ELSEVIER

Artificial Intelligence 83 (1996) 241-266

Artificial Intelligence

Active surface estimation: integrating coarse-to-fine image acquisition and estimation from multiple cues *

Subhodev Das *, Narendra Ahuja

Coordinated Science Laboratory and Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Received June 1992; revised February 1995

Abstract

This paper is concerned with the problem of surface reconstruction from stereo images for large scenes having large depth ranges with depth discontinuities. The passive stereo paradigm is inadequate for this problem because of the need to aim cameras in different directions and to fixate at different objects. We present an active approach that involves the following steps. First, a new fixation point is selected from among the nonfixated, low-resolution scene parts of current fixation. Second, a reconfiguration of the cameras is initiated for refixation. As reconfiguration progresses, the images of the new fixation point gradually deblur and the accuracy of the position estimate of the point improves allowing the cameras to be aimed at it with increasing precision. In the third step, the improved depth estimate is used to select focus settings of the cameras, thus completing fixation. Finally, stereo images are acquired and segmented into fixated and nonfixated parts of the scene that are analyzed in parallel.

1. Introduction

Many algorithms have been proposed in early computational vision to reconstruct visible three-dimensional (3-D) surfaces. Such surfaces serve as representation of the scene being viewed and are useful to a number of machine vision applications. Now, a single probe of the environment may not be sufficient to collect enough data for reconstructing an entire scene that is both wide and deep (i.e., most real-world scenes).

* The support of the Defense Advanced Research Projects and National Science Foundation under grant IRI-89-02728 is gratefully acknowledged.

* Corresponding author. Current address: PVI, 47 Hulfish St, Princeton, NJ 08542, USA.

SSDI 0004-3702(95)00036-4

This limitation is posed by the visual sensors which are sensitive to a finite range of environmental characteristics, such as possessing a limited field of view or depth of field. Besides, even within a sensor's range of operation, a fine tuning of its parameters may be necessary to obtain the most accurate data set, such as sharply focused features. Both of these factors motivate the adoption of an *active* strategy to adaptively control the sensors and to analyze the data. Much contemporary research in machine vision has involved analysis of passively sampled data acquired using a single visual cue, such as stereo, motion, focus, or vergence. On the other hand, an active system that deploys and controls *passive* sensors and utilizes multiple visual cues is at the heart of the research reported in this paper.

To analyze a typical real-world scene, cameras have to be directed at different parts of the scene in a sequential manner and images have to be acquired. The important issues related to this problem are *how* to process the data and *where* to look next. At any given time during imaging of a scene, sharp images can be acquired only for narrow parts of the visual field, capturing limited depth range. The acquired images are stereo analyzed to obtain high-resolution, accurate surface maps for the sharply-imaged parts of the scene. Simultaneously, low-resolution, inaccurate surface maps are obtained for the rest of the scene which may then be used to direct movement of the cameras to new, unmapped portions of the scene. The global surface map of the scene is synthesized from partial high-resolution maps extracted from the individual fixations. Elsewhere, we have presented an approach to the reconstruction of a single contiguous surface, where the cues of focus, vergence, and stereo are integrated to define an active, self-calibrating system [1]. The current approach essentially addresses the continuation of surface reconstruction across depth discontinuities. Overall, the following processes take place in parallel: single-surface estimation by integrating focus, vergence, stereo, and calibration; and integration of coarse-to-fine image acquisition and coarse-to-fine surface reconstruction. It is interesting to note that active control of cameras which is primarily motivated by the need to process large scenes actually provides additional opportunities and advantages, such as coarse-to-fine computation, fusion of depth estimates from multiple sources.

In the next section, we summarize the past research related to the work reported in this paper. Section 3 describes in greater detail the motivation behind the work reported in this paper; it is argued that an integration of the processes of camera reconfiguration and surface reconstruction helps to obtain improved structural information at a minimal cost of the mechanical movements and the computational steps. Section 4 presents an algorithm that achieves this desired integration to derive accurate descriptions for the fixated surfaces in the scene. Section 5 gives details of the implementation and the experimental results. Section 6 sketches a proof of the convergence (termination) of the surface estimation algorithm. Section 7 presents concluding remarks.

2. Past research

The bulk of past research in computer vision has considered the use of individual cues for deriving 3-D information. 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, especially in a mutually cooperative mode such as discussed in [3,18,24]. Ahuja and Abbott [1] have argued that to reconstruct surfaces for large scenes having large depth ranges, 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. A coarse estimate of the surface under reconstruction from focus was used by stereo to predict matches for the surface features, generate depth points from the matches, and interpolate a more accurate surface to these depth points [17]. The scope of that algorithm is limited to the reconstruction of a single continuous surface, e.g., one 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. Other efforts have concentrated on modeling biological mechanisms of interactions among vergence, accommodation, and stereopsis. Erkelens and Collewijn [15] discuss interactions between vergence and stereo for biological systems. Sperling [24] presents a model for the interaction of vergence, accommodation (focus), and binocular fusion in human vision.

This research pursues the basic theme of active, intelligent data acquisition which has been facilitated in recent years with the availability of sophisticated hardware for controlling imaging elements [6,9,11,18]. For example, incorporation of anthropomorphic features, such as spatially varying sensors [4,25], can reduce irrelevant sensory information. In their analysis of surface reconstruction from stereo images, Marr and Poggio [19] point out the role of eye movements in providing large relative image shifts for matching stereo images having large disparities, thus implying the need for active data acquisition. 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. Ballard and Ozcandarli [7] also 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. Geiger and Yuille [16] describe a framework for using small vergence changes to help disambiguate stereo correspondences.

3. Integration of camera control and surface estimation

In this section, we first discuss the need for and consequences of "saccadic" changes in fixation points, which is the primary motivation and emphasis of this paper. In particular, we observe that the *necessity* of changing fixation points across surface boundaries actually provides opportunities of integrating computational processes and depth information from multiple sources, leading to more efficient algorithms and more robust estimates (Section 3.1). Then, we present a specific analytical formulation of such integration which is used to define the approach pursued in the rest of the paper.

3.1. The context and motivation for integration

Consider the initial state of the active stereo system in which the system is fixated on one among multiple objects in a scene 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 description. 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 (which we call the *central visual field*). Any parts of the scene that may lie within the visual field of a camera but are outside the depth of field (which we call the *peripheral visual field*) appear out-of-focus, with the degree of blur determined by the distance from the fixation point. Stereo analysis of the out-of-focus peripheral image regions would result in surface estimates which would be of lower resolution and inaccurate due to poor localization of features. A detailed analysis of the performance of stereo, focus, and vergence in estimating depth is provided in [14]. Here, we utilize the results of that analysis.

Our use of the terms "central" and "peripheral" is suggested by the use of the same terms in human vision in which they refer to the presence of graded resolution. The terms "central" and "peripheral" have a slightly different meaning in our adaptation. The loss of resolution in our case is due only to the depth of field and not to the location of an object within the image. Thus objects within the depth of field, irrespective of their locations, will appear sharp, and in our use of the term, in the "central" visual field. The exact location of the "peripheral" visual field with respect to the point of fixation will depend upon where the scene depth ceases to be in the current depth of field, and will thus be determined by the scene.

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, selected from the periphery of the current visual field. This presents a dilemma since the exact locations and shapes of "new objects" are unknown (otherwise there would be no need for fixation and subsequent surface reconstruction). Fortunately, the availability of coarse peripheral maps would make it possible to select a new fixation point on a new object [13].

The selection of a new fixation point initiates the "saccadic" movements necessary to *foveate* the new object currently out-of-focus. While moving from one fixation point to the next, the mechanical reconfiguration of the sensor planes (e.g., a CCD array) to register a sharp image of the new fixation point, i.e., to focus the fixation point, is not instantaneous. Intermediate images are obtained with decreasing blur which may be continuously stereo analyzed to improve the estimate for the new point and the accuracy with which the two optic axes can be intersected at the fixation point using the vergence process. In this homing process, 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 those obtained by the artificial blurring used in coarse-to-fine stereo algorithms. Further, the image blurring operation is integrated with the reconfiguration of the cameras, and thus with image acquisition. A succession of images of increasing resolution can be acquired while the cameras verge and focus on the new

fixation point. This enables the inherently serial, coarse-to-fine analysis of stereo pairs [17] 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 sensor plane reconfiguration required; the larger the amount of camera reconfiguration, the greater would be the opportunity to acquire intermediate resolution images. At the completion of sensor plane reposition, a focus criterion function can be invoked to minimize the image blur. This completes the interaction among the three spatial cues of stereo, focus, and vergence, during the process of foveation. 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 precision of the surface information available for those other objects which are now closer to the fixation point also improves.

The image sequence obtained during foveation corresponds to the temporal interleaving of coarse-to-fine images of the individual objects, in which the peripheral objects gradually move in to occupy the central visual field. Just as in human vision the fovea has the highest neural connectivity per receptor, the central visual field that requires the most computation per pixel has to have a judicious allocation of computational resources to ensure the efficient processing of sensor data. In the case of a nonstationary visual target or sensory platform, incorporation of gaze stabilization may be necessary to keep the active vision system continuously fixated on the target [12].

3.2. *Analytical formulation for integration*

This section presents an analytical formulation for the intuitive scheme for the integration of camera adjustments and surface estimation described above. An assumption we use in this formulation is that the surfaces are *piecewise* smooth and that they have sufficient image detail. To ensure that there are no depth boundaries in the region under analysis, images are partitioned into central and peripheral visual fields. Two surfaces from the opposite sides of a depth discontinuity are almost never in-focus simultaneously. Thus, stereo analysis of all the contiguous feature points from the surface in-focus (belonging to the central visual field) avoids surface interpolation across depth discontinuities. The assumption about image details or intensity discontinuities is particularly important because an edge-based stereo, like ours, operates on these discontinuities; the focus and vergence cues require a high signal-to-noise ratio for reliable operation, a ratio which depends on image details.

A part of the visual field that has not yet been fixated but has appeared as the peripheral visual field during a fixation, will provide coarse (inaccurate) structural information. The stereo-based depth estimate of the peripheral target point is inaccurate due to the optical blurring of the peripheral features in the vicinity of the target point during the current fixation. Assuming that the lens blurring function is a Gaussian, the extent of this optical blur may be described by the spread parameter σ_L of the Gaussian for a lens of aperture A and focal length f [14]:

$$\sigma_L = kD = k\frac{A}{v_0}|v - v_0| = k\frac{Af}{Z_0}\left(\frac{|Z - Z_0|}{Z - f}\right), \tag{1}$$

where D is the diameter of the blur circle, the defocused image of the new target point, and k is a constant of proportionality; Z_0 and Z are the distances of the current fixation and the new target points, respectively, from the projection center of the lens measured along the optic axis and v_0 and v are the corresponding sensor plane positions. In addition, a Laplacian of Gaussian ($\nabla^2 G$) having a spread parameter $\sigma = \sigma_G$ is used to detect these features leading to further location errors of the detected features. The Gaussian expressing the optical and computational blurring effects at the given peripheral point has a spread parameter of $\sigma_t = \sqrt{\sigma^2{}_L + \sigma^2{}_G}$ and a quantized blur parameter of $d_t = 2\sigma_t$ pixels. The uncertainty in stereo depth, Z, of the new target point on the peripheral object may be expressed as [14],

$$U_S(Z) = Pd_t^2 + Q\alpha^2, \tag{2}$$

where α is the precision of the angular positioners of the verging camera system. The variables P and Q are functions of the camera system parameters and Z and determine the relative importance of d_t and α, i.e., feature localization and mechanical configuration, in influencing the uncertainty. By bringing the new target point into focus, σ_L and hence $U_S(Z)$ are reduced.

The fixation of a target point, during which the point has to be brought into focus and the optic axes of the two cameras must intersect at the point, is initiated by changing the direction of the optic axis of each camera. The latter requires a change in the camera pan as well as the tilt angle. Let the change in the pan angle be $\Delta\varphi = \kappa_p \Delta\phi_p$ and that in the tilt angle be $\Delta\tau = \kappa_t \Delta\phi_t$, where ϕ_p and ϕ_t are the angular positions of the pan and tilt actuators, and κ_p and κ_t are the corresponding constants of proportionality. A change in the direction of the optic axis is followed by a change in the left and right vergence angles denoted by $\Delta\theta_L = \kappa_v \Delta\phi_{vL}$ and $\Delta\theta_R = \kappa_v \Delta\phi_{vR}$. The large angular movements of the camera help to bring the image of the target point in the vicinity of the two image centers. According to Eq. (1), a reduction of σ_L requires bringing v_0 closer to v. The intermediate images, obtained during this sensor plane reconfiguration at intervals of $\Delta\sigma_L$, may be stereo analyzed to give increasingly accurate range estimate of the new target point. A change, Δv, in the sensor plane position (or focus setting) is related to a change in the corresponding actuator position, ϕ_f, as $\Delta v = \kappa_f \Delta\phi_f$. Suppose that the entire range of σ_L is partitioned into m intervals of length $\Delta\sigma_L$ each. Let v_{0i} ($v_{01} = v_0$) be the estimated sensor plane position for the target point at the beginning of the ith interval and $\Delta v_i = |v_{0i} - v_{0\overline{i+1}}|$. The total movement of the sensor plane positioner for each camera over the m intervals is

$$\sum_{i=1}^{m} \Delta\phi_{fi} = \frac{\Delta\sigma_L}{\kappa_f kAv} \sum_{i=1}^{m} v_{0i} v_{0\overline{i+1}}. \tag{3}$$

Finally, after the estimated optical blur has been eliminated by reconfiguring the sensor plane, precise focusing needs to be done.

Each step of mechanical readjustments of the cameras is followed by processing of data for intermediate surface synthesis. The increasingly focused stereo images can be

of different resolution; for example, the images that are less blurred and therefore have higher spatial frequency contents can be finely sampled. If $H_i \times H_i$ is the image resolution of the ith interval and c_S is the cost of stereo analysis per pixel, then the total cost of analyzing m stereo pairs is $\sum_{i=1}^{m} c_S H_i^2$. Vergence involves image registration in which one image (say, right) is shifted with respect to the other (i.e., left) until the overlapping regions around the left and right image centers are most similar. The computational cost of the registration process, using a $W_r \times W_c$ (W_r pixel rows and W_c pixel columns) neighborhood about the right image center, is $c_V W_r W_c$, where c_V is the cost of evaluating the similarity criterion function (e.g., normalized cross correlation) at one pixel. The precise focusing scheme requires the evaluation of a focus criterion function (e.g., gradient-squared sum) at each position of the sensor plane during reconfiguration. Let N be the number of such positions. If c_F is the cost of evaluating the square of the image gradient at every pixel within a $W_F \times W_F$ window, then the total cost at N positions is $c_F W_F^2 N$.

Since visible surfaces are interpolated from the depth points derived using stereo, an accurate reconstruction of these surfaces would require minimal uncertainty in the depth estimates of the underlying points. It has been noted above that one of the ways the uncertainty can be reduced is by bringing the surfaces into focus. In case of stereo, there are other sources of uncertainty such as lack of prior depth information to constrain the correspondence search in the images and occlusion. Both of these problems can be handled by performing stereo analysis of the parts of the scene for which approximate 3-D information is available, such as the defocused, low-resolution peripheral visual field. Consequently, the objective of our approach is to reduce the uncertainty in the structural information already available in the vicinity of the new fixation point. As has been argued earlier, the integration of camera motion and surface estimation while moving from one fixation point to the next provides the opportunity for this uncertainty reduction. The only remaining problem is to select this new fixation point from the candidate set belonging to the peripheral visual field. In the next section, we present an algorithmic solution to this problem and also to the integration scheme.

4. An algorithm for active surface estimation

This section describes an algorithm to achieve the desired integration of camera reconfiguration and coarse-to-fine surface reconstruction, wherein the coarse peripheral maps emerge to provide new targets and the cameras refixate to home on to these targets. To describe the algorithm, consider the state in which a fine surface map has been constructed for the central visual field along with a coarse map for the peripheral visual field with respect to the current fixation point. Then, the algorithm for iteratively extending the surface map consists of the following steps: (1) An unoccluded peripheral point, whose selection involves minimum lateral movement of the cameras and reconfiguration of their sensor planes, is chosen as the new target point; (2) a sequence of images of increasing resolution is acquired and stereo analyzed using the largest available focal length for the cameras, thus obtaining surface maps with increasing accuracy, during the time the cameras verge and focus on the new target point; (3) the improved depth

Surface estimation algorithm

Repeat until done

1. Select target based on current peripheral surface description
2. Home-in on target
   - 2.1 Aim both cameras at target point and set lenses to full zoom
   - 2.2 Repeat until target is fixated
     - 2.2.1 Acquire images of target area and subsample at a density inversely proportional to blur
     - 2.2.2 Obtain surface estimate from stereo analysis of subsampled images
     - 2.2.3 Aim cameras at target point indicated by stereo-based estimate
     - 2.2.4 Vary vergence to register image centers
     - 2.2.5 Adjust sensor plane positions to minimize blur at image centers
3. Identify central and peripheral visual fields
   - 3.1 Reduce zoom and acquire stereo images
   - 3.2 Segment stereo images into in-focus and out-of-focus regions
4. Identify occluded regions
5. Perform surface estimation
   - 5.1 Extract features in central visual field and obtain fine surface map from stereo
   - 5.2 Simultaneously, extract features in the peripheral visual field and obtain coarse surface map from stereo

Fig. 1. Algorithm for surface estimation using active stereo.

estimate of the fixation point is used to select focus settings, thereby completing fixation; (4) acquired stereo images are segmented into central and peripheral visual fields and a fine surface map is obtained for the former while a coarse map is obtained for the latter. Surface reconstruction for the central and peripheral visual fields can be performed in parallel, with more computational resources allocated per unit surface area to the former because of its greater computational requirements. These steps of the algorithm are summarized in Fig. 1 and elaborated upon in the following subsections.

4.1. Target selection

The extension of the surface map resumes by identifying an object in the peripheral field for fixation. The availability of the peripheral surface map makes the selection of a new fixation point possible, albeit with limited accuracy, and thus helps to avoid the need for knowing object depths before they are estimated!

Given an approximate surface map in the peripheral visual field, how should we select a fixation point? In [1], some criteria were identified for the selection of a fixation point which were motivated by known characteristics of fixation in human vision as well as computational considerations. Shmuel and Werman [23] have considered the related problem during surface map generation from multiple viewpoints; they use iterative Kalman-filtering techniques to predict a new camera pose for maximal reduction of

uncertainty in depth information. Burt [9] has considered hierarchical approaches to the target selection process; pyramid-based implementation searches for information pertinent to a chosen task by processing images at multiple spatial resolutions. In [22], an augmented hidden Markov model has been used to learn and generate a sequence of controlled movements to direct cameras at selected parts of the scene. Some recent studies have considered higher-level criteria for fixation, e.g., in object recognition [8]. Our selection criterion involves minimization of two costs: mechanical and computational. The mechanical cost reflects the effort required for geometrically reconfiguring the cameras—it involves changing the direction of gaze, viz., the optic axes, of the cameras and fixating them on the new object. The computational cost is proportional to the amount of computation required for the reconfiguration and for surface estimation during camera reconfiguration.

In the previous section, we outlined the steps of integrating camera reconfiguration and surface reconstruction. We now further analyze the mechanical and the computational aspects of each integration step, and derive the expressions for each cost. Since, we must later combine the two costs, we express each in terms of the amount of time required. Clearly, the time required depends upon the amount of change made, the speed at which it is made, and whether the different changes are made in parallel or sequentially. Here we will assume that the changes are made sequentially, although their parallel executions would be desirable and easy to do.

The extent of change in the electromechanical actuator settings, $\boldsymbol{q}$, necessary to reorient and fixate the cameras is given by

$$\Delta|\boldsymbol{q}| = \left[\Delta\phi_p + \Delta\phi_t + \Delta\phi_{vL} + \Delta\phi_{vR}\right] + \left[\left(\sum_{i=1}^{m}\Delta\phi_{fi}\right)_L + \left(\sum_{i=1}^{m}\Delta\phi_{fi}\right)_R\right] + \left[\Delta\phi_{fL} + \Delta\phi_{fR}\right]. \tag{4}$$

The first term includes the changes in the pan, tilt and vergence angles (necessary) to reorient the cameras. The second term is the total change in the sensor plane positions for coarse focusing of the left and right cameras in m steps. The third term is the amount of sensor plane adjustments for fine focusing. Let the angular speeds of the pan, tilt, vergence, and focus actuator motors be denoted by ω_p, ω_t, ω_v, and ω_f, respectively. Then the total mechanical cost (time) of reconfiguration and fixation is

$$C_m = \left[\frac{\Delta\phi_p}{\omega_p} + \frac{\Delta\phi_t}{\omega_t} + \frac{\Delta\phi_{vL}}{\omega_{vL}} + \frac{\Delta\phi_{vR}}{\omega_{vR}}\right] + \left[\frac{1}{\omega_{fL}}\left(\sum_{i=1}^{m}\Delta\phi_{fi}\right)_L + \frac{1}{\omega_{fR}}\left(\sum_{i=1}^{m}\Delta\phi_{fi}\right)_R\right] + \left[\frac{\Delta\phi_{fL}}{\omega_{fL}} + \frac{\Delta\phi_{fR}}{\omega_{fR}}\right]. \tag{5}$$

If the changes represented by the three terms could be carried out in parallel, we would need to retain only the maximum of the summands within each bracket.

The total computational cost (time) of reconfiguration and surface reconstruction is

$$C_c = c_V W_r W_c + c_S \sum_{i=1}^{m} H_i^2 + \left[c_F W_F^2 N_L + c_F W_F^2 N_R\right]. \tag{6}$$

The first term is the total cost of evaluating the correlation criterion function for vergence registration with c_V time units per pixel. The second term is the cost of obtaining surfaces from stereo in m steps with c_S time units per pixel. The third term is the cost of evaluating the focus criterion function in N_L and N_R steps for the left and right cameras, respectively, with c_F time units per pixel. Thus, the total cost or duration of camera reconfiguration and fixation is

$$C = C_m + C_c. \tag{7}$$

Suppose, there are N objects in a scene represented by N vertices of a graph. Then the problem of surface reconstruction for the scene is one of fixating these objects successively, moving the fixation point along each object surface, and reconstructing surface around the fixation points. Assuming that no object is fixated twice, to determine the order in which objects are fixated amounts to finding a *path* in the graph. If a cost is assigned to every change of fixation from one object to another, then the solution is one of finding the *minimal cost path* in the graph. However, since the number of objects and their spatial relationships are unknown, identification of the globally minimal cost path is not possible. While selecting the next point of fixation, only the M objects ($M \ll N$) present in the current peripheral field of view are considered, and the path having the minimal cost is found. This selected path is called *locally* minimal. To find this path, each edge of the graph connecting the currently fixated object point to one of its M neighbors is assigned a cost C_i according to Eq. (7), and the edge having the minimal cost, $\min_{i \in M} C_i$, is used to select the next object for fixation.

4.2. Target homing

Once a target point has been selected on a new object, the cameras need to be reconfigured to fixate on the point. Let A be the 3-D point at which the cameras are fixated and B be the new target point. For illustrative purposes, B is shown to be further away from the cameras than A in Fig. 2. Stereo analysis from images obtained using a focal length of f_{stereo} yields an inaccurate depth estimate of the peripheral target point because of blurring of the peripheral features in the vicinity of the target point. Following our discussions in Section 3.2, let σ_t be the spread parameter of the Gaussian that causes inaccurate localization of the peripheral features. The subsampling of the peripheral image regions further reduces the accuracy. The result is a very coarse estimate B_1' (of B) which may not correspond to a point on the actual surface of the object. This is illustrated in Fig. 2(a). However, the estimate serves as valuable initial source of information about how the cameras should start to reconfigure, i.e., in which direction the vergence and focus settings should be directed. In this sense, the coarse estimate may be viewed as qualitative. Once the reconfiguration process starts, however, the images in the vicinity of the target point become sharper and hence the accuracy

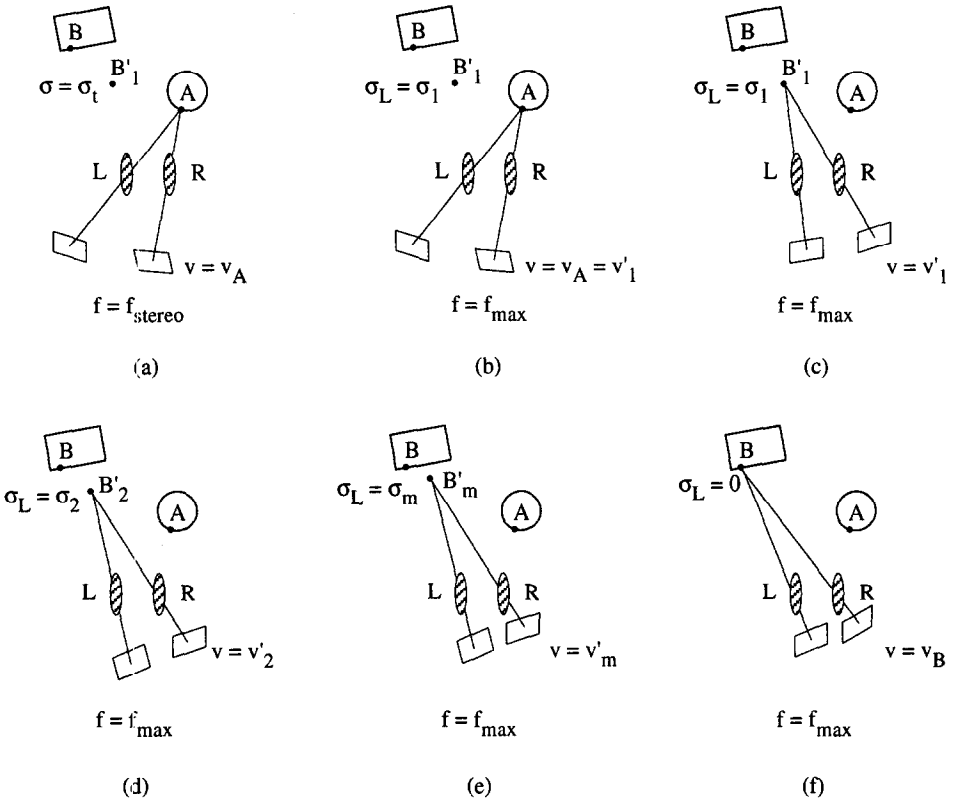


Fig. 2. The selection of a new target B is followed by changes in the camera orientations and focus axis settings ((a)–(e)) to home onto B. The final step in camera reconfiguration is the fixation of B (f). The larger the object distance, the closer is the sensor plane to the lens.

of estimate of the location of B improves. The goal of the camera reconfiguration process is modified accordingly. Thus, as the cameras continue to reconfigure to aim at the new target point, a coarse-to-fine image sequence of the new object in the vicinity of B is obtained along with a coarse-to-fine sequence of surface estimates in the vicinity of B. The increasing accuracy of the estimate of B and the frequent updating of target location information ensure that the cameras actually converge on the target eventually. The rest of this subsection presents details of the above target homing process.

Target homing is attempted using the largest available focal length, $f = f_{\max}$, which facilitates stereo analysis at subpixel (effectively) resolution. Since the estimate B_1' is inaccurate, the reconfiguration is performed incrementally to reduce the overall cost by integrating it with the process of improving the estimate of B. While still focused at the current fixation point, the change to large focal length from f_{stereo} (Fig. 2(b)) causes increased blurring of the new target point. Let the optical blur of the target point at the beginning of the target homing phase have a $\sigma_L = \sigma_1$:

$$\sigma_1 = k\frac{A f_{\max}}{Z_A}\left(\frac{\left|Z_{B_1'} - Z_A\right|}{Z_{B_1'} - f_{\max}}\right). \tag{8}$$

The cameras are initially oriented to aim at B_1', whose 3-D location is $X_{S1}' = (X_{S1}', Y_{S1}', Z_{S1}')$. The sensor plane locations of the two cameras are $v = v_A = v_1'$. This configuration is shown in Fig. 2(c). Since the depth estimate is only approximate, the two image centers may not contain the projection of the same surface point. A coarse vergence registration to find the offset for the right window of fixation that has the best similarity with the window of fixation in the left image is performed next. It evaluates the normalized discrete correlation function which is the similarity function of vergence registration at every pixel within the correlation neighborhood around the right image center. The projections of the points $X_{S1}' - \Delta X_{S1}'$ and $X_{S1}' + \Delta X_{S1}'$ determine the correlation neighborhood along the epipolar line in the right image. The variable $\Delta X_{S1}' = (\Delta X_{S1}', \Delta Y_{S1}', \Delta Z_{S1}')$ is the error in 3-D location of B due to stereo. Here, $\Delta Z_1'$ is the maximum error in depth, and $\Delta X_{S1}'$ and $\Delta Y_{S1}'$ are the corresponding maximum errors along the X and Y dimensions as determined from the perspective projection model of a pin-hole camera. Once the cameras are aimed at B_1' after the translational registration, a vergence-based estimate, Z_{V1}', of B is obtained. To do finer registration requires further interleaving of the processes of vergence registration and depth estimation from finer stereo.

As the sensor planes are gradually reconfigured, the new target point becomes less and less blurred; the image sequence acquired during the reconfiguration thus comprises a multiresolution (coarse-to-fine) image sequence of the target area. Each pair of optically blurred images is subsampled, reducing the degree of subsampling as images become less blurred, i.e., σ_L decreases. Let $H_i \times H_i$ denote the resolution of the sampled images at the ith stage during reconfiguration. Then,

$$\frac{H_1}{M} = \frac{\sigma_t}{n\sigma_1} \quad \text{and} \quad \frac{H_i}{H_{i+1}} = \frac{\sigma_{i+1}}{\sigma_i}, \tag{9}$$

where the peripheral resolution is $M \times M$ and $f_{\max}/f_{\text{stereo}} = n$, $n > 1$. The inverse dependence of the image resolution on the degree of blur is due to the fact that blurring causes a reduction in spatial frequency, thus lowering the sampling rate. The result of stereo analysis of the images blurred with $\sigma_L = \sigma_1$ is an improved estimate B_2' of B (Fig. 2(d)). Since the optically blurred images are obtained continuously, the improvement in the stereo-based depth estimate of the target point from the analysis of two consecutive image pairs is significant only when the difference $\Delta\sigma_L = \sigma_i - \sigma_{i+1}$ is significant. Let $\Delta\sigma_T$ be the chosen significant value of $\Delta\sigma_L$. Therefore, the intermediate images in which the blur of the target point is between σ_i and $\sigma_{i+1} = \sigma_i - \Delta\sigma_T$ are skipped. When the cameras are aimed at B_2', which is the currently estimated location of B, the new sensor plane position that is used to acquire the next pair of optically blurred images is v_2'. The blur of the target point in these images has a $\sigma_L = \sigma_2 = \sigma_1 - \Delta\sigma_T$.

The surface estimates derived from an image pair at any stage during camera reconfiguration serve as coarse estimates for surface reconstruction from later images acquired with smaller σ_L. After the ith image is obtained, an estimate Z_{Si}' from stereo and an ad-

ditional estimate Z'_{V_i} from vergence are obtained for the target point. The more precise of the two estimates, one that has the lower variance, is then used to aim the cameras at the target point. If at any stage, m, during this coarse-to-fine image acquisition interleaved with surface reconstruction the blur of the target point, σ_m, is smaller than or equal to a preset lower bound, $\sigma_{\min}$, surface reconstruction is discontinued, but the sensor plane reconfiguration is continued till the estimated blur of the target point is zero. (This stage of the target homing process is illustrated in Fig. 2(e).) The significance of the lower bound, $\sigma_{\min}$, is that it corresponds to the smallest discernible blur circle (cf. Eq. (1)). The images acquired after the estimated σ_L is zero are stereo analyzed using the finest resolution grid. The resulting stereo estimate of the fixated point is used to determine the final sensor plane position v_m.

4.3. Target fixation

The target homing stage terminates with the cameras oriented such that the estimated fixation point location falls at the center of each image. In [1], the *initial* fixation is intended to be achieved by ensuring that the focus-based depth estimates of the fixation point obtained from the two cameras are consistent, and that the fixation windows centered at the image centers in the two images are in registration. However, this state may lead to an arbitrarily long sequence of vergence adjustments since focus and vergence adjustments may continue to miss the desired state of fixation. To avoid this, [1] uses a threshold on the difference between the left and right focus-based estimates to terminate the fixation process. Also, both focus and vergence mechanisms are based on image intensities and hence are affected by scene illumination. On the other hand, the effect of illumination is much less pronounced on stereo. The increasing precision of the stereo estimate obtained during target homing brings the two cameras close to the state of fixation with the estimated blur of the target point zero and the two cameras aimed at the same 3-D point. To focus the cameras based on the evaluation of the gradient-based criterion function, a small interval of sensor plane positions, $[v_{1m}, v_{2m}]$, is established about v_m, where $v_{1m} > v_m$ and $v_{2m} < v_m$. This interval, which is the depth of focus at v_m, is finely quantized and searched for a peak of the focus criterion function. To reduce the effect of noise on the localization of the peak of the criterion function, each of the left or right images is actually an ensemble average of several images taken at successive time instants using identical imaging parameters. The peaks in the left and right images correspond to the best sensor plane positions, v_{B_L} and v_{B_R}. The final camera configuration, shown in Fig. 2(f), is used to initiate surface reconstruction for the new object.

4.4. Multiresolution surface estimation

Stereo images are acquired using an intermediate focal length $f = f_{\text{stereo}}$ and the visual field is segmented into central and peripheral parts. 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 further such that the corresponding depths of field

are adjacent to the depth of field of the fixation point. If a lens is to be focused at three different object distances, Z_{0n}, Z_0, and Z_{0f} so that their associated depths of field border one another, we can derive the following expressions for Z_{0n} and Z_{0f} in terms of Z_0 using Eq. (1):

$$Z_{0n} = \frac{Z_0 f(A + D_0)}{f(A - D_0) + 2D_0 Z_0}, \tag{10}$$

$$Z_{0f} = \frac{Z_0 f(A - D_0)}{f(A + D_0) - 2D_0 Z_0}, \tag{11}$$

where $f = f_{\text{stereo}}$. In Eqs. (10) and (11), D_0 is the diameter of the *circle of confusion*, which is the blur circle associated with a point object located anywhere within the depth of field.

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. To determine whether the 3-D point projecting on a pixel in the in-focus image is within the depth of field, its depth value, if available, is used. When the range is within the depth of field, the pixel is marked as in-focus. In the absence of any depth information, the focus criterion function is evaluated. The output of the function at every pixel in the in-focus image is compared with the outputs at the corresponding pixel location in the out-of-focus images. If the former value is the largest, the pixel is considered to be imaged sharply, otherwise it is defocused. The process is repeated by taking each of the out-of-focus images and comparing it to the remaining images at those pixels which are not yet classified. At the end of the segmentation process, every pixel is marked as either near-focused, in-focus, or far-focused. The in-focus regions of the image constitute the central field of view and the defocused regions comprise the peripheral field. The exact location of the peripheral visual field with respect to the point of fixation will depend upon where the scene depth crosses out of the current depth of field, and will thus be determined by the scene. Two surfaces separated by a depth larger than the current depth of field cannot be included in the central field simultaneously. A depth discontinuity contour separating these surfaces becomes a part of the boundary of the central region. In other words, discontinuity of focus is coincident with discontinuity of depth or disparity. Occluded parts of the scene which are located near depth discontinuities are identified by projection ray tracing of all points on a depth discontinuity contour to the projection centers of left and right cameras using available surface information and positions of the cameras. Following this, surface reconstruction for the unoccluded parts of the central visual field is initiated.

Parts of the central visual field have highly accurate estimates obtained during high-zoom target homing. This is illustrated in Fig. 3. For the rest of the central visual field, surface reconstruction on an $N \times N$ grid within the stereo images takes place using a small value of σ (σ_{ctl}) for the Laplacian of Gaussian ($\nabla^2 G$) feature detector. The choice of σ_{ctl} gives the best trade-off between the localization and stability of the detected zero-crossings. The surface reconstruction for regions not analyzed during target homing begins with initial surface estimates obtained in two different ways. Certain parts

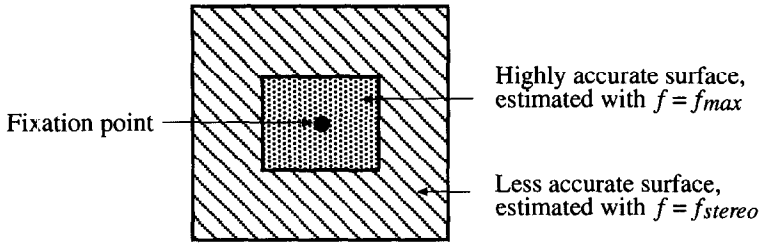


Fig. 3. Multiple surface estimates of varying accuracy within central visual field.

of the central visual field have only coarse estimates available from the previous fixation at which time these parts belonged to peripheral field. Other parts of the central visual field may have entered the visual field during refixation and thus do not have any associated estimates; for these parts, the depth estimate of the current fixation point obtained at the end of target homing is used as the initial depth estimate during stereo matching of features. Stereo matching of the detected features yields 3-D points. These points are clustered and a surface smoothness constraint is enforced to remove the false ones. Quadratic surface patches are fit to the cluster of points. Finally, range values are interpolated from these smooth surfaces. The result of stereo reconstruction is a high-resolution (*fine*) surface map for the central visual field.

A σ_{pfl} larger than σ_{ctl} is used for the peripheral feature detector to introduce smoothing in addition to that caused by optical blurring so that the number of matchable features is small. In addition to smoothing, the periphery is subsampled using an $M \times M$ grid ($M < N$). The effects of blurring and subsampling significantly degrade the accuracy of stereo and lead to a low-resolution (*coarse*) surface map for the peripheral visual field.

An implementation of the algorithm described in this section is presented in the next section. A proof of the convergence of the algorithm is outlined in the following section.

5. Implementation and results

This section presents the details and results of implementing the active stereo algorithm described in the previous section. The salient points of our particular implementation are given first, followed by the presentation of the results.

5.1. Implementation details

The algorithm was implemented on the University of Illinois' Active Vision system. The system consists of two Cohu 4815 CCD cameras mounted on a stereo platform and equipped with Vicon V17.5-105M motorized version of Fujinon C6 × 17.5B lenses. The cameras have a resolution of 754H × 488V pixels, where each pixel is 11.7μm wide and 13.5μm high. The lenses have computer-controlled focus, aperture, and zoom settings with 17.5–105mm focal length and a maximum f-stop of 1.8. Minimum focusing distance is 1.3m. High-precision stepper-motor rotational units are used to independently

Table 1
The lens and position controllers of the imaging system; the speeds take into account the initial setup time for the motors; the maximum speed of each positioner unit is ten times the listed speed

Lens controllers		Positioners		
Unit	Speed (steps/s)	Unit	motor step	speed (steps/s)
Focus	5000	Vergence	0.01°	330
Aperture	4670	Pan	0.002°	950
Zoom	4000	Tilt	0.001°	830
		Translation	10μm	900

control pan, tilt, and vergence angles. The system can also translate horizontally with one degree of freedom. Some operational details of the lens and positioner actuators are listed in Table 1. The imaging system is controlled by a Sun Microsystems 3/160 workstation.

For the left and right cameras, calibrated focal lengths of $f_{\text{stereo}} = 47.7\text{mm}$ and 47.2mm are used to acquire the stereo images, and $f_{\max} = 105.4\text{mm}$ and 101.0mm (full zoom) are used in the fixation process. The calibrated image centers are $(r_{0L}, c_{0L}) = (270.1, 243.0)$ and $(r_{0R}, c_{0R}) = (241.0, 255.0)$. (Actually, the system undergoes a self-calibration phase following every fixation and stereo analysis [1].) The approximate distance between the cameras is 28cm. The translational unit of the stereo cameras is not used in our implementation. Since the baseline changes as the cameras converge and diverge, the transformation of each camera-centered coordinate frame with respect to a base coordinate frame is explicitly calculated to obtain the baseline whenever the cameras reconfigure. Images are digitized to 512×512 pixels of 256 gray levels each. The sensor plane position is adjusted through the focus setting controller (equivalent to the focusing ring on a camera lens). Empirically, the relation between the sensor plane position v (expressed in units of length) and the focus setting p (expressed as a number) is $v = ap + b + f$ for a zoom setting f. The calibrated parameters for full zoom are $a = -6.08E-07$ and $b = 0.009$ (left camera), and $a = -5.2E-07$ and $b = 0.008$ (right camera). The circle of confusion is experimentally determined to be 2 pixels of the CCD imaging array. The relation between the diameter D of the blur circle and the spread parameter σ_L of the Gaussian associated with lens defocusing is experimentally found to be $\sigma_L = kD + \sigma_0$. The calibrated values of k and σ_0 are 0.35 and 1.09, respectively.

The values of $\sigma_{\text{ctl}} = 6$ for the central visual field and $\sigma_{\text{pfl}} = 9$ for the peripheral field are used in the implementation of our algorithm. The central field is stereo analyzed using an $N \times N = 256 \times 256$ grid; the stereo grid for the peripheral field is $M \times M = 128 \times 128$. In our implementation, the angular displacement of each motor is measured in units of a motor step and the speed in motor steps per second. Thus, the angular speed of the pan unit from Table 1 is about 1.9°/s and that of the tilt unit is about 0.8°/s. (Compare these with the saccadic angular speed of 400°/s in humans.) The unit cost of stereo analysis during target homing is $c_S = 20\text{ms/pixel}$ ($c_S = 0.5\text{ms/pixel}$ using a parallel architecture [10]).

To efficiently allocate resources, we exploit the fact that the central and the peripheral visual fields can be processed in parallel. Each visual field in turn is divided by a two-

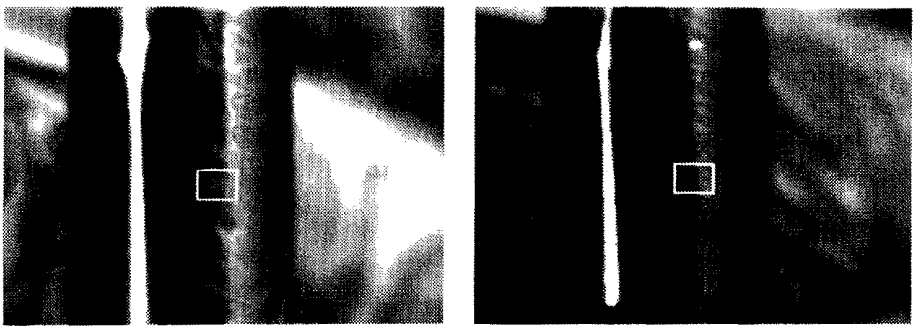
dimensional grid into regions within each of which computation proceeds independently of others. The stereo reconstruction can therefore be performed on a parallel architecture. Since the computation depends on the number of features present inside the grid regions, it is important to correctly partition and distribute the data to the processors to obtain significant performance improvement (speedup). Partitioning the data uniformly among the processors allows good speedups for the data independent tasks, such as feature detection and image segmentation. However, dynamic scheduling and computational load balancing are important for the data dependent tasks of stereo matching and surface fitting. In our simulated parallel implementation, six MC68020 processors (a network of SUN 3's) comprised the multiprocessor system with distributed memory in which the 3/160 workstation served as the central processor.

5.2. *Experimental results*

We first present experimental results demonstrating some basic capabilities of the dynamic imaging system such as target homing. In this experiment, the imaging system is coarsely aimed at a soda can. The focal length of each camera is set to full zoom, resulting in the optically blurred images of Fig. 4 when the target homing process is initiated. The 3-D coordinates (hand-measured) of a point on the can that is projected at the center of the left image are $(X, Y, Z) = (0.584, 0.344, 1.821)$. The coordinate values are all in meters and expressed in a reference world coordinate system associated with the experimental setup. The initial focus setting is $p = 0$ and the optical blur has $\sigma_L = \sigma_1 = 8$ pixels. The subsampled size for these blurred images is chosen as $H_1 = 64$. The Nevatia–Babu line extraction algorithm [20] is used to detect features in these subsampled images, which are matched to obtain a coarse stereo map of the soda can. The updated stereo estimate for the target point is used to reaim the cameras at it, using coarse vergence registration. The window for evaluating the correlation function during vergence registration is $W_V \times W_V = 49 \times 49$ pixels and the neighborhood around the right image center is $W_r = 30$ pixels high. The reduction step in the blur to acquire successive pairs of images for coarse-to-fine stereo analysis is made to be dependent on the most recent blur; in particular, $\Delta\sigma_T = \sigma_i/2$. The images acquired during successive sensor plane reconfiguration are shown in Figs. 5 and 6. The lower bound on the optical blur, beyond which only coarse-to-fine image acquisition occurs without image analysis, is $\sigma_{\min} = 3$. The coarse-to-fine stereo analysis to fixate the target point are summarized in Table 2.

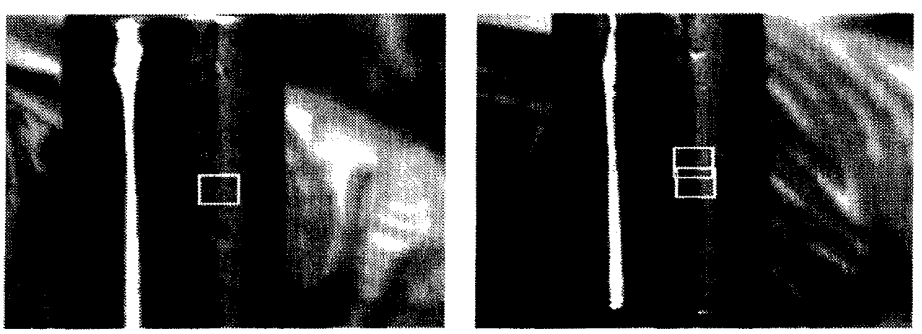
Once the estimated blur σ_i reaches the lower limit of $\sigma_{\min}$, only sensor plane reconfiguration takes place without stereo analysis until $\sigma_i = 0$. At this point, the cameras are aimed at the target using the finest stereo estimate. Finer adjustments of the sensor planes as indicated in Table 3 are performed to focus the target exactly and to complete the fixation process.

In the second set of experiments, the active camera system was made to scan an indoor scene. A perspective of the scene from the left viewpoint (camera) is shown in Fig. 7. It shows a box and a barrel, the back rest of a chair, and a beverage container in the foreground with a rear wall in the background. Artificially textured surfaces were used to ensure deviation from *uniformity* and to derive *dense* depth maps using feature-based



(a) (b)

Fig. 4. The (a) left and (b) right images of the initial ($i = 1$) stereo pair. The rectangular boxes are centered at the calibrated left and right image centers.

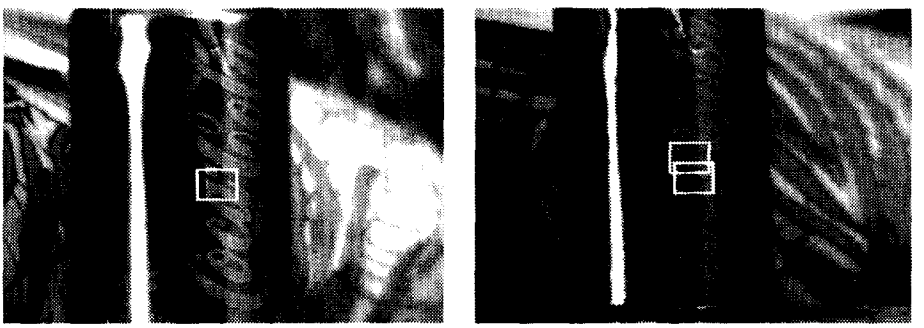


(a) (b)

Fig. 5. The (a) left and (b) right images after aiming the cameras at the target whose location is estimated using coarse stereo. The focus setting is also changed ($i = 2$) to obtain sharper images of the target.The lower window in the right image encloses the calibrated image center and the upper window represents the best match to the center of the left image obtained during image registration.

stereo, as density of maps depends on the degree of image detail. However, as previous experimental results indicate, such surfaces are not crucial for the approach. Restrictions were imposed on the travel limit of the cameras to reconstruct only the scene depicted in Fig. 7.

The system is initially fixated at the box (the rightmost window on the box). Successive fixations, indicated in Fig. 7 by three windows aligned nearly horizontally, continue scanning of the box wherein the surface map is extracted through the iteration of the



(a) (b)

Fig. 6. The (a) left and (b) right images after minimizing the estimated optical blur ($i = 3$). Analysis of this pair gives the finest stereo estimate. Focus criterion function is evaluated next over the rectangular box in each image.

Table 2
The results of stereo analysis and vergence registration during the successive phases (denoted by i) of homing on the object shown in Fig. 4

i	Focus axis		Focused depth		Blur	Sampling	Stereo depth	Registration window	Vergence depth
	p_L	p_R	Z_{F_L} (m)	Z_{F_R} (m)	σ_i	H	Z_S (m)	W_C (pixels)	Z_V (m)
1	0	0	1.644	1.640	8	64	1.818	30	1.818
2	1465	1514	1.765	1.764	4	128	1.821	10	1.823
3	2195	2112	1.836	1.820	0	256	1.819	1	1.821

Table 3
The results of fine focusing the cameras; the window for evaluating the focus criterion function is $W_F \times W_F = 49 \times 49$ which encloses each image center

Focus adjustments							
Left focus setting			Left focus depth	Right focus setting			Right focus depth
p_{1m}	p_{2m}	p_L	Z_{F_L} (m)	p_{1m}	p_{2m}	p_R	Z_{F_R} (m)
1756	2382	2043	1.820	1550	2411	2229	1.832

steps of visual target selection, fixation, and stereo reconstruction [1]. During the third fixation of the box, the stereo images of Fig. 8 are acquired using a focal length of $f = f_{\text{stereo}}$. Here, the box, being in-focus, occupies the central visual field. Features are detected in the central visual field using the $\nabla^2 G$ feature detector that has a $\sigma = \sigma_{\text{ctl}}$. Smooth surfaces are obtained as a result of stereo matching these features, and range

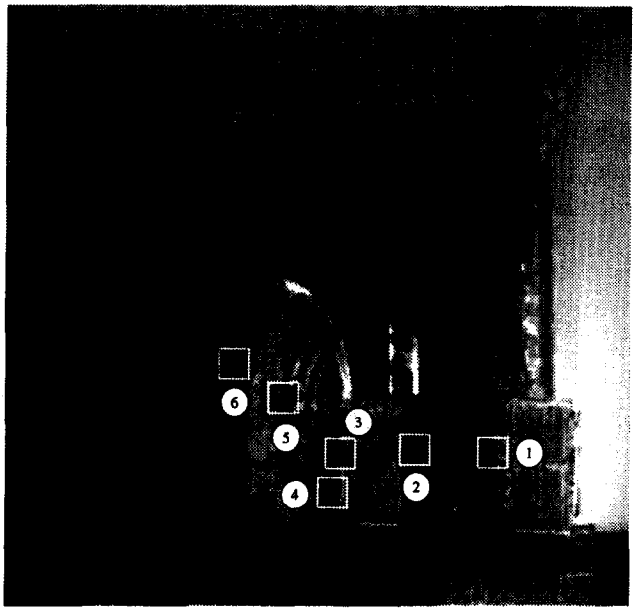
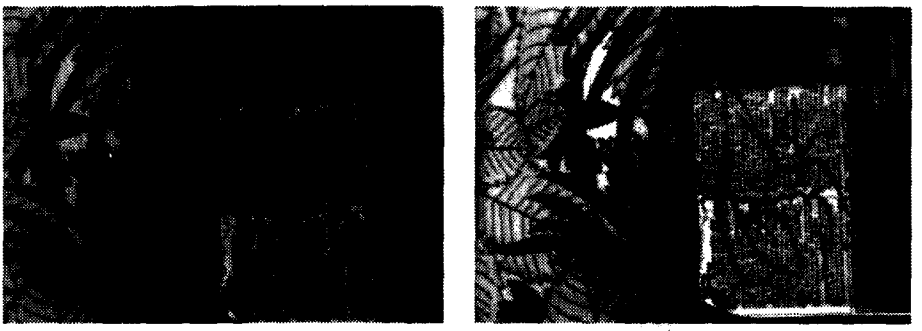


Fig. 7. An overview of the scene that has four objects—a box, a chair, a container and a barrel—in front of a rear wall. Several fixation points are indicated by enclosing windows.

values are interpolated. The fine central range map is added to the evolving composite map in Fig. 9(a) which shows the extension of the depth map of the box. Apart from the box, the stereo images for this fixation show the presence of two more objects of the scene—the back rest of the chair and the beverage container—located in the peripheral visual field.

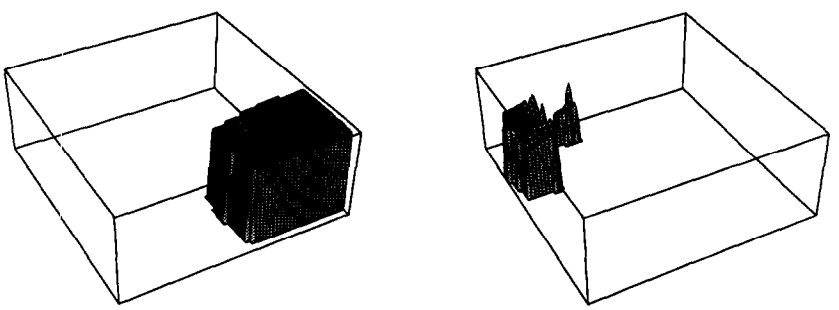
The target selection process fails to locate a new target on the box at the end of this fixation. The search is subsequently extended to the peripheral visual field where coarse depth maps of the chair and the container are now available. These maps are shown in Fig. 9(b). The target selection method identifies a point on the chair, highlighted by the lowermost window in Fig. 10, that minimizes the readjustments of the mechanical components of the imaging system and the computational cost during the target homing process. The predicted changes in the mechanical components are summarized in Table 4. As a comparison, the reconfiguration parameters estimated during the target selection process for the point on the container (highlighted in Fig. 10) which has the minimum cost among all the points belonging to the container for which depth values are available from the peripheral map are also included in the table. The world coordinates of the new target point on the chair are $(X, Y, Z) = (0.671, 0.457, 2.014)$, as estimated from the coarse peripheral depth map.

Next, the cameras are reoriented to aim at the new target point on the chair. After three steps of target homing in which sensor plane reconfiguration is integrated with coarse-to-fine stereo analysis, a depth estimate of 2.014m is obtained from stereo and



(a) (b)

Fig. 8. Stereo image pair, (a) left and (b) right, acquired during the third fixation of the box. The chair and the container are the peripheral objects.



(a) (b)

Fig. 9. (a) The composite depth map showing extension of the reconstructed surface of the box during the third fixation. (b) The coarse depth map for the current peripheral visual field. The leftmost object is the back rest of the chair while the remaining object is the container.

Table 4
The estimated changes in the camera configuration necessary to fixate the new target point on the chair and their respective durations; the entry new′ corresponds to the target point on the container

Unit	Settings in steps			Time	Settings in steps		Time
	old	new	new − old	(s)	new′	new′ − old	(s)
Left focus	2173	3920	1747	0.35	4862	2689	0.54
Right focus	2275	3772	1497	0.30	4781	2608	0.52
Left vergence	−287	−230	57	0.17	−190	97	0.29
Right vergence	600	554	−46	0.15	517	−83	0.25
Pan	−7591	−6776	815	0.86	−7387	204	0.21
Tilt	2265	1163	−1102	1.33	3757	1492	1.80

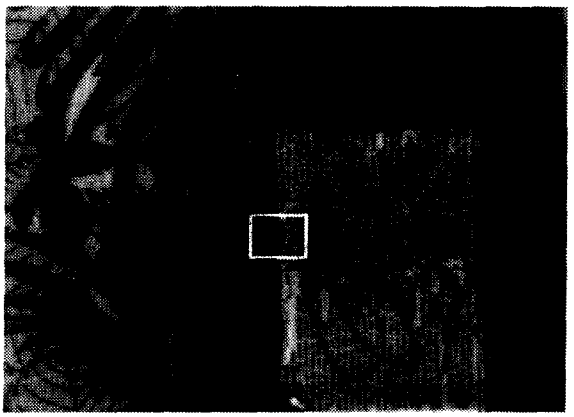
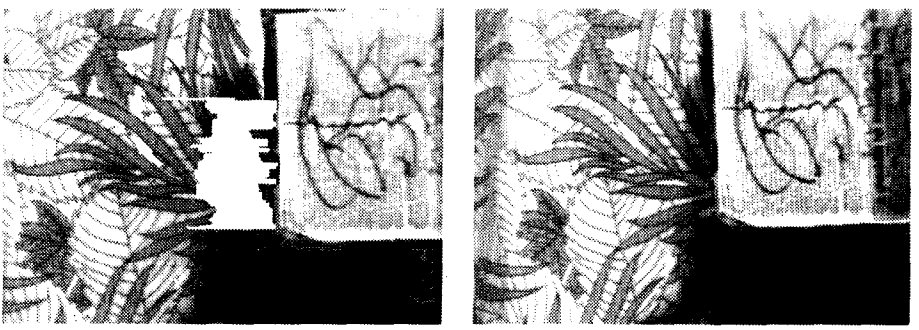


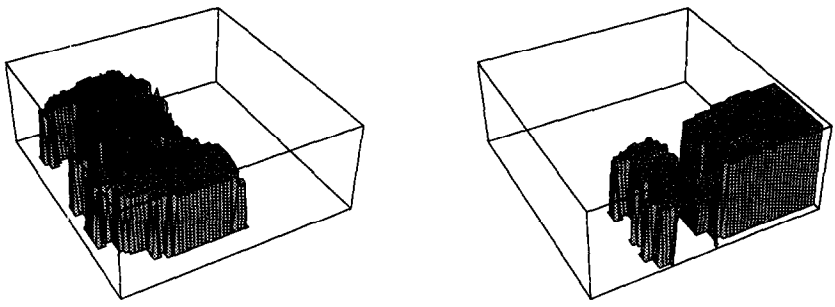
Fig. 10. The middle window encloses the current fixation point on the box and the lower window marks the new target point on the chair selected by the target selection criterion. For the new target, $\sigma_{LS} = 5$ pixels and $\sigma_I = 10$ pixels. The upper window highlights the point for which the criterion function has the smallest value among all the depth points on the container.



(a) (b)

Fig. 11. Stereo images after fixating the chair. (a) The regions occluded from the right camera (marked white) superposed on the left image and (b) right image.

an estimate of 2.025m is derived using vergence for the target point. The estimation error is 0.001m for both cues. The focus-based depth estimates after fine adjustments of the sensor planes at the completion of target homing are 2.006m and 1.982m using left and right cameras, respectively. The final estimated position of the fixation point is



(a) (b)

Fig. 12. (a) The depth map of the chair combining the high-resolution map from target homing and the relatively low-resolution map from post-fixation stereo. (b) The depth map of the chair added to the composite map previously containing the box.

$(\hat{X}, \hat{Y}, \hat{Z}) = (0.671, 0.459, 2.023)$. For the surface reconstruction phase, stereo images of Fig. 11 are acquired. Parts of the back rest occluded from the right camera by the box are identified using stereo estimates for the box and the back rest. These are highlighted in Fig. 11. The surface map for the central visual field obtained from stereo analysis is combined with the high-resolution map from the target homing process in Fig. 12(a). The combined map is added to the composite map previously containing the depth map of the box only. The result is shown in Fig. 12(b).

6. Convergence of the surface estimation algorithm

In this section we analyze the convergence condition for the surface estimation algorithm. We assert that the algorithm terminates only when the surface scanning process is completed, i.e., there exists no object in the scene which has not been scanned. We assume that every object in the scene has sufficient image details for stereo, focus, and vergence to work. Additionally, our assumption is that the surface estimation process is not subjected to calibration or stereo matching errors, and that the error in the depth estimate of a surface point is due only to the random errors in stereo, focus, or vergence. Since the termination of the algorithm implies no addition to the evolving composite surface map, the convergence condition for the surface estimation algorithm may be expressed by the following statement: *if* there is no addition to the composite surface map, *then* scanning is completed. In the following, we justify the validity of this statement.

We assume that the scene contains a finite number of objects and each object has a finite nonzero surface area. The projections of such objects on an infinite image plane are completely enclosed by finite-length boundaries. Let all the objects be represented by the vertices of a graph. The fixation of an object is equivalent to the labeling of an unmarked vertex and connecting it to the vertex corresponding to the object just scanned. Thus, when the scene is completely scanned, there exists a path through all the vertices of the graph.

At the initial fixation when a single object in the scene is fixated, the correct matches for features in the vicinity of the fixation points are found at the locations predicted by focus and vergence, and a surface patch is subsequently obtained from stereo [1]. For the second fixation, a point is chosen on the boundary of the surface patch such that this new target point is not on an occluding contour. As a result, there is a nonzero increment in the composite surface map after the second fixation. If the patch at the first fixation has holes, then these are filled up in subsequent fixations [1]. The surface patch does not grow outwards but there are additions to the composite surface map corresponding to the holes. By induction, after the n_ith fixation of the ith object (that corresponds to the ith vertex of the graph) either the surface map is extended in the $(n_i + 1)$th fixation or a hole is filled up. In other words, at the $(n_i + 1)$th fixation, the surface area to be scanned is less than that at the n_ith fixation. This implies that the surface scanning process for a single object converges.

Suppose that, after the N_ith fixation, there are no holes in the surface patch and all potential fixation points are on the occluding contour. Surface scanning of the ith object is therefore completed, and scanning must resume by fixating another object lying across the occluding contour. There can be at most a finite number of such peripheral objects adjacent to the finite-length occlusion boundary. The target selection process terminates in a finite amount of time with the identification of the new target. With every new fixation, the number of objects that remain to be scanned decreases.

The convergence of the target homing process is a direct consequence of the fact that it monotonically reduces the error in the estimate of the fixation point. Therefore, it terminates with zero image blur and disparity at the image centers in a finite number of steps. Consequently, a vertex representing the fixated object is labeled and included in the path.

If an object surface has been reconstructed and its peripheral objects have already been fixated, then the target selection process would fail. This will require backtracking to the previous vertex in the path and examining the boundary of the corresponding object to determine whether there exists at least one unmapped object adjacent to that boundary. If true, one such unmapped object is fixated. Otherwise, the backtracking continues. If in this process of backtracking the starting vertex is reached, then there exists no mapped object that is adjacent to an unmapped object, i.e., the scene has been completely scanned. Notice that there is no addition to the composite surface map during the search.

7. Conclusions

This paper has described an active approach to surface estimation from stereo images of large scenes having large depth ranges and depth discontinuities. The following are some salient features of the approach presented:

- adaptively controlling the imaging parameters to acquire the best possible sensory data,
- optimally reconfiguring the cameras to sense different parts of the scene—such reconfiguration must utilize the processed data efficiently to reduce the total time for computation and mechanical reconfiguration,

- acquiring and maintaining surface maps from different viewpoints to build a consistent description of the scene,
- parallelizing and interleaving the numerous operations that are involved in a meaningful way to reduce the complexity of the problem.

Finally, it has been proved that the active surface estimation process terminates when the entire scene has been scanned.

There are three distinct facets of the approach presented. First, during any given fixation, multiple 3-D cues cooperate to synthesize an accurate surface map for the fixated high-resolution part of the scene; relatively inaccurate maps are derived in parallel for the defocused low-resolution parts of the scene. Second, the low-resolution peripheral surface maps help the active vision system to select targets in a cost-effective manner. Finally, the refixation step involves the integration of the mechanical reconfiguration of cameras and multiresolution surface estimation with the gradual deblurring (optical) of the new fixation point.

At the beginning of the surface estimation process, the fixated parts of the scene, referred to as the central visual field, and the nonfixated parts, referred to as the peripheral visual field, are distinguished. Next, occluded regions adjacent to depth discontinuity contours are identified. Features belonging to these regions are excluded from stereo analysis. The selection of new target points on objects other than the one currently being scanned is cast as an objective function to be minimized. The criteria favor visual targets that are near the current fixation point, both laterally and in depth. The objective function value is proportional to the time taken to mechanically refixate the cameras and to stereo analyze the images acquired during such refixation. The refixation phase is modeled as an error reduction process, iterating over the steps of surface estimation and camera readjustments. The integration of the cues of stereo, focus, and vergence is performed to achieve fixation. The small error reduction step avoids overshooting the minima, thereby reducing the total number of camera readjustments necessary to fixate a point. Experimental results from a particular implementation of the approach have been presented.

Perhaps the most important aspect of integration that is illustrated by the work reported is how the scope of an individual process may be extended in a cooperative environment. Examples are the use of focus to determine the fixated parts of the scene and the depth discontinuity contours and to acquire optically blurred multiresolution images during target homing; the use of focus and vergence together to identify occluded regions of the central visual field; the use of stereo to determine the size of the vergence registration window and the search interval of focus settings. Such extensions reduce the operational costs of individual cues making an integration scheme beneficial.

The prospect of a parallel implementation of the algorithm [10] is an indication of its potential to support real-world applications, such as navigation and manipulation, requiring surface information. The resolution of the surface can be appropriately selected based on speed-accuracy trade off for a particular application. A natural extension of the approach is the utilization of space-variant sensors [4,25] to further speed up processing and to exhibit greater anthropomorphic behavior during visual attention.

References

[1] N. Ahuja and A.L. Abbott, Active stereo: Integrating disparity, vergence, focus, aperture, and calibration for surface estimation, *IEEE Trans. Pattern Anal. Mach. Intell.* **15** (1993) 1007–1029.

[2] Y. Aloimonos, I. Weiss and A. Bandyopadhyay, Active vision, in: *Proceedings First International Conference on Computer Vision*, London (1987) 35–54.

[3] R. Bajcsy, Active perception, *Proc. IEEE* **76** (1988) 996–1005.

[4] R. Bajcsy, Active observer, in: *Proceedings DARPA Image Understanding Workshop*, San Diego, CA (1992) 137–147.

[5] R. Bajcsy, E. Krotkov and M. Mintz, Models of errors and mistakes in machine perception, Part 1: First results for computer vision range measurements, in: *Proceedings DARPA Image Understanding Workshop*, Los Angeles, CA (1987) 194–204.

[6] D.H. Ballard, Reference frames for animate vision, in: *Proceedings IJCAI-89*, Detroit, MI (1989) 1635–1641.

[7] D.H. Ballard and A. Ozcandarli, Eye fixation and early vision: kinetic depth, in: *Proceedings Second International Conference Computer Vision*, Tarpon Springs, FL (1988) 524–531.

[8] R.M. Bolle, A. Califano and R. Kjeldsen, Data and model driven focus of attention, in: *Proceedings 10th International Conference on Pattern Recognition*, Atlantic City, NJ (1990) 1–7.

[9] P.J. Burt, Algorithms and architectures for smart sensing, in: *Proceedings DARPA Image Understanding Workshop*, Cambridge, MA (1988) 139–153.

[10] A.N. Choudhary, S. Das, N. Ahuja and J.H. Patel, A reconfigurable and hierarchical parallel processing architecture: Performance results for stereo vision, in: *Proceedings 10th International Conference on Pattern Recognition*, Atlantic City, NJ (1990) 389–393.

[11] J.J. Clark and N.J. Ferrier, Modal control of an attentive vision system, in: *Proceedings Second International Conference on Computer Vision*, Tarpon Springs, FL (1988) 514–523.

[12] D.J. Coombs and C.M. Brown, Intelligent gaze control in binocular vision, in: *Proceedings Fifth IEEE International Symposium on Intelligent Control*, Philadelphia, PA (1990) 239–245.

[13] S. Das and N. Ahuja, Multiresolution image acquisition and surface reconstruction, in: *Proceedings Third International Conference on Computer Vision*, Osaka (1990) 485–488.

[14] S. Das and N. Ahuja, Performance analysis of stereo, vergence, and focus as depth cues for active vision, *IEEE Trans. Pattern Anal. Mach. Intell.* **7** (1995) 1213–1219.

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

[16] D. Geiger and A. Yuille, Stereopsis and eye-movement, in: *Proceedings First International Conference on Computer Vision*, London (1987) 306–314.

[17] W. Hoff and N. Ahuja, Surfaces from stereo: Integrating feature matching, disparity estimation, and contour detection, *IEEE Trans. Pattern Anal. Mach. Intell.* **11** (1989) 121–136.

[18] E.P. Krotkov, *Active Computer Vision by Cooperative Focus and Stereo* (Springer, New York, 1989).

[19] D. Marr and T. Poggio, A computational theory of human stereo vision, *Roy. Soc. London B* **204** (1979) 301–328.

[20] R. Nevatia and K.R. Babu, Linear feature extraction and description, *Comput. Graph. Image Process.* **13** (1980) 257–269.

[21] T.J. Olson and R.D. Potter, Real-time vergence control, in: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA (1989) 404–409.

[22] R.D. Rimey and C.M. Brown, Selective attention as sequential behavior: Modelling eye movements with an augmented hidden markov model, Tech. Rept. 327, University of Rochester, New York (1990).

[23] A. Shmuel and M. Werman, Active vision: 3d from an image sequence, in: *Proceedings 10th International Conference on Pattern Recognition*, Atlantic City, NJ (1990) 48–54.

[24] G. Sperling, Binocular vision: A physical and a neural theory, *Am. J. Psych.* **83** (1970) 461–534.

[25] M. Tistarelli and G. Sandini, Robot navigation using an anthropomorphic visual sensor, in: *Proceedings IEEE Conference on Robotics and Automation*, Cincinnati, OH (1990) 374–381.