Active Surface Reconstruction by Integrating Focus, Vergence, Stereo, and Camera Calibration

A. Lynn Abbott
Virginia Polytechnic Institute
and State University
Blacksburg, Virginia USA

Narendra Ahuja
University of Illinois
Urbana, Illinois USA

ABSTRACT

This paper describes a method for estimating surfaces from stereo images. A single pair of stereo images can yield surface reconstruction for only a small volume of space. Since the scene may be large, images are obtained dynamically using different camera configurations through the control of focus and camera vergence. For smooth objects, this results in the reconstruction of surface patches that are contiguous and overlapping. These sequentially acquired surface maps are merged into a central, composite representation. During camera reconfiguration, unpredictable calibration errors may arise. Therefore, this method permits small changes in calibration parameter values to achieve agreement between adjacent reconstructed patches in the areas of overlap. To integrate the different sources of depth information, and to balance the effects they have on the resulting surface estimate, the surface reconstruction problem is formulated as one of optimization.

1. INTRODUCTION

Most passive stereo methods attempt to estimate three-dimensional (3D) scene locations from feature correspondences detected within a single pair of two-dimensional (2D) images. Due to a camera's limited field of view, and because of occlusion, it is likely that some scene points of interest are not visible in both images. It is also possible that some scene points are not clearly imaged due to the limited depth of field of the lenses. For these scene locations, it is not possible to perform accurate stereo reconstruction. This represents a limitation of stereo surface reconstruction from a single image pair.

It is therefore necessary to obtain additional images using new camera configurations and viewpoints. Such an approach is called *active* (or *dynamic*), especially when task-oriented criteria are used to select new camera configurations [1-5, 8, 10]. In addition to the extrinsic imaging parameters of camera location and orientation, intrinsic parameters such as focal length, focus, and lens aperture may also be variable and subject to algorithmic control in an active system.

This paper presents an approach to active surface reconstruction from stereo images. In this approach, a system automatically selects an unmapped area of the scene, dynamically adjusts camera parameters to obtain images for this portion of the scene (this is the *fixation* process), and synthesizes a surface estimate in the vicinity of the fixated location. Thus, the method presented here consists of the iterative execution of two steps: 1) selection of new visual targets, and 2) surface reconstruction in the target area.

Important aspects of this approach are now summarized. Depth information is extracted from the visual cues of focus, vergence, and stereo and is integrated to produce local surface maps. As surface patches are reconstructed for different parts of the surface, they are successively merged so that a global, composite map is created. When neighboring patches overlap in the composite map, the overlapping regions typically will not be in complete agreement. This is due in part to camera calibration inaccuracies, which at best can only approximately relate the internal imaging model to the physical world. Furthermore, with time and as camera configurations change, any initially calibrated parameters will represent the physical system with decreasing accuracy. This can result from such problems as gear slippage and component wear. In this approach, the system performs local self-calibration using the overlapping surface areas before and after each camera movement. Along with the visual cues of surface depth, these calibrated imaging parameters are incorporated into the overall optimization process, so that the resulting composite surface estimate represents a balanced combination of these sources of depth information. Therefore, this approach integrates image acquisition, camera calibration, and surface reconstruction into a single process which is formulated as an optimization problem.

The method described here is limited to the reconstruction of a single contiguous surface; we do not consider fixation points across object (depth) discontinuities. An extended framework for multiple object reconstruction is presented in [6].

In Section 2 we summarize the different sources of surface depth that are integrated. Section 3 presents an optimization formulation of the integration process, and in Section 4 we outline a method for simplifying the optimization problem to obtain a solution. Section 5 describes an implementation and results, and Section 6 presents a summary. Due to space limitations, many details have been omitted and may be found in [2].

2. THE INTEGRATED CUES

In this section, we briefly describe the components which are integrated to guide the process of surface estimation.

Stereo. Stereo disparity provides a powerful cue to object depth. Surface reconstruction from stereo requires in-focus images, an initial surface estimate, and the determination of correspondences. The stereo system utilized in this research is an extension of that described in [7], and imposes a hierarchical surface-smoothness constraint to guide the matching process. For a single fixation, the system produces a dense, local surface map.

This work was supported by the National Science Foundation under grant IRI-89-11942 and by the Army Research Office under grant DAAL 03-87-K-0006.

Vergence. Camera vergence refers to relative camera rotations which cause both optical axes to be aimed at a single scene location. Vergence movements may serve as cues to distance when the amount of rotation is known. The goal is to rotate the cameras until the binocular disparity is zero at the image centers.

Focus. By changing the focus setting of a lens so that image blur is minimized, it is possible to obtain an estimate of object depth. This is an attractive source for distance information, since it is monocular, having no analog to the correspondence problem of stereopsis.

Self-Calibration. The problem of camera calibration has been studied extensively, but researchers have only begun to address the problem of *self*-calibration for stereo cameras [2, 9, 11]. We view camera parameters as additional variables that need to be computed to yield the best integrated surface estimate. An initial calibration procedure provides the system with a vector $\hat{\beta}$ of calibrated imaging parameters. As each new stereo image pair is acquired, the system uses $\hat{\beta}$ to obtain an initial estimate of the surface map for this image pair. If this map overlaps the previously obtained surface map, the system permits small changes in the calibrated system parameters so that the newly obtained surface map best agrees with the overlapping composite map. This means that *new* system parameters β are estimated as a perturbation of $\hat{\beta}$. The final composite map results from the superposition of all sequentially acquired local maps. This method differs from most existing self-calibration methods in that no easily identifiable anchor points in the scene are needed.

3. OPTIMIZATION FORMULATION

In this section, we present a formulation for surface reconstruction that is intended to satisfy constraints on the surface structure imposed by the different depth cues discussed in Section 2. The problem of estimating surfaces is formulated as an optimization problem, based on an objective function which is to be minimized. This function is a linear combination of several terms (components):

$$\min_{\mathbf{q},\beta,S} \left(\lambda_c E_c + \lambda_f E_f + \lambda_v E_v + \lambda_s E_s + \lambda_p E_p + \lambda_a E_a\right) \qquad (1)$$

The coefficients λ_i determine the relative contribution of each component. The first component, E_c, is concerned with image contrast, and the next four components represent integration of different constraints for surface estimation. The last component, E_a, permits adaptive self-calibration for merging surfaces. The vector $\mathbf{q}$ represents *actuator settings*, reflecting the current state of the physical actuators (such as focus settings or camera tilt). As described earlier, β represents a set of calibrated system parameters which are updated during the optimization process. A local surface map S is synthesized by this process, and is merged with a composite map S_C. We now describe each of the components E_i. The functions $w_i(x, y)$ represent relative weights over the images.

Optimize image contrast: Without appropriate contrast, image features may be difficult to detect. This is optimal only when the lens aperture is set properly, matching the level of image irradiance to sensor sensitivity. For two cameras, the apertures can be independently controlled so that the following criterion is minimized:

$$E_c = \left| E_{c0} - \iint w_c I_L \, dxdy \right| + \left| E_{c0} - \iint w_c I_R \, dxdy \right|$$

where the constant E_{c0} is the desired average intensity level for the image region.

Minimize image blur: When an image is in sharp focus, the energy in the image intensity gradient will be maximum. In the ideal, noiseless case, the following function tends to be minimum when both images are in sharpest focus:

$$E_f = -\iint w_f \left[\| \nabla I_L \|^2 + \| \nabla I_R \|^2 \right] dxdy$$

Minimize disparity at image centers: Cross-correlation measures may be used to quantify the degree of similarity between two image regions, and may therefore be used to obtain a disparity measure for the image centers. When the following function is minimized, both cameras should be aimed at a single scene point:

$$E_v = -\frac{\left[\iint w_v I_L \, I_R \, dxdy \right]^2}{\left[\iint w_v I_L{}^2 \, dxdy \right] \left[\iint w_v I_R{}^2 \, dxdy \right]}$$

Optimize surface smoothness: When constructing a surface estimate $S(x, y)$ from stereo images, the system favors those surfaces which are smoother by penalizing high-frequency fluctuations in the surface.

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

Depth estimates from separate depth cues should agree: The cues of focus, vergence and stereo provide four different estimates of the location of the point of fixation. Typically when the system is properly fixated, the depth estimates for all visual cues should agree. The degree to which these depth estimates differ is added as a penalty within the overall objective function. Let the function δ represent the absolute difference of the reciprocal of the depths for two 3D points. If the four point estimates are denoted $\mathbf{p}_{Lf}$, $\mathbf{p}_{Rf}$, $\mathbf{p}_v$, and $\mathbf{p}_s$, for left focus, right focus, vergence, and stereo, then the agreement function is

$$E_p = \delta(\mathbf{p}_{Lf}, \mathbf{p}_{Rf}) + \delta(\mathbf{p}_{Lf}, \mathbf{p}_v) + \delta(\mathbf{p}_{Rf}, \mathbf{p}_v) + \delta(\mathbf{p}_s, \mathbf{p}_v)$$

The first term is instrumental in detecting the presence of occlusions. The second two terms are used to verify that the vergence estimate agrees with the estimates from focus. The final term compares depth from the stereo process with the vergence estimate.

Optimize calibrated imaging parameters: Any discrepancy between S (the "local" surface map) and S_C (the composite map) is used to form an error term which guides the selection of new values for β. If the area of the overlapping region is of size A, then the following term can be used as a measure of disagreement between the two estimated surfaces:

$$E_{a1} = \frac{1}{A} \iint w_{a1} \left| S - S_C \right|^2 dxdy$$

The integration is performed only where both S and S_C are defined. An additional term is needed to penalize large deviations of β from the initially calibrated values $\hat{\beta}$:

$$E_{a2} = \sum_i w_{\beta i} \left| \beta_i - \hat{\beta}_i \right|^2$$

The constants $w_{\beta i}$ determine the willingness of the system to permit changes in the calibrated values. The component for self-calibration is therefore $E_a = E_{a1} + E_{a2}$.

4. PIECEWISE OPTIMIZATION

Optimization of the objective function given in Section 3 is a complex process that involves interleaving imaging with analysis. For ease of implementation, we have decomposed the global optimization of (1) into two independent subproblems:

$$\begin{cases} \min\limits_{q_1, q_2} \; E_c \\ \min\limits_{\beta, S \mid \mathbf{q}^*} \left[\lambda_s E_s + \lambda_a E_a + \left[\min\limits_{\mathbf{q}^*} \; (\lambda_f E_f + \lambda_v E_v + \lambda_p E_p) \right] \right] \end{cases} \qquad (2)$$

The variables q_1 and q_2 represent aperture settings, and $\mathbf{q}^*$ represents the remaining actuator controls for lens control and camera orientation. This decouples aperture control from surface reconstruction. In the second equation, the focus and vergence components are first optimized to obtain fixated images, and then the local surface estimate is obtained and merged with the composite map. This simplifies the computation and results in piecewise optimization. Additional details of the algorithm are given in [2].

5. IMPLEMENTATION AND RESULTS

The algorithm was implemented with the University of Illinois Vision System (Figure 1) [2]. This system acquires stereo images from cameras which can tilt, pan, and verge under computer control. The host can also control lens settings of focus, aperture, and zoom. For the results shown here, the system performed several fixations on a single cylindrical object.

First fixation. The system performs an exploratory fixation sequence, during which focus ranging and camera rotations are used to aim the cameras at a single object point. This involves minimization of the terms E_c, E_f, E_v, and E_p in (2). A segmentation process then uses changes in focus to retain image areas only for the object portions which lie within the depths of field of the cameras. The stereo module is then invoked, and depth maps are obtained using coarse-to-fine reconstruction. The resulting surface map is shown in Figure 2.

Second fixation. The next step in the automated mapping process is to select a new fixation point from that map. The goal is to select a new target so as to smoothly extend the scene description. In this case a target is chosen along the lower edge of the surface map. The system now fixates a surface location in the vicinity of this target, again using focus and vergence information. Distance information from the previous fixation is used where possible to assist in the construction of the second local map. Features from the previous images are not used. The resulting combined surface for the first two fixations is shown in Figure 3. The mean-squared difference between the two overlapping regions is 0.000028 m^2.

If the system were not permitted to utilize the calibration parameters β in the optimization of (2), then the combined surface from the two fixations appears as in Figure 4. The mean-squared error for the overlapping regions of the two maps is now 0.000071 m^2. This is a degradation of a factor of 2.54 compared to the case with optimization.

Subsequent fixations. The algorithm now continues, automatically selecting scene targets, fixating, and incrementally building a composite surface map. For each fixation, depth estimation from focus, vergence, and stereo is integrated with aperture control and camera calibration to yield a smooth composite surface estimate. Figure 5 shows the state of the composite map after 8 fixations.

6. SUMMARY

This paper has described a vision system which integrates target selection, image acquisition, surface reconstruction, and camera calibration to maintain an evolving, global, composite surface map. The system automatically scans a scene area to build a surface estimate for a single, contiguous scene region.

REFERENCES

[1] A. L. Abbott and N. Ahuja, "Surface Reconstruction by Dynamic Integration of Focus, Camera Vergence, and Stereo," *Proceedings: Second International Conference on Computer Vision*, pp. 532-543, Dec. 1988.

[2] A. L. Abbott, "Dynamic Integration of Depth Cues for Surface Reconstruction from Stereo Images," Ph.D. Dissertation, University of Illinois, 1990.

[3] R. Bajcsy, "Active Perception vs. Passive Perception," *Proceedings: Workshop on Computer Vision*, pp. 55-59, Oct. 1985.

[4] D. H. Ballard and A. Ozcandarli, "Eye Fixation and Early Vision: Kinetic Depth," *Proceedings: Second International Conference on Computer Vision*, pp. 524-531, Dec. 1988.

[5] J. J. Clark and N. J. Ferrier, "Modal Control of an Attentive Vision System," *Proceedings: Second International Conference on Computer Vision*, pp. 514-513, Dec. 1988.

[6] S. Das and N. Ahuja, "Integrating Multiresolution Image Acquisition and Coarse-to-Fine Surface Reconstruction from Stereo," in *Proceedings: IEEE Workshop on Interpretation of 3D Scenes*, Austin, TX, pp. 9-15, Nov. 1989.

[7] W. Hoff and N. Ahuja, "Surfaces from Stereo: Integrating Feature Matching, Disparity Estimation and Contour Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-11, pp. 121-136, Feb. 1989.

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

[9] P. Liang, Y. L. Chang, and S. Hackwood, "Adaptive Self-Calibration of Vision-Based Robot Systems," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 4, July/Aug. 1989.

[10] A. Shmuel and M. Werman, "Active Vision: 3D from an Image Sequence," *Proceedings: 10th International Conference on Pattern Recognition*, pp. 48-54, June 1990.

[11] H. Takahashi and F. Tomita, "Self-Calibration of Stereo Cameras," *Proceedings: Second International Conference on Computer Vision*, pp. 123-128, Dec. 1988.

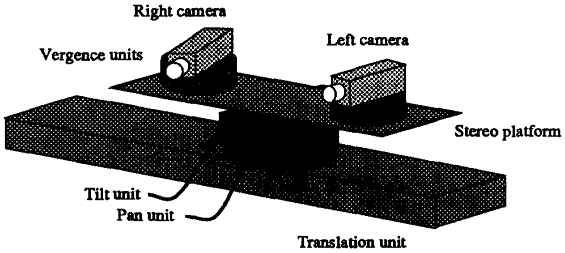

Figure 1. A dynamic stereo camera system. Vergence units rotate the cameras independently. A stereo platform supports these units, and can undergo tilt, pan, and translation movements.

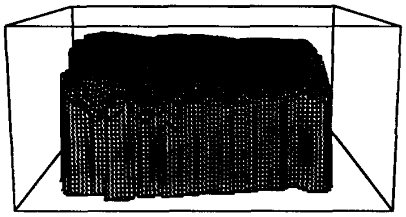

Figure 3. Composite surface map after second fixation (view from left side). The local map from the first (second) fixation is on the left (right) side of the figure. Where the two maps overlap, the minimum depth value is shown.

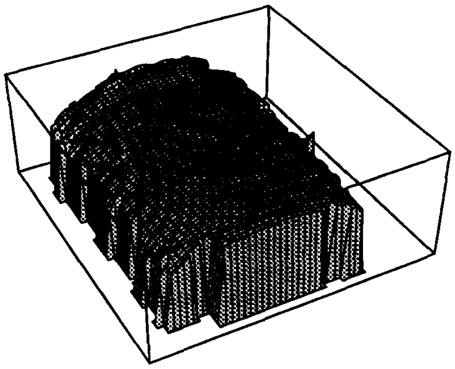

Figure 5. Composite surface map after eight fixations. This map is incrementally updated after each fixation. When overlapping areas are present for subsequent fixations, the mean depth values are retained.

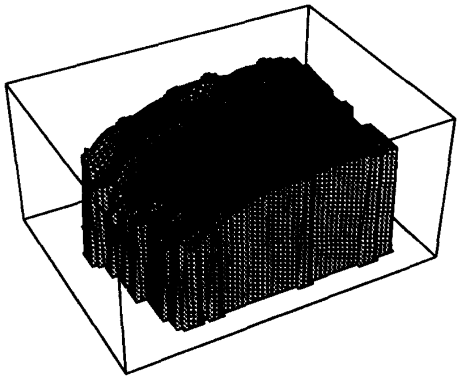

Figure 2. Surface map from first fixation, resolution level 256 x 256. Range values are referenced to the coordinates of the left image. Points on the object which are nearest the camera are on the right side of the figure.

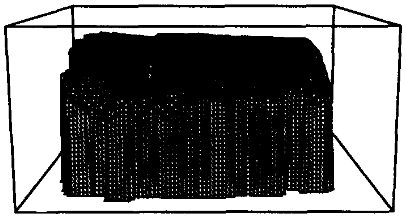

Figure 4. Merge of first two surface maps without optimization. The second map has been constructed without any knowledge of the first. A more pronounced seam is visible where the two maps join.