acquired after fixation of the barrel.



(a) (b)

Figure 4: The (a) left and (b) right images with cameras focused beyond the farthest point of the depth of field.

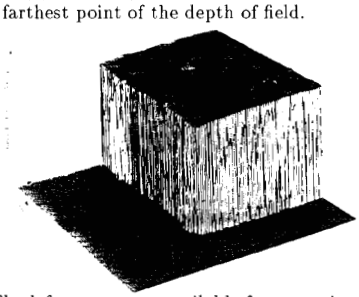
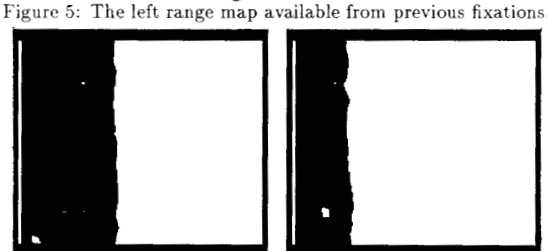


Figure 5: The left range map available from previous fixations.



(a) (b)

Figure 6: The central regions (white pixels) after segmenting the (a) left and (b) right images.

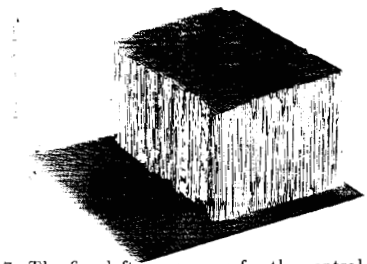


Figure 7: The fine left range map for the central field.

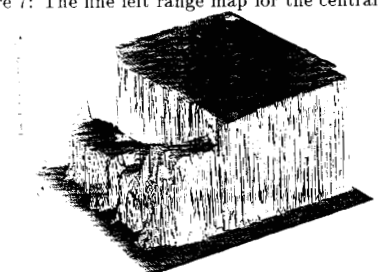
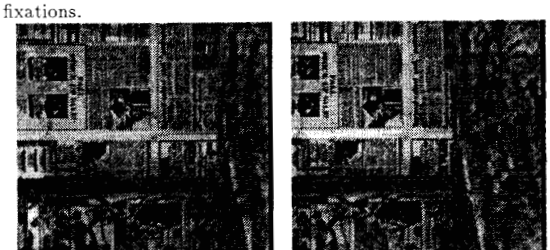


Figure 8: The combination of coarse peripheral and fine central left range maps that also includes information from previous fixations.



(a) (b)

Figure 9: Stereo image pair: (a) left, (b) right, acquired after fixation of the back wall.

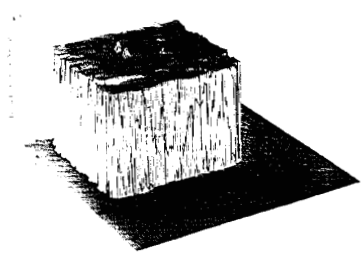


Figure 10: The fine left range map for the central field using focus based estimate.

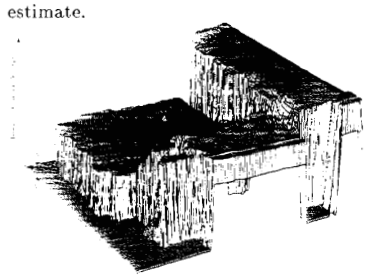


Figure 11: The composite left range map using coarse stereo estimate for the wall.