Dynamic integration of visual cues for position estimation

Subhodev Das and *Narendra Ahuja*

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

ABSTRACT

Three-dimensional (3D) position estimation using a single passive sensor, particularly vision, has frequently suffered from unreliability and has involved complex processing methods. Past research has combined vision with other active sensors in which the emphasis has been on data fusion. This paper attempts to integrate multiple passive 3D cues - camera focus, camera vergence and stereo disparity - using a single sensor. We argue that in the active vision paradigm an estimate of the position is obtained in the process of fixation in which the imaging parameters are dynamically controlled to direct the attention of the imaging system at the point of interest. Fixation involves integration of the passive cues in a mutually consistent way in order to overcome the deficiencies of any individual cue and to reduce the complexity of processing. Taking into account their reliabilities, the individual position estimates from the different cues are combined to form a final, overall estimate.

1 INTRODUCTION

Visual processing has an important role in the functioning of an intelligent autonomous system. Sensing three-dimensional structure is necessary for planning navigation and end-effector motion. Traditionally, non-contact active sensors such as, ultrasound, laser, collimated light, have been preferred in spatial planning over more economical passive sensors. The reason being three-dimensional position estimation using passive sensors has frequently suffered from unreliability and has involved complex processing methods. Inaccuracy may primarily be attributed to the limitations of the sensors as real devices interacting with a real world and greater sophistication of processing can hardly serve as an antidote. Most of the past algorithms that have used the camera as single passive sensor have utilized a single visual cue such as, stereo disparity, camera focus, or camera vergence, for extracting 3D information. Thus they all suffer from the limitations of a single sensor. Multiple sensors or visual cues working conjunctively, however, can reduce this problem by providing large sets of competitive data whose consistency can be enforced via mutual constraints.

The restrictive nature of a single sensor, the deficiencies of a single visual cue employing such a sensor in particular, has prompted many researchers in the past to combine vision with a number of different active sensors. For example, touch and stereo [2], acoustic and vision [10], laser and vision [17], and thermal and vision [13]. The emphasis in these approaches has been more on *different* sensor modalities than on cooperative interaction among the sensors. Some efforts [9,19] have concentrated on modeling biological mechanisms of interactions among vergence, accommodation (focus) and stereopsis. Others have advocated active, intelligent data acquisition [3,4] while using a single sensor. Computational active vision has become more feasible in the recent years with the availability of sophisticated hardware for controlling imaging elements [1,5,6,7,11]. Thus it has been possible for a sensing device such as, a stereo camera, to apply different visual cues while operating in the single mode and to have intra-modal cooperation among these cues that is difficult to achieve with multiple disparate sensors operating in different modes. Examples of such intra-modal cooperation are stereo and focus [11], and stereo, focus an vergence [1].

This paper is concerned with the dynamic integration of camera focus, camera vergence and stereo disparity for estimating the position of a 3D point. We present an approach to acquiring structural information in the vicinity of a selected scene point at increasing resolution and using that information for coarse-to-fine control of imaging

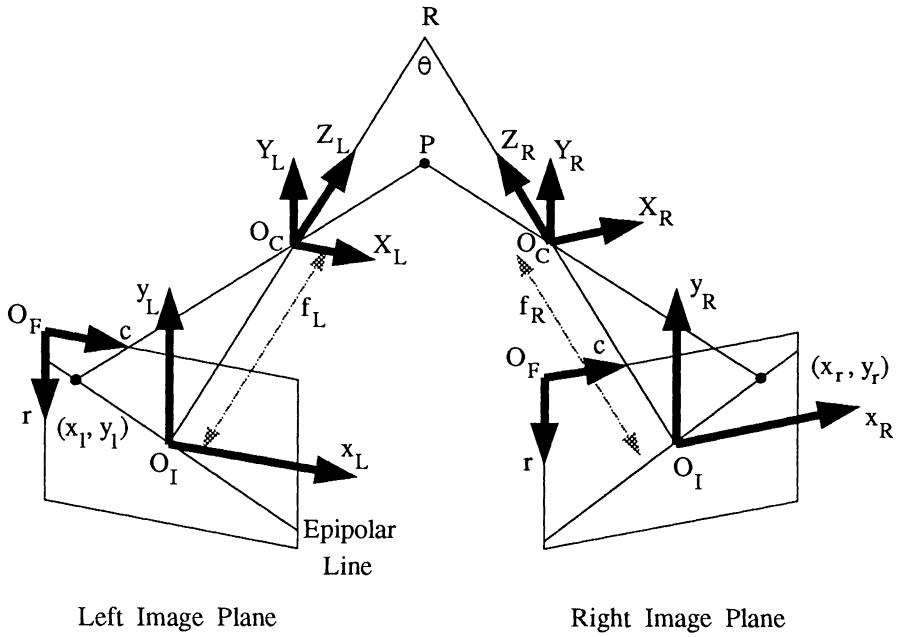


Figure 1: A stereo geometry in which the two optic axes and the baseline are coplanar. The cameras cannot rotate around their optic axes.

parameters to aim the cameras at the point. After the cameras have *fixated* at the point its depth estimates from focus, vergence and stereo cues are combined to form a final, overall estimate for the point.

Section 2 discusses stereo, focus and vergence as independent sources of depth information, and the accuracies of the depth estimates derived from them; knowing their individual performances is necessary for combining them. Section 3 describes in greater detail the motivation behind the work reported in this paper. Section 4 presents an algorithm that performs two types of integration. First, it interleaves coarse-to-fine control of the cameras with the acquisition of coarse-to-fine structural information about the point. Second, it achieves fusion of depth information available from different sources of depth to derive a single accurate estimate. Section 5 gives details of implementation and the experimental results. Section 6 presents concluding remarks.

2 STEREO, FOCUS, AND VERGENCE AS 3D CUES

The binocular cues of stereo disparity and vergence and the monocular cue of focus have long been recognized as important sources of 3D information. In this section we will be discussing how each of them is used and what its limitations are.

2.1 Stereo

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

A stereo camera configuration is shown in Figure 1, in which θ denotes the angle between the two optic axes i.e., the vergence angle. The 3D coordinate systems of the left and right cameras are related by the transformation

$$\mathbf{X}_R = R\mathbf{X}_L + \mathbf{T} \tag{1}$$

where $\mathbf{X}_i = [X_i, Y_i, Z_i]^T$, R is a 3×3 orthonormal rotation matrix determined by θ, and $\mathbf{T}$ is a 3×1 translation vector. Using the homogeneous coordinate representation, R and $\mathbf{T}$ are combined into a single 4×4 matrix A and the same left to right transformation is represented as

$$\begin{bmatrix} s\mathbf{X}_R \\ s \end{bmatrix} = {}^R A_L \begin{bmatrix} s\mathbf{X}_L \\ s \end{bmatrix}, \tag{2}$$

where s is an arbitrary scale factor. Notationally, (2) is $\mathbf{X}_R^{(h)} = {}^R A_L \mathbf{X}_L^{(h)}$ and ${}^j A_i$ is the transformation of the coordinates of a point from reference frame i to reference frame j. A 2D image is digitized and stored in the computer frame memory. Given an object point $P = \mathbf{X} = (X, Y, Z)$ in the camera 3D coordinate system $XYZO_C$ (XYZ axes with origin O_C), let $\mathbf{r} = (r, c)$ denote the row and column number of a pixel in the frame coordinate system rcO_F. Using homogeneous coordinates, the transformation from the camera 3D coordinates to the frame coordinates is then $\mathbf{r}^{(h)} = {}^F A_C \mathbf{X}^{(h)}$, upto a scale factor s.

To obtain depth requires selecting feature points (mostly detected intensity edges) from the two images and matching points belonging to one image with those from the other. The search in the right image for the match of a given pixel in the left image (say $\mathbf{r}_L$) can be restricted to a unique line called the *epipolar line*, determined by the camera geometry. Let $\mathbf{r}_R$ be the unique match in the right image. Then the coordinates of the corresponding 3D point can be computed from the homogeneous transformation equations for the left and right pixels

$$\mathbf{r}_L^{(h)} = {}^F A_C^L \, \mathbf{X}_L^{(h)}, \qquad \mathbf{r}_R^{(h)} = {}^F A_C^R \, \mathbf{X}_R^{(h)} \tag{3}$$

when these equations are related by (2).

2.1.1 Accuracy of range estimation from stereo

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

The relationship between the errors in stereo camera geometry and the estimated range has been widely investigated [18,20]. Because of the discrete nature of the image plane and frame memory coordinates, point projections can be expressed in terms of pixels only. This quantization affects the coordinate values by a maximum of $\Delta r_{qnt} = \Delta c_{qnt} = \pm 1/2$ pixel.

Non-sharp image features are often poorly localized. If the blurring function is modeled as a Gaussian then the location uncertainties are greater for larger values of the spread of the Gaussian. When features are matched, this location error $\Delta \mathbf{r}_{loc}$ leads to 3D location error.

If $\Delta \mathbf{r} = \Delta \mathbf{r}_{qnt} + \Delta \mathbf{r}_{loc}$ denotes the uncertainty in pixel coordinates due to quantization and feature localization then the discrepancy between the location of a 3D point computed from

$$\mathbf{r}_L^{(h)} + \Delta \mathbf{r}_L^{(h)} = {}^F A_C^L \, \mathbf{X}_L^{(h)}, \qquad \mathbf{r}_R^{(h)} - \Delta \mathbf{r}_R^{(h)} = {}^F A_C^R \, \mathbf{X}_R^{(h)} \tag{4}$$

and that from (3) is the error due to the stereo model.

2.2 Focus

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

The cone of light rays emanating from an arbitrary point object is brought to focus at a point in the image plane by an ideal lens. For a lens embedded in a uniform medium, the lens equation relating the distance u of an object point on the optic axis of the lens to the distance v of the corresponding image point is

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f}, \tag{5}$$

where f is the focal length of the lens. For such a point, the object distance can be computed using the above equation if v and f are known. The distance v is changed by adjusting the focus setting (denoted by p). This directly affects the degree of image blur for objects in the field of view. The distance v is linearly related to p; the exact relation between u and p is determined empirically.

An object point which is not in focus is imaged as an ellipse (for an object point on the optic axis, the projection is a circle known as the *blur circle*), which is the intersection of the cone of light rays and the image plane perpendicular to the optic axis. Such imaging amounts to low-pass filtering of the ideal (point) image. Let v_0 be the position of the image plane when an object point O at a distance u_0 is brought to focus at I. If the image

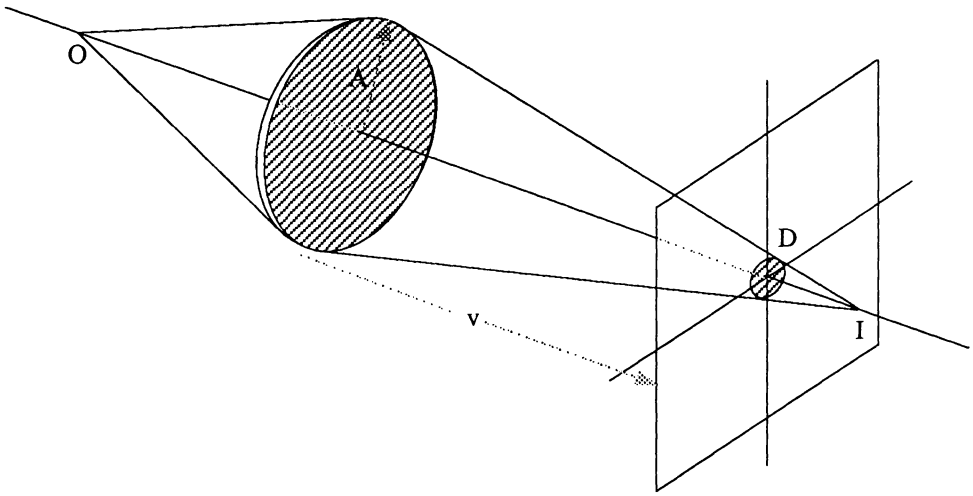


Figure 2: Defocusing of a point by an ideal lens. A is the aperture diameter of the lens.

plane is now displaced to a new position v, then from Figure 2 the diameter of the blur circle in the image plane at position v is

$$D = \frac{A}{v_0}|v - v_0| = \frac{Af}{u_0}\left(\frac{|u - u_0|}{u - f}\right). \tag{6}$$

The blur circle increases as the distance offset $|u - u_0|$ from the focused point increases. The intensity within the blur circle is non-uniform, diminishing towards the edge of the circle. It approaches a relatively constant value when large amount of defocusing is present. If the magnitude of the intensity distribution within the circle is modeled by a 2D Gaussian then the spread parameter σ_l of the Gaussian signifies the degree of optical blurring of the defocused point. The parameter σ_l and the diameter of the blur circle are related by

$$\sigma_l = kD, \quad k > 0 \tag{7}$$

where the constant of proportionality, k, is a characteristic of the imaging system. Thus σ_l is proportional to the focal length, aperture and the distance of the defocused point from the focused point.

2.2.1 Accuracy of range estimation from focus

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

Let us assume that for any image the differences in sharpness cannot be distinguished for blur circles with diameter smaller than C, referred to as the *circles of confusion*. For a given location of the image plane v_0, corresponding to object distance u_0, the projections of all objects located in the interval $[u_2, u_1]$ appear equally sharp. This interval is the depth of field. The expressions for the near and far extremes of the depth of field are derived from (6):

$$u_1 = \frac{u_0 Af}{Af + C(u_0 - f)}, \qquad u_2 = \frac{u_0 Af}{Af - C(u_0 - f)}. \tag{8}$$

The depth of field is

$$\Delta z = u_2 - u_1 = \frac{2ACu_0 f(u_0 - f)}{A^2 f^2 - C^2(u_0 - f)^2}. \tag{9}$$

2.3 Vergence

Vergence mechanism is necessary for a stereo-camera system to direct both the cameras at the same world point. Once the cameras have verged on to a point it is possible to compute the 3D location of the point from the knowledge of the vergence angle.

Vergence is a special case of stereo in the sense that it can provide 3D information about one particular point in the visual field viz., the point where the optic axes of the two cameras intersect. Referring to Figure 1 this point is R which is also known as the *point of vergence*. When the point of vergence lies on some physical surface it is referred to as the *point of fixation*. Following Figure 1, R projects at the center of each image plane; it is therefore represented as $\mathbf{x} = 0$ in image coordinates and $\mathbf{r} = (r_0, c_0)$ in frame coordinates. The 3D coordinates of R are obtained by the following substitutions: $\mathbf{r}_L = (r_{0L}, c_{0L})$ and $\mathbf{r}_R = (r_{0R}, c_{0R})$, in (3).

The fundamental problem in using vergence as a depth cue is to ensure that the two cameras are in fact aimed at the same 3D point. A common approach to this problem is to estimate the disparity between the left and right image centers and to use vergence movements to null out this disparity [15]. Another approach is to shift one image with respect to the other until the overlapping regions are most similar. This later approach is known as translational image registration. The most commonly employed similarity criterion functions are minimum-distance and cross-correlation. In the cross-correlation approach the following similarity measure is maximized:

$$d(s,t) = \int\int I_L(x,y) I_R(x+s, y+t) dxdy \tag{10}$$

where I_L and I_R are the left and right images, and s and t are offset of the right image with respect to the left along the x- and y-directions. The integration is performed over some selected window.

2.3.1 Accuracy of range estimation from vergence

The method of translational registration works best when the cameras are oriented such that the same scene area approximately projects at the two image centers. Otherwise, the sample windows used for comparison need to be selected carefully. If the sample windows are too large they are likely to contain objects at different depths, in which case occlusion may prevent fixation of scene points and results may be ambiguous. On the other hand, if the windows are too small, two problems may arise. First, they may not contain enough detail to provide adequate basis for similarity comparison. Second, they may not overlap at all, in which case no criterion function evaluation will be meaningful. For close-range applications, registration methods can fail when the surface gradient is sufficiently large relative to the image planes.

Given that the two image centers are correctly aligned such that the computation of depth of the fixation point is possible, the accuracy of the vergence-based position estimate is limited by the precision of the imaging system in setting up the correct vergence angles to aim at the fixation point. The error in the vergence can be estimated from the difference between the depth computed by backprojecting the left and right features that are matched by the registration process and the depth of the fixation point at which the two optic axes actually intersect after aiming the cameras at the backprojected point.

3 MOTIVATION

The estimation of the 3D location of a point may be possible with any one of the passive visual cues of stereo disparity, focus or vergence, operating independently. However, each of these visual processes has its own requirements and restrictions. The accuracy of depth estimates from focus is limited by the depth of field effect of the lens which increases as object distance increases, as the aperture diameter decreases, and as the focal length is reduced. Besides, the search for the best focus setting to register the sharpest image of an object point is slow because of the large range of axis settings that needs to be examined. The process of range estimation from stereo disparity is most robust and least complex only when an estimate of the disparity value for the object point is available. Further, since the stereo analysis depends on the locations of features and their correspondences, it is necessary that the features appear sharp and well localized. Features are sharp only for objects within the depth of field of the lens. It is required that the optic axes of two verging cameras intersect at an object point in order to compute the depth of that point from vergence. To fulfill this requirement the object point must be visible to both cameras. The translational image registration method to align the two image centers as described above cannot guarantee a correct or unique result. Failures are likely when the surface gradient near the object point is sufficiently large relative to the image planes, or when the image contains periodic patterns or insufficient details.

In the active vision paradigm the method of position estimation of a 3D point is one of *foveation* that brings the point into attention of the visual system at a high resolution. As mentioned above, none of the visual depth cues can reliably extract structural information if used alone. However, the strengths and shortcomings of these cues are mutually complementary. Therefore, a cooperation between them may overcome many of their individual shortcomings. An active control of the imaging parameters makes such a cooperative integration possible. The foveation (or fixation) of a scene point is attempted using the largest available focal length to provide the highest resolution and to maximally exploit the focus cue. Since the search for the best focus axis setting must be initiated at the smallest axis setting for each new object point, the large focal length of the lens causes significant blurring of any defocused point at the onset of the search process. The defocused images from the two cameras may be stereo analyzed to obtain a coarse depth estimate for the center of one of the cameras (say left). This coarse estimate may be used to set up the approximate vergence angle of the two cameras and to predict the best focus setting. As image planes are reconfigured for this axis setting intermediate images are obtained with decreasing blur which may be continuously stereo analyzed to improve the estimate for the object point. In this process, the computational blurring operation is replaced by instantaneous optical blurring. The number of stereo pairs acquired before fixation is achieved would depend on the amount of image plane configuration required; the larger the amount of this reconfiguration, the greater would be the opportunity to acquire intermediate resolution images. At the end of the fixation process a reliable depth estimate of the fixated point is obtained from each of the different cues.

The use of stereo, focus and vergence processes motivates fusion of depth information from all of them, thus improving the accuracy of the final estimate. Stereo-based estimates have errors determined by the feature location and quantization errors. The focus-based depth estimates on the other hand suffer from the depth of field effect of the lens. Finally, the accuracy of the vergence-based estimate is limited by the angular quantization of the camera positioners. Using the uncertainty characteristics of each of these sources that are discussed in Section 2, the estimates may be combined using an optimal estimator.

4 ALGORITHM

In this section we describe an algorithm to achieve the desired integration of the processes of stereopsis, camera focusing and camera vergence in fixating a 3D point on the surface of a physical object, followed by the fusion of depth estimates from these different visual cues in estimating the position of the point. We assume that the direction of gaze for the cameras or the direction in which the target object point lies, is specified initially. A stereo camera system can be attended to a selected part of the scene by executing a sequence of controlled movements that are learned and generated based on an explicit model for such sequences [16] or by obtaining coarse structural information about the selected scene area [8]. The availability of the gaze direction allows the selection of camera orientations such that the target point, though defocused, is visible to both viewpoints. Additionally, we assume that one of the cameras, the left, is the dominant one and the target point projects at its image center.

4.1 Integration of Depth Cues in Fixation

In order to fixate the object point, the point has to be brought into sharp focus using the largest available focal length, f_{max}, for the lens and the stereo cameras have to be oriented such that the estimated point projection falls at the center of each image. Initially, the depth of the point in the 3D world is unknown, therefore the search for the best focus axis setting has to be initiated at the smallest available axis setting, p_1. The selection of the large focal length causes significant blurring of the corresponding image point. The magnitude of the image blur σ_{lf}^1 is, however, unknown at this stage because of the unknown depth of the scene point. We define the *window of fixation* as a $w \times w$ window centered at the image center of each camera within which the focus criterion function is evaluated. The defocused window of fixation in the left image is subsampled coarsely using an $H_1 \times H_1$ grid, where H_i denotes the resolution of the window of fixation at the ith stage ($\sigma_{lf} = \sigma_{lf}^i$) during reconfiguration. The defocused right image is also subsampled using the same $H_1 \times H_1$ grid. Features are detected in the subsampled left window of fixation and the right image using the Nevatia-Babu line extraction algorithm [14]. These features are matched using the stereo geometry of the current camera configuration, and a surface patch is fit to the resulting 3D points. A coarse depth estimate, Z_s^1, of the scene point that is projected at the left image center is interpolated

from this patch. This depth estimate is now used to compute

$$\sigma_{lf}^1 = kD = k\frac{Af_{max}}{Z_f^1}\left(\frac{\left|Z_s^1 - Z_f^1\right|}{Z_s^1 - f_{max}}\right) \tag{11}$$

where $Z_f^1 = u_1$ is the object distance corresponding to p_1.

Once a depth estimate for the target point has been obtained the cameras can be reoriented such that the two optic axes actually intersect at this estimated 3D point. 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 uses a $w_v^1 \times w_v^1$ subimage of the full-resolution right image around the right image center to evaluate the similarity criterion function which is a normalized version of (10):

$$d^2(s,t) = \frac{\left[\sum\sum I_L(r,c) I_R(r+s, c+t)\right]^2}{\left[\sum\sum {I_L}^2(r,c)\right]\left[\sum\sum {I_R}^2(r+s,c+t)\right]}. \tag{12}$$

The projections of the points $(X_s^1, Y_s^1, Z_s^1 - \Delta_s^1)$ and $(X_s^1, Y_s^1, Z_s^1 + \Delta_s^1)$ (Δ_s^1 being the error in Z_s^1) in the right image determines the $w_v^1 \times w_v^1$ registration window. To do finer registration needs further interleaving of the processes of vergence registration and depth estimation from finer stereo.

As the image planes are gradually reconfigured by changing the focus settings, the new target point becomes less and less blurred. The two windows of fixation in each pair of optically blurred images are subsampled, reducing the degree of subsampling as images become less blurred (σ_{lf} decreases) according to $H_i/H_{i+1} = \sigma_{lf}^{i+1}/\sigma_{lf}^i$. 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 - larger the blur smaller is the spatial frequency - thus lowering the sampling rate. 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 $\Delta\sigma_{lf}^i = \sigma_{lf}^i - \sigma_{lf}^{i+1}$ is significant. Let $\sigma_{lf}^i/2$ be the chosen significant value of $\Delta\sigma_{lf}^i$. The intermediate images in which the blur of the target point is between σ_{lf}^i and $\sigma_{lf}^{i+1} = \sigma_{lf}^i - \Delta\sigma_{lf}^i = \sigma_{lf}^i/2$ are skipped for stereo analysis. 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_{lf}. The improved estimate ${Z_s}^{i+1}$ of the target point aims the cameras at the point with increasing precision. This process of coarse-to-fine image acquisition and depth estimation from stereo interleaved with vergence registration is continued till $\Delta\sigma_{lf}^i$ reaches a lower bound on $\Delta\sigma_{lf}$, $\Delta\sigma_T$, which corresponds to the smallest error in stereo disparity that can be determined experimentally, viz., error of one pixel. Beyond this stage only coarse-to-fine image acquisition is continued until $\sigma_{lf} = 0$, when the cameras are approximately focused. The final focus axis setting is p_m. These focused images are stereo analyzed using the finest resolution grid. At this stage reliable depth estimates of the target point are obtained from stereo (Z_s) and vergence (Z_v) using the final camera configuration.

In order to focus the cameras accurately, the depth estimate Z_s of the target point is used to establish an interval of focus axis settings $[p_{1m}, p_{2m}]$ symmetric about p_m ($p_{1m} < p_m$ and $p_{2m} > p_m$). This interval is finely quantized and searched for a peak of the focus criterion function evaluated over the window of fixation in each image. The best focus axis settings p_l and p_r corresponding to the peaks in the left and right images, respectively, yield focus-based depth estimates Z_{f_L} and Z_{f_R} for the target point. This completes the fixation of the point and the integration of the processes of stereo, focus and vergence.

4.2 Fusion of Depth Estimates

We first describe an error model associated with each observation process and a model for combining these observations. To begin with, let us assume that each estimate Z_i is the observed value of a random variable z_i that has a Gaussian distribution $N(Z, \sigma_i^2)$, where Z is the *true* depth of the fixation point and $\sigma_i^2 = E[(z_i - Z)^2]$. Thus the variance σ_i^2 is proportional to the observation error $\Delta_i = |Z_i - Z|$. Let $\mathbf{z} = [z_1, z_2, \ldots, z_N]^T$ be an N-dimensional random vector whose elements are statistically independent. We model the fusion problem as one of *estimating* the parameter Z from the observations $\mathbf{Z} = \{Z_i\}^T$. We also assume that the probability density for

the mapping from parameter space to observation space is

$$p(\mathbf{Z}|Z) = [(2\pi)^{N/2}|\mathbf{\Lambda}|^{1/2}]^{-1} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\frac{(Z_i - Z)^2}{{\sigma_i}^2}\right]. \tag{13}$$

The covariance matrix $\mathbf{\Lambda}$ is a diagonal matrix whose elements are ${\sigma_i}^2$. Since Z is *not* a random variable an unbiased estimate for it is the *maximum likelihood estimate* $\hat{Z}_{MLE}$, at which the *likelihood function* $\ln p(\mathbf{Z}|Z)$ is a maximum. Thus

$$\left.\frac{\partial \ln p(\mathbf{Z}|Z)}{\partial Z}\right|_{Z=\hat{Z}_{MLE}} = 0. \tag{14}$$

Substituting $p(\mathbf{Z}|Z)$ from (13) we obtain,

$$\left.\frac{\partial}{\partial Z}\left[\frac{1}{2}\sum_{i=1}^{N}\frac{(Z_i - Z)^2}{{\sigma_i}^2} + \ln[(2\pi)^{N/2}|\mathbf{\Lambda}|^{1/2}]\right]\right|_{Z=\hat{Z}_{MLE}} = \left.\sum_{i=1}^{N}\frac{Z_i - Z}{{\sigma_i}^2}\right|_{Z=\hat{Z}_{MLE}} = 0$$

or

$$\hat{Z}_{MLE} = \frac{\sum_{i=1}^{N}\frac{Z_i}{{\sigma_i}^2}}{\sum_{i=1}^{N}\frac{1}{{\sigma_i}^2}}. \tag{15}$$

Fusion of stereo and focus data using this model has also been reported in [12].

Here, the four depth estimates Z_s, Z_v, Z_{f_L} and Z_{f_R} are the observed values of the random variables $z_s(\sim N(Z,\sigma_s^2))$, $z_v(\sim N(Z,\sigma_v^2))$, $z_{f_L}(\sim N(Z,\sigma_{f_L}^2))$ and $z_{f_R}(\sim N(Z,\sigma_{f_R}^2))$, respectively. Since each observation process z_i outputs a single data and the sensor model is non-stationary it is difficult to obtain exact values of the σ's either on-line or by off-line calibration; the uncertainties in stereo (Δ_s), vergence (Δ_v) and focus-based estimates (Δ_L, Δ_R) may be used instead. We observe that the stereo estimate is interpolated from the depth values computed at feature locations in the neighborhood of the fixation point. Thus, while obtaining depth values at the neighborhood features from stereo, error estimates due to image plane quantization are also obtained at these locations. These error estimates are then interpolated to the image center to obtain Δ_s. The value of Δ_v is estimated from the difference between the depth of the 3D point obtained from the stereo pair matched by translational registration and that of the 3D point where the two optic axes have actually intersected. Finally, the values of Δ_{f_L} and Δ_{f_R} are estimated from the depth of field of the lens during the Z_{f_L} and Z_{f_R} measurements. The hypothesis that all the distributions have the same mean is verified by doing a pairwise comparison of the random variables z_i and z_j in the following way: the random variable $x = z_i - z_j/\sqrt{\sigma_i^2 + \sigma_j^2}$ is tested to be $\sim N(0,1)$. This is done by checking that $|X| \leq X_\alpha$ for $(1-\alpha)$ degree of confidence, where X is the observed value of x:

$$X = \frac{Z_i - Z_j}{\sqrt{\Delta_i^2 + \Delta_j^2}}. \tag{16}$$

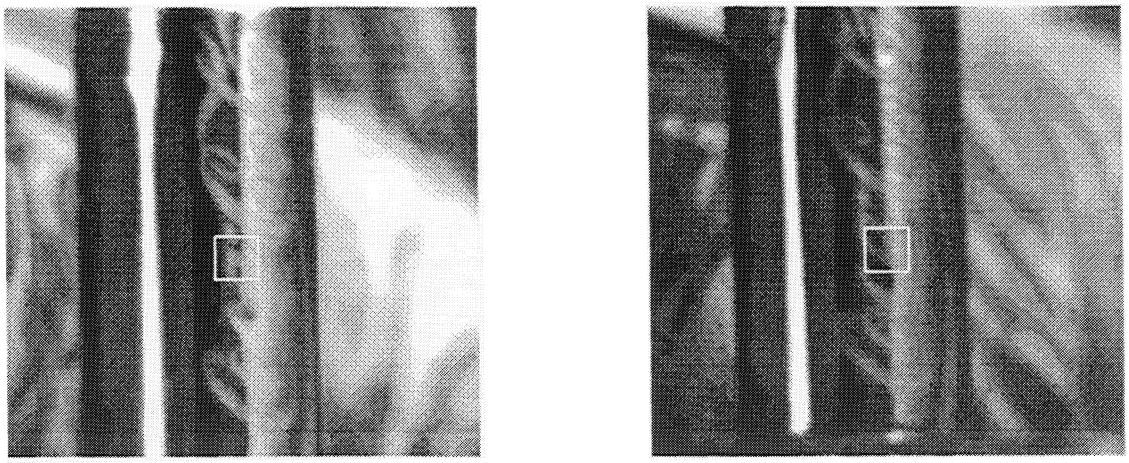
If the hypothesis is not acceptable then the estimate with the smallest error is used as the final estimate. If the hypothesis is valid, the best (maximum likelihood) estimate of Z is obtained from (15) by weighted averaging:

$$\hat{Z} = \frac{\Delta_v^2\Delta_{f_L}^2\Delta_{f_R}^2 Z_s + \Delta_s^2\Delta_{f_L}^2\Delta_{f_R}^2 Z_v + \Delta_s^2\Delta_v^2\Delta_{f_R}^2 Z_{f_L} + \Delta_s^2\Delta_v^2\Delta_{f_L}^2 Z_{f_R}}{\Delta_v^2\Delta_{f_L}^2\Delta_{f_R}^2 + \Delta_s^2\Delta_{f_L}^2\Delta_{f_R}^2 + \Delta_s^2\Delta_v^2\Delta_{f_R}^2 + \Delta_s^2\Delta_v^2\Delta_{f_L}^2}. \tag{17}$$

Finally, $\hat{Z}$ is used to compute the X and Y coordinates in the reference coordinate system.

5 IMPLEMENTATIONS AND RESULTS

In this section we present details and results of implementing our position estimation algorithm on a dynamic imaging system. The system consists of two Cohu 4815 CCD cameras mounted on a stereo platform and equipped with Vicon V17.5-105M motorized zoom lenses. High-precision stepper-motor rotational units are used to control independent pan, tilt and vergence angles. The imaging system is controlled by a Sun Microsystems 3/160 workstation.



(a) (b)

Figure 3: 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.

5.1 Implementation Details

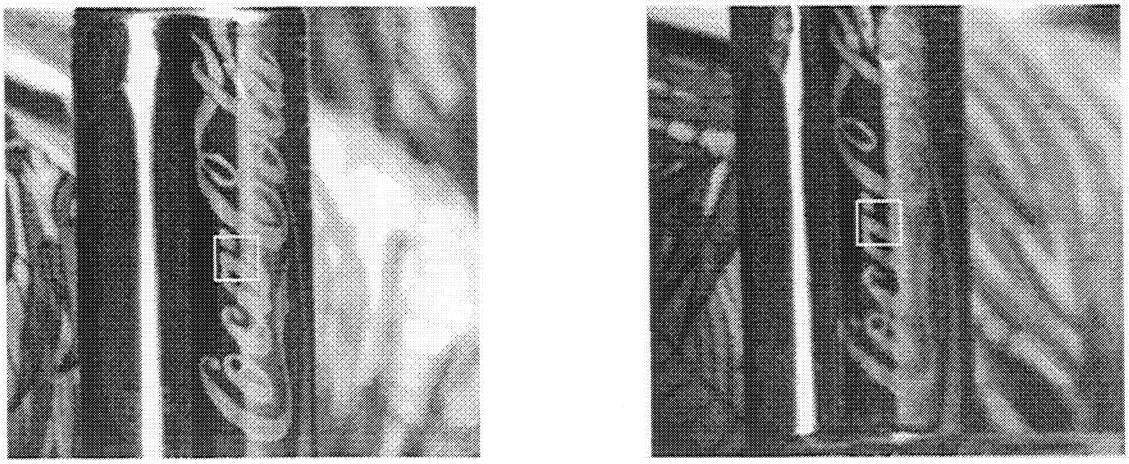
For the left and the right cameras focal lengths (calibrated) of $f_{max} = 105.4$ mm and 101.0 mm are used in the fixation process. The smallest resolution used in the coarse-to-fine stereo reconstruction of the target area is $H_1 = 64$ and the window of fixation is $w \times w = 48 \times 48$. It is experimentally observed that an increment of $\Delta\sigma_T = 3$ leads to a maximum of one pixel increase in stereo disparity error. Empirically, the relation between object distance u and focus setting p is determined to be $u = f(ap + b + f)/(ap + b)$ for a zoom setting f. The calibrated parameters of this expression 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 0.5 and 0.67, respectively. A threshold of $X_\alpha = 1.96$ for an $\alpha = 0.05$ significance level is used to test the pairwise consistency of data.

5.2 Experimental Results

The dynamic imaging system is coarsely aimed at a coke can. The 3D coordinates (hand-measured) of a point on the can that is projected at the center of the left image are $(X, Y, Z) = (0.584, 0.343, 1.816)$. The coordinate values are all in meters and expressed in a reference world coordinate system. The initial stereo images in which the target point is defocused are shown in Figure 3. (The original images are of size 512×512 while the displayed images are 256×256.) Fixation begins with the initial focus axis setting of $p = 0$. The results of successive image plane reconfiguration by changing the focus axis setting and coarse-to-fine stereo analysis to fixate the target point are summarized in the following tables:

Table I: Initial configuration, $i = 1$

focus axis		focused depth		blur	sampling	stereo		registration	vergence	
						depth	error	window	depth	error
p_L	p_R	Z_{fL} (m.)	Z_{fR} (m.)	σ_{lf}	H	Z_s (m.)	Δ_s (m.)	w_v (pixels)	Z_v (m.)	Δ_v (m.)
0	0	1.644	1.640	8	64	1.818	0.092	30	1.818	0.007



(a) (b)

Figure 4: 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.

Table II: Second reconfiguration, $i = 2$

focus axis		focused depth		blur	sampling	stereo		registration window	vergence	
						depth	error		depth	error
p_L	p_R	Z_{fL} (m.)	Z_{fR} (m.)	σ_{lf}	H	Z_s (m.)	Δ_s (m.)	w_v (pixels)	Z_v (m.)	Δ_v (m.)
1465	1514	1.765	1.764	4	128	1.821	0.030	10	1.823	0.007

Table III: Final reconfiguration, $i = 3$

focus axis		focused depth		blur	sampling	stereo		registration window	vergence	
						depth	error		depth	error
p_L	p_R	Z_{fL} (m.)	Z_{fR} (m.)	σ_{lf}	H	Z_s (m.)	Δ_s (m.)	w_v (pixels)	Z_v (m.)	Δ_v (m.)
2195	2112	1.836	1.820	0	512	1.811	0.005	1	1.823	0.007

Once the image planes are reconfigured to approximately focus the target ($\sigma_{lf} = 0$) using the finest stereo estimate and the cameras are aimed at the target, finer adjustments of the image planes are performed to exactly focus the target and to complete the fixation process. The results of this finer focusing are included in the next table.

Table IV: Focus adjustments

left focus axis			left focus		right focus axis			right focus	
			depth	error				depth	error
p_{1m}	p_{2m}	p_L	Z_{fL} (m.)	Δ_{fl} (m.)	p_{1m}	p_{2m}	p_R	Z_{fR} (m.)	Δ_{fR} (m.)
1756	2382	2043	1.820	0.065	1550	2411	2229	1.832	0.087

(a) (b)

Figure 5: 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 citerion function is evaluated next over the rectangular box in each image.

At the completion of the fixation process, depth estimates are obtained from stereo, focus and vergence. These estimates are pairwise compared using (16) and are all accepted, thus indicating the success of the integration method. The estimates are then fused using (17) the results of which are tabulated.

Table V: Fusion of depth estimates

stereo		vergence		left focus		right focus		fused estimate
depth Z_s (m.)	error Δ_s (m.)	depth Z_v (m.)	error Δ_v (m.)	depth Z_{fL} (m.)	error Δ_{fl} (m.)	depth Z_{fR} (m.)	error Δ_{fR} (m.)	$\hat{Z}$ (m.)
1.811	0.005	1.823	0.007	1.820	0.065	1.832	0.087	1.815

The final estimated position of the fixation point is $(\hat{X}, \hat{Y}, \hat{Z}) = (0.583, 0.344, 1.815)$.

6 SUMMARY

In this paper we have described an approach to 3D location estimation of an arbitrary point using the passive depth cues of camera focus, camera vergence and stereo disparity. We have argued that any one of these cues is inadequate for the given task, and a mutual cooperation among the cues is more appropriate. We have stated that within the active vision paradigm the task of position estimation of a 3D point is best realized as a process of fixation in which the imaging parameters are dynamically controlled to bring the point into focus and to aim the cameras at it. Our implementation of the fixation process incorporates the integration of the cues of stereo, focus and vergence. Additionally, our algorithm combines the estimates from these different cues into a single estimate. Experimental results demonstrate how a dynamic imaging system is used in our implementation to fixate an object. Further speedup of our simplistic yet robust algorithm is possible with special-purpose hardware to compute the focus and vergence criterion functions and to detect features for stereo.

7 ACKNOWLEDGEMENTS

This research was supported in part by the National Science Foundation under grant IRI-89-11942, Army Research Office under grant DAAL 03-87-K-0006, and State of Illinois Department of Commerce and Community Affairs

under grant 90-103.

References

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

[2] P. Allen and R. Bajcsy. Two sensors are better than one: Example of integration of vision and touch. Technical Report MS-CIS-85-29, Computer Science Department, University of Pennsylvania, Philadelphia, PA, 1985.

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

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

[5] D. H. Ballard. Reference frames for animate vision. In *Proc. 11th IJCAI*, pages 1635–1641, Detroit, MI, August 1989.

[6] P. J. Burt. Algorithms and architectures for smart sensing. In *Proc. DARPA Image Understanding Workshop*, pages 139–153, Cambridge, MA, April 1988.

[7] J. J. Clark and N. J. Ferrier. Modal control of an attentive vision system. In *Proc. Second Intl. Conf. on Computer Vision*, pages 514–523, Tarpon Springs, FL, December 1988.

[8] S. Das and N. Ahuja. Multiresolution image acquisition and surface reconstruction. In *Proc. Third Intl. Conf. on Computer Vision*, Osaka, Japan, December 1990.

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

[10] S. Y. Harmon, G. L. Bianchini, and B. E. Pinz. Sensor data fusion through a distributed blackboard. In *Proc. IEEE Conf. on Robotics and Automation*, pages 1449–1454, San Francisco, April 1986.

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

[12] E. P. Krotkov and R. Kories. Adaptive control of cooperating sensors: Focus and stereo ranging with an agile camera system. In *Proc. IEEE Intl. Conf. on Robotics and Automation*, pages 548–553, Philadelphia, PA, April 1988.

[13] N. Nandhakumar and J. K. Aggarwal. Multisensor integration - experiments in integrating thermal and visual sensors. In *Proc. First Intl. Conf. on Computer Vision*, pages 83–92, London, UK, June 1987.

[14] R. Nevatia and K. R. Babu. Linear feature extraction and description. *Computer Graphics and Image Processing*, 13:257–269, 1980.

[15] T. J. Olson and R. D. Potter. Real-time vergence control. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 404–409, San Diego, CA, June 1989.

[16] R. D. Rimey and C. M. Brown. Selective attention as sequential behavior: Modelling eye movements with an augmented hidden markov model. Technical Report 327, University of Rochester, February 1990.

[17] S. A. Shafer, A. Stentz, and C. E. Thorpe. An architecture for sensor fusion in a mobile robot. In *Proc. IEEE Conf. on Robotics and Automation*, pages 2002–2011, San Francisco, April 1986.

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

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

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