Correspondence

Performance Analysis of Stereo, Vergence, and Focus as Depth Cues for Active Vision

Subhodev Das and Narendra Ahuja

Abstract—**This paper compares the performances of the binocular cues of stereo and vergence, and the monocular cue of focus for range estimation using an active vision system. The performance of each cue is characterized in terms of sensitivity to errors in the imaging parameters. The effects of random, quantization errors are expressed in terms of the standard deviation of the resulting depth error. The effect of systematic, calibration errors on estimation using each cue is also studied. Performance characterization of each cue is utilized to evaluate the relative performance of the cues. Also discussed, based on such characterization, are ways to select a cue taking into account the computational and reliability aspects of the corresponding estimation process.**

Index Terms—**Active vision, range from stereo, range from vergence, range from focus, performance evaluation, uncertainty analysis.**

I. Introduction

Recent research in 3D vision employing active camera systems has demonstrated the integrated use of the binocular cues of stereo and vergence and the monocular cue of focus in estimating scene surfaces or obtaining range measurements [1], [3], [6]. Judging reliability of the overall integrated system would require a careful evaluation of the performance of the individual components. The objective of this paper is to analyze and compare the uncertainties of stereo, vergence, and focus, in estimating scene surfaces. For ease of analysis, this is done by considering the estimation of range of a scene point, thus excluding the changes in the range values that would result from the use of surface smoothness constraint during surface fitting.

The uncertainty of the range estimated by the individual cues is determined from the errors in their respective imaging parameters. Individual error analyses of stereo- and focus-based range estimation methods have been reported in the past [2], [5], [6], [7], [9], [10]. We present simulated performance curves based on our analysis which indicate that every active vision system has a performance crossover point, determined by the system's parameters, on either side of which some of the cues employed by the system are more reliable than others. Since the use of a depth cue has an associated computational cost, we propose a model for combining the computational and reliability aspects of a cue that defines the operational cost of the cue in the active vision system. Finally, we discuss how to exploit these results to preferentially select or combine the various depth cues in practical situations.

This paper is organized in the following way. The next section reviews the range estimation method using each cue and analyzes the uncertainties of the estimated range values. A comparison of the performance of these different methods is presented in Section III, following which are discussed the various ways such an analysis can be utilized in practice, including the model for expressing the operational cost of a cue. Section IV presents the concluding remarks.

Manuscript received Jul. 7, 1992; revised Feb. 14, 1995.

S. Das was with the Beckman Institute, Coordinated Science Laboratory, University of Illinois, Urbana-Champaign, IL 61801; he is now with PEB Inc., Princeton, NJ 08542; e-mail: subhodev@peb.com

N. Ahuja is with the Beckman Institute, Coordinated Science Laboratory and Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, IL 61801.

To order reprints of this article, e-mail: transactions@computer.org, and reference IEEECS Log Number P95092.

II. Individual Performance of the Cues

In this section, we will be discussing stereo and focus as independent sources of depth information with vergence treated as a special case of stereo and the uncertainties of the depth estimates derived from them. Let the range estimated using stereo, vergence or focus be a non-linear function of the respective imaging parameters: $Z = f(a_1, \cdots, a_n)$. The first order relative uncertainty of the range Z is then given by

$$\frac{\Delta Z}{Z_0} = \frac{\partial Z / Z_0}{\partial_{a_1} / a_{01}} \frac{\Delta_{a_1}}{a_{01}} + \cdots + \frac{\partial Z / Z_0}{\partial_{a_n} / a_{0n}} \frac{\Delta_{a_n}}{a_{0n}} = \sum_{i=1}^{n} S_i \frac{\Delta_{a_i}}{a_{0i}}, \quad (1)$$

where $\Delta Z = Z - Z_0$ and Z_0 is the most probable range value; $\Delta a_i = a_i - a_{0i}$, a_i is the observed parameter value, and a_{0i} is the most probable parameter value. Here, $S_i = (\partial Z/Z_0)/(\partial a_i/a_{0i})$ is the linear sensitivity (dimensionless) of Z to a_i, and $\Delta a_i/a_{0i}$ is the relative uncertainty of a_i. Since the number of parameters using any cue is small and the sensitivities can be evaluated analytically, sensitivity-based uncertainty analysis of the cues is a reasonable approach in our case.

For analysis purpose, the imaging parameters of stereo and focus methods will be classified into two categories: *intrinsic*, the parameters that are internal to the camera and are not required to be changed for estimating range values at different scene locations, such as focal length and aperture; *extrinsic*, the parameters that are external to the camera or the internal parameters that may vary for different scene points, such as relative orientation of two cameras for fixating scene points or sensor plane position for focusing these points. Uncertainties in the depth estimates are caused by two types of error sources: (a) *systematic* errors which have their origin in the calibration of the intrinsic parameters and those extrinsic parameters which remain unchanged for different scene points, and (b) *random* errors which are associated with the variable extrinsic imaging parameters and cause random fluctuations in the computed range values. In general, the systematic errors introduce biases in the computed range values and thus determine the *accuracy* of the range estimation method, while the random errors limit the *precision* of the method.

In the following two subsections, we will consider each of the range estimation methods separately. For each, we will first review the estimation procedure, and then discuss its effectiveness. To characterize the effectiveness, first the significant imaging parameters (intrinsic/extrinsic or constant/variable) will be listed. Then the impact on performance of the (systematic) errors in the constant imaging parameters and that of the (random) errors in the variable imaging parameters will be discussed separately. In the former case, the sensitivities of the range value to the calibrated parameters will be studied. The reliability of any estimate in the presence of random errors will be described in terms of the standard deviation of the relative uncertainty described by (1). In reality, there may be other noise sources contributing to the random errors. These are considered to be secondary in nature and are the "unimportant" parameters of our current uncertainty model.

A. Stereo

Consider a binocular stereo camera configuration in which the optic axes of the two cameras intersect at an angle θ. The left and right camera centers have a relative displacement (i.e., baseline) of τ. The cameras are constrained such that the left and right optic axes, which make angles θ_L and θ_R, respectively, with the baseline vector, are coplanar with the latter, i.e., $\theta = 180° - \theta_L - \theta_R$. Suppose, a 3D point projects to (x_L, y_L) and (x_R, y_R) in the left and right image planes, respectively, and let (r_L, c_L) and (r_R, c_R) be the corresponding pixel locations in the frame memories. The transformation from image plane to frame memory coordinates may be expressed as $r = -\langle k_y y\rangle + r_0$ and $c = \langle k_x x\rangle + c_0$, where $\langle . \rangle$ denotes the rounding off operation and (r_0, c_0) are the pixel coordinates of image plane center in the frame memory. The parameters k_x and k_y are the ratios of number of frame pixels to sensor dimension along row and column, respectively. The depth of the 3D point with respect to the left camera center is

$$Z_L = \frac{r_R V - c_R U}{c_R A - r_R B}, \tag{2}$$

where $A = f_R/f_L(r_L - r_{0L}) + r_{0R} \sin\theta/(f_L k_x)(c_L - c_{0L}) + r_{0R}\cos\theta$, $B = (f_R k_x \cos\theta + c_{0R}\sin\theta)/(f_L k_x)(c_L - c_{0L}) - f_R k_x \sin\theta + c_{0R}\cos\theta$, $U = r_{0R}\cos\theta_R$, and $V = \tau f_R k_x \sin\theta_R + \tau c_{0R}\cos\theta_R$. Here, f_L and f_R are the focal lengths of the left and the right cameras, respectively.

In a practical dynamic stereo imaging system, the center of rotation of a camera does not usually coincide with the origin of the camera-based 3D coordinate system. As a result, the baseline changes as the cameras converge and diverge. To account for this variation, the baseline is expressed as $\tau = \tau_0 + \delta_L \cos\theta_L + \delta_R \cos\theta_R$, where δ_L and δ_R are the offsets of the centers of rotation from the origins of the 3D coordinate systems along the optic axes in the left and right cameras, respectively. These offsets are considered to be positive in the viewing direction and negative in the opposite direction. The term τ_0 represents the baseline length when the optic axes are parallel.

Vergence geometry is a special case of the stereo geometry in the sense that it can provide 3D information about one particular point in the visual field for a given camera configuration, viz., the *point of fixation* at which the optic axes of the two cameras intersect. The range of this point is

$$Z_L = \tau \frac{\sin\theta_R}{\sin\theta}. \tag{3}$$

This can also be obtained from (2) using the following substitutions: $r_L = r_{0L}$, $c_L = c_{0L}$, $r_R = r_{0R}$, and $c_R = c_{0R}$. For analysis purpose, we will also be considering another special case of stereo, that of parallel stereo, in which $\theta_L = \theta_R = 90°$ and $\theta = 0°$, and $\tau = \tau_0$. The object distance, $Z_L = Z_R = Z$, in this case evaluates to

$$Z = \frac{k_x f \tau_0}{c_R - c_L} \tag{4}$$

when $f_L = f_R = f$, $r_{0L} = r_{0R}$ and $c_{0L} = c_{0R}$ are substituted in (2).

The various parameters of the stereo method of interest for uncertainty analysis can be grouped in the following way: intrinsic – f; extrinsic – τ_0, δ_L, δ_R (constant), x_L, x_R, θ_L, θ_R (variable).

Systematic Errors. Following (4), the relative uncertainty of Z due to the uncertainty in the intrinsic parameter f is given by

$$\varepsilon_{iS} = \left(\frac{\Delta Z}{Z}\right)_{int} = S_f \frac{\Delta f}{f}, \tag{5}$$

where $S_f = 1$. Thus, the calibration error has an effect on the range uncertainty that decreases with increasing focal length f. According to (3), the uncertainty of Z_L due to the calibration errors in the constant extrinsic parameters τ_0, δ_L, and δ_R is given by

$$\varepsilon_{xS} = \left(\frac{\Delta Z_L}{Z_L}\right)_{cext} = S_{\tau 0}\frac{\Delta\tau_0}{\tau_0} + S_{\delta_L}\frac{\Delta\delta_L}{\delta_L} + S_{\delta_R}\frac{\Delta\delta_R}{\delta_R}, \tag{6}$$

where
$S_{\tau_0} = \tau_0/\tau$, $S_{\delta_L} = \delta_L \cos\theta_L/\tau$, and $S_{\delta_R} = \delta_R \cos\theta_R/\tau$.
Since $\tau \approx \tau_0$ except for very small baseline lengths, S_{τ_0} is the dominant sensitivity term. This sensitivity term is nearly constant and so also is the relative uncertainty ε_{xS} for the same relative uncertainty of τ_0. Hence, the effect of the calibration error on the range uncertainty decreases with increasing baseline τ_0.

Random Errors. There are two types of possible errors in the variable extrinsic parameters x_L and x_R (related to c_L and c_R) of (4)—quantization error and localization error. The precision of the estimated range in presence of quantization errors in feature locations, using the parallel stereo geometry, has been widely investigated [2], [5], [9]. The localization error is particularly associated with feature-based stereo methods and occurs when the locations of the detected features, e.g., zero-crossings of the Laplacian of the Gaussian operator, do not coincide with the true intensity discontinuities. Let Δx_L and Δx_R denote the image plane coordinate error variables that are uniformly distributed within the interval $[-D/2, +D/2]$ and are independent of each other. The uniformity and the independence assumptions are valid except for the cases when $|x_L - x_R|$ is very small [2]. Here, $D \geq w$, w being the width of a square pixel, is determined by the combined quantization and localization errors. The detected features can be further localized with subpixel accuracy [4]. Consequently, the feature location error is bounded by $[-d/2, +d/2]$, where $d = D/n$ and n (> 1) is the subpixel resolution. Thus, the uncertainty of range Z is given by

$$\frac{\Delta Z}{Z} = S_{x_L}\frac{\Delta x_L}{x_L} + S_{x_R}\frac{\Delta x_R}{x_R}, \tag{7}$$

where $S_{x_L} = (x_L Z)/(\mathcal{f})$ and $S_{x_R} = -(x_R Z)/(\mathcal{f})$. Observing that the errors in the different parameters are independent, we express the reliability of the range estimated using parallel stereo as

$$\sigma_{rS_\parallel}(Z) = \left[S_{x_L}^2 \sigma_r^2(x_L) + S_{x_R}^2 \sigma_r^2(x_R)\right]^{1/2}, \tag{8}$$

with $\sigma(x_L) = \sigma(x_R) = d/\sqrt{12}$ (noise-free) or $\left(d^2/12 + \sigma_n^2\right)^{1/2}$ (additive zero-mean Gaussian noise, $N(0, \sigma_n)$). The quantization of the vergence angle due to the quantized steps of the angular positioners of the left and the right cameras can also affect the precision of the range estimated using (3). Suppose, the angular error in θ_L as well as θ_R is limited to $[-\alpha/2, +\alpha/2]$, and $\Delta\theta_L$ and $\Delta\theta_R$ are the error variables. Then, the uncertainty of the range Z_L is described by

$$\frac{\Delta Z_L}{Z_L} = S_{\theta_L}\frac{\Delta\theta_L}{\theta_L} + S_{\theta_R}\frac{\Delta\theta_R}{\theta_R}, \tag{9}$$

where $S_{\theta_L} = \theta_L(\delta_L \sin\theta_L/\tau + \cot\theta)$ and $S_{\theta_R} = \theta_R(-\delta_R \sin\theta_R/\tau + \sin\theta_L/(\sin\theta_R \sin\theta))$. The reliability of the vergence-based range estimate is

$$\sigma_{rV}(Z_L) = \left[S_{\theta_L}^2 \sigma_r^2(\theta_L) + S_{\theta_R}^2 \sigma_r^2(\theta_R)\right]^{1/2}, \tag{10}$$

with $\sigma(\theta_L) = \sigma(\theta_R) = \alpha/\sqrt{12}$.

B. Focus

To estimate depth from focus, usually the distance between the lens center and the sensor plane of a camera system is varied to register a sharp image of an object point. (Alternatively, depth may also be obtained from defocusing [8], but that case will not be addressed here.) The distance, v, measured from the second principal plane and yielding the sharpest image depends on the distance of the object point, u, measured from the first principal plane. The sensor plane distance can be used to estimate the object distance using the relation

$$\frac{1}{u}+\frac{1}{v}=\frac{1}{f}, \tag{11}$$

in which f is the focal length of the lens. The object distance (range) measured from the projection center of the lens is $Z = u + t$, where t is the offset of the principal plane from the projection center and is positive in the viewing direction. The sharpness of an image is usually estimated by a criterion function that measures the high-frequency content of the image.

The different parameters of interest for the focus-based method are: intrinsic $-f$, t; extrinsic $-v$ (variable).

Systematic Errors. The uncertainty of the range, Z, due to the calibration errors in the intrinsic parameters f and t is given by

$$\varepsilon_{iF}=\left(\frac{\Delta Z}{Z}\right)_{int}=S_f\frac{\Delta f}{f}+S_t\frac{\Delta t}{t}, \tag{12}$$

where $S_f = (Z - t)^2/(Zf)$ and $S_t = t/Z$. Typically, $t \ll Z$ and $f \ll Z$, thus $S_f \approx Z/f$. Hence, the uncertainty of depth due to the systematic errors is more critically dependent on the calibration of f than that of t. Moreover, this uncertainty is more pronounced at larger distances and for smaller focal lengths.

Random Errors. The error in localizing the peak of the criterion function affects determination of the correct sensor plane position v. The two sources of error are the discrete sampling of a continuous function and the flatness of the peak. The discrete samples of the criterion function are obtained at the integer-valued steps of the sensor plane positioner to which v is linearly related. This uncertainty is assumed to be bounded by $[-B/2, +B/2]$, where B denotes a quantized step of the mechanical positioner. By interpolating around the detected peak of the focus criterion function, the localization of the true peak can be further improved to a subquantization step. As a result, the localization error can be bounded by $[-\beta/2, +\beta/2]$, where $\beta = B/n$ and n (> 1) is the desired resolution of the subquantization step. The flatness of the peak is attributed to a phenomenon commonly known as the *depth of focus*. Suppose, the sensor plane could be positioned anywhere within an interval $[v_2, v_1]$ about v_0, where $v_2 < v_0 < v_1$, such that projections of all objects within the corresponding interval $[u_2, u_1]$ about u_0, where $u_2 > u_0 > u_1$, appear equally sharp. The interval $[v_2, v_1]$ is the depth of focus and is expressed as

$$w(v_0)=v_1-v_2=\left(2AD_0f^3\right)/\left(A^2f^2-D_0^2(t+f)^2\right)\cdot Z_0/(Z_0-t-f),$$

for a thick lens. Here, D_0 is the diameter of the smallest resolvable circle which is the image of a point. The sensor noise will not seriously affect the identification of the in-focus image using the criterion function if the peak is sharp, i.e., the depth of focus is small. This is because the large number of pixels from which the criterion function is usually computed (unless the window for evaluating the function is small, such as in the vicinity of a depth discontinuity) may average out the noise effect and allow following of the peak. However, if the depth of focus is large, then the criterion function may not change much over the entire depth of focus. Assuming a zero-mean, white, additive sensor noise, it may be shown that for a criterion function based on squared image gradient values, the random noise will cause the criterion function peak to be located anywhere within the depth of focus with uniform probability. This is the basis of our uncertainty analysis for focus when depth of focus is considered.

The relative uncertainty of the range value due to the localization uncertainty of the sensor plane with respect to the true focused position is expressed as

$$\frac{\Delta Z}{Z}=S_v\frac{\Delta_v}{v}, \tag{13}$$

where the linear sensitivity $S_v = -(Z - t)(Z - t - f)/(Zf)$. The reliability of the focus-based range estimate is then expressed as

$$\sigma_{rF}(Z)=S_v\sigma_r(v), \tag{14}$$

where

$\sigma(v) = \beta/\sqrt{12}$ (if $w(v) < \beta$)

or

$$D_0\left[A^2f^2+3D_0^2(t+f)^2\right]^{1/2}/\sqrt{3}\left[A^2f^2-D_0^2(t+f)^2\right]\cdot f^2Z/(Z-t-f)$$

(otherwise). In a typical imaging system, the aperture A may be in the order of several centimeters while the diameter of the confusion circle D_0 is in the order of several microns. According to (14), $\sigma_{rF} \approx (D_0Z)/\left(\sqrt{3}Af\right)$ when $w(v) > \beta$, i.e., for large distances. Hence, the reliability of the focus-based depth estimate which degrades with distance can be improved by 1) increasing the focal length f, 2) increasing the aperture A, and 3) reducing D_0 by increasing the resolution of the sensor plane.

III. Relative Performance of the Cues

In this section, we will compare the performance of stereo disparity, vergence, and focus in the presence of systematic and random errors in the imaging parameters. We first discuss the relative performance with respect to the systematic errors for which we will compare the relative uncertainties of the computed range due to these errors. Next, we describe the relative performance of the cues in the presence of random errors by comparing the statistics of the errors in range estimates.

We observe that the relative uncertainty of the range estimated from vergence, ε_{xV}, due to the calibration error in the baseline is expressed in (6). With the relative uncertainty of range from focus given by (12), the ratio of these two uncertainties for the same relative uncertainties of the input parameters is

$$\frac{\varepsilon_{xV}}{\varepsilon_{iF}}=\frac{S_{\tau_0}+S_{\delta_L}+S_{\delta_R}}{S_f+S_t}=\frac{f}{Z} \tag{15}$$

for $Z \gg t$. Since, for a typical imaging system $Z \gg f$, we find that $\varepsilon_{xV} \ll \varepsilon_{iF}$. In other words, calibration of focal length is relatively more important for focus than baseline is for vergence.

To compare stereo disparity and focus, we will consider the simple case of parallel stereo. Let the relative uncertainty of range from parallel stereo due to the calibration errors in f and τ_0 be denoted by ε_S. According to (5) and (6), the linear sensitivities of range to these parameters are $S_f = 1$ and $S_{\tau_0} = \tau_0/\tau = 1$, respectively. Thus, the ratio of the relative uncertainties of stereo and focus-based estimates due to the systematic errors is

$$\frac{\varepsilon_s}{\varepsilon_{iF}} = \frac{S_{fS}+S_r}{S_{fF}+S_t} = \frac{2f}{Z} \tag{16}$$

when the relative uncertainties of the corresponding input parameters of both cues are assumed to be equal, and $Z \gg t$. Once again, for $Z \gg f$, the systematic errors have more pronounced effect on focus than on stereo.

To compare the relative performance of vergence and focus in presence of random errors, we utilize (10) and (14). The ratio of the reliability measures of vergence and focus or the *sensitivity* of focus with respect to vergence, $S_{F/V}$, is obtained as

$$S_{F/V} = \frac{\sigma_{rV}}{\sigma_{rF}} = \begin{cases} \left(\frac{\alpha f}{\beta}\right)\left(\frac{f}{Z}\right)\sqrt{\cot^2\theta+\csc^2\theta}, & \text{if } w(v) < \beta \\ \left(\frac{\alpha A}{2D_0}\right)\left(\frac{f}{Z}\right)\sqrt{\cot^2\theta+\csc^2\theta}, & \text{otherwise} \end{cases} \tag{17}$$

when we assume that $\theta_L = \theta_R$, δ_L, $\delta_R \ll \tau_0$, $A \gg D_0$, and $Z \gg t + f$. Now, a value of $S_{F/V}$ considerably larger than unity would indicate that focus is more sensitive (better) than vergence. A value of $S_{F/V}$ considerably smaller than unity indicates that vergence is more sensitive. Thus, the condition for the higher sensitivity of vergence-based measurements is

$$\frac{\sqrt{\cot^2\theta+\csc^2\theta}}{Z} < \frac{\beta}{\alpha f^2}, \text{ or } \frac{\sqrt{\cot^2\theta+\csc^2\theta}}{Z} < \frac{2D_0}{\alpha A f}. \tag{18}$$

For very small range values, $\theta_L = \theta_R \to 0^\circ$ or $\theta \to 180^\circ$, thus both inequalities are unlikely to be satisfied. In other words, focus-based estimates are likely to be more reliable at very close range. However, with increasing range, particularly when $Z \gg \tau$, the term $\cot^2\theta + \csc^2\theta$ can be approximated in the following way:

$$\cot^2\theta+\csc^2\theta = \frac{8Z^4-4Z^2\tau^2+\tau^4}{\tau^2\left(4Z^2-\tau^2\right)} \approx \frac{2Z^2}{\tau^2},$$

where the expression following the equality sign is obtained by letting $\theta_L = \theta_R = \theta_0$ and noting that $Z = \tau/(2\cos\theta_0)$ from (3). Consequently, the inequalities of (18) can be approximated as

$$\frac{\alpha}{\tau} < \frac{\beta}{\sqrt{2}f^2}, \quad \text{or} \quad \frac{\alpha}{\tau} < \sqrt{2}\,\frac{D_0}{Af}. \tag{19}$$

The parameters on the left-hand side (LHS) of the inequality sign are associated with vergence, while those on the right-hand side (RHS) are associated with focus. With proper selection of the vergence parameter values, it is possible to satisfy either of the inequalities for practical active vision systems. (At least, there is no restriction, technological or otherwise, on how large the baseline τ can be.) In other words, the sensitivity of focus with respect to vergence, $S_{F/V}$, is likely to exhibit a *crossover point* with respect to $S_{F/V} = 1$ as the range increases. This is illustrated in Fig. 1. The location of the crossover point is completely determined by the imaging parameters and therefore varies from system to system.

To compare the reliabilities of stereo disparity and focus methods, we utilize the results of (8) and (14). The sensitivity of focus with respect to parallel stereo is given by

$$S_{F/S_\parallel} = \frac{\sigma_{rS}^{\parallel}}{\sigma_{rF}} = \begin{cases} \left(\sqrt{2}fd\right)/(\tau\beta), & \text{if } w(v) < \beta \\ (Ad)/\left(\sqrt{2}\tau D_0\right), & \text{otherwise} \end{cases} \tag{20}$$

where we assume that $A \gg D_0$ and $Z \gg t + f$. The corresponding condition for the higher sensitivity of stereo is

$$\frac{d}{\tau} < \frac{\beta}{\sqrt{2}f}, \text{ or } \frac{d}{\tau} < \frac{\sqrt{2}\,D_0}{A}. \tag{21}$$

The LHS and RHS parameters of the inequality sign are associated with stereo and focus, respectively. Comparing (18) and (21), we find that the inequalities of (21) are independent of the range and are completely determined by the stereo and focus parameter values. In other words, the sensitivity measure $S_{F/S_\parallel}$ does not exhibit any crossover point as the range changes, provided that the range is not too small to invalidate the assumption $Z \gg t + f$. Consequently, the sensitivity is either greater or less than unity, which is illustrated in Fig. 1. Additionally, we note that D_0 is in the order of the size of a CCD element while d can be considerably smaller than D_0 under subpixel localization. Thus, if τ is selected to be significantly larger than A (say, $\tau \gg 2A$), then parallel stereo can be made more sensitive than focus. Finally, it is intuitive from the behavior of $S_{F/V}$ and $S_{F/S_\parallel}$ that the sensitivity of focus with respect to verging stereo, $S_{F/S}$, should also exhibit a crossover point with increasing range value. This observation is confirmed by the simulation results of Fig. 1. Once again, the actual location of the crossover point is system dependent.

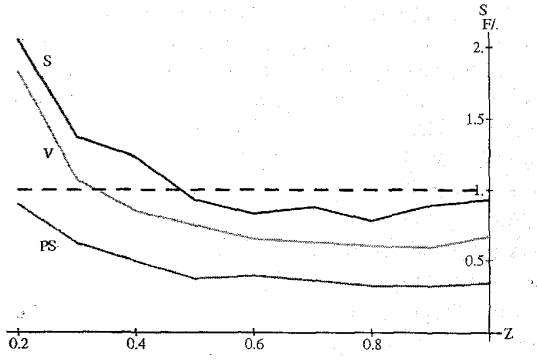
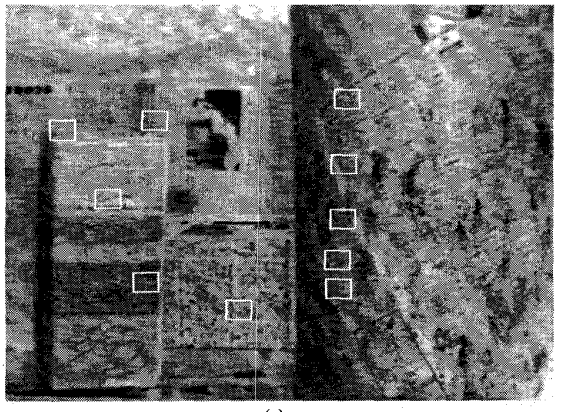


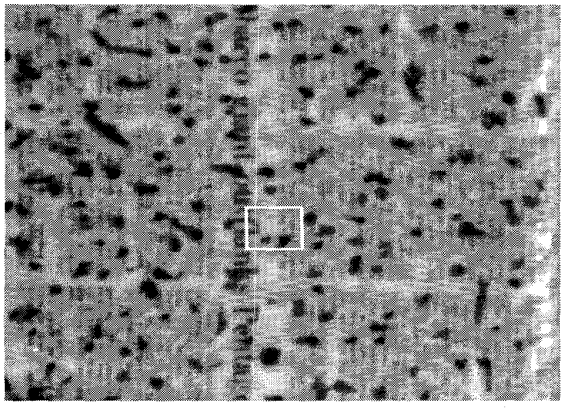
Fig. 1. Simulation results showing the relative performance of focus and stereo as a function of the range: sensitivity of focus with respect to vergence, $S_{F/V}$, has a crossover point; sensitivity of focus with respect to parallel stereo, $S_{F/PS}$, is less than unity; sensitivity of focus with respect to verging stereo, $S_{F/S}$, too has a crossover point. The same imaging parameter values, typical of an active vision system, are used for the three different situations. The range, Z, is in meters.

To demonstrate the utility of the relative performance analysis of the cues in practical situations, we consider the task of controlling an active vision system for surface reconstruction of scenes that are wide and deep, such as the one shown in Fig. 2a. At any given time during imaging of such a scene, sharp features are acquired for a limited depth range and only a small part of the scene is visible. Thus, it is required that the cameras be aimed in different directions and fixate upon objects at different distances. In our first example, a verging stereo camera system is directed at various objects in the scene sequentially and the decision to utilize focus or vergence-based estimate in the fixation of each object is sought. The baseline length of the camera system is $\tau_0 = 28$ cm, the circle of confusion has a calibrated diameter of $D_0 = 24$ μm or about 2 pixels, and the angular resolution of the vergence stepper motors is $\alpha = 1.7 \times 10^{-4}$ radians/step. Fixation is attempted using a focal length of $f = 105$ mm and an aperture diameter of $A = 50$ mm. Substituting these parameter values in (19), we

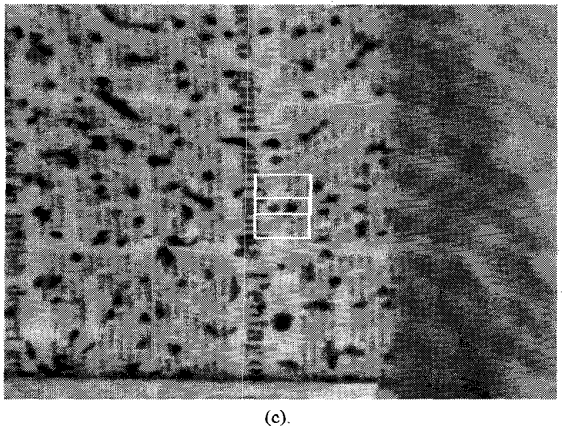
observe that $\alpha / \tau < \sqrt{2} D_0 / Af$ for object distances greater than the baseline length. Consequently, for these object distances, the estimate of vergence is more reliable than that of focus and hence it is used in the fixation process shown in Fig. 2b, and 2c.



(a).



(b).

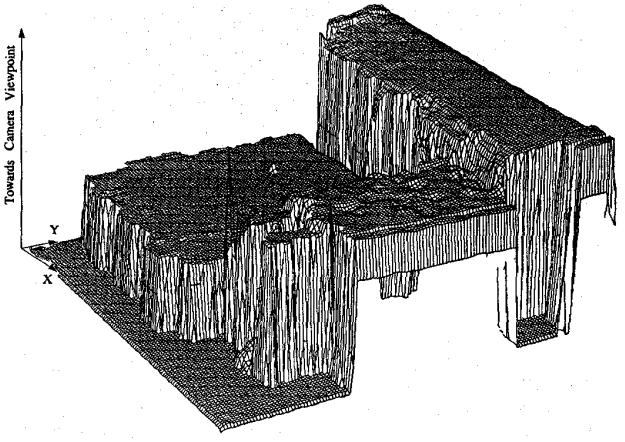


(c).

Fig. 2 . Experimental results utilizing relative performance analysis of focus and vergence: (a) An overview of a scene requiring fixation and stereo analysis of individual objects (rectangles enclose fixation points); fixation of the horizontal box using vergence-based depth estimate, (b) left, and (c) right images showing image centers and the best match to the left center (upper rectangle in right).

The second example employs a parallel stereo camera system. Here, the problem is selecting focus or coarse stereo-based range information as the initial estimate for stereo-based surface reconstruction for the in-focus part during any given fixation of the scene shown in Fig. 3a. The coarse stereo information is derived during an earlier fixation when the current in-focus part happened to be part of the peripheral visual field and, therefore, was subjected to lens defocusing [3]. Inorder to have overlapping visual fields, the baseline length of the parallel stereo camera system is considerably smaller than that of the verging camera system and is $\tau_0 = 5$ cm With an aperture diameter of $A = 34$ mm, the feature localization uncertainty of stereo d has to be less than $2D_0$ or 4 pixels according to (21) for greater reliability of coarse stereo over focus. However, lens defocusing (assumed to be Gaussian in nature) of peripheral features together with a Gaussian-based feature detector would cause the localization uncertainty to be greater than the above limit (typically, the uncertainty is in the range of 6–12 pixels). As a result, focus-based depth is the preferred initial estimate for stereo analysis and the corresponding result is shown in Fig. 3b.

(a).



(b).

Fig. 3. Experimental results utilizing relative performance analysis of focus and parallel stereo: (a) An overview of a scene requiring piecewise stereo analysis of different parts; (b) The composite range map corresponding to the left viewpoint obtained by merging the individual stereo analysis results in which depth from focus has been the initial estimate.

The knowledge of relative performance of the different depth cues is also helpful to determine the operational criteria of the cues as each range estimation method has an associated computational cost. Active vision offers the opportunity to obtain multiple measurements of a scene point such as from different viewpoints. Now, it is well known that if σ denotes the standard deviation of a population given by the probability function of a random variable, then the standard deviation of the mean of n independent samples drawn from the population is $\sigma / \sqrt{n}$. Suppose, n measurements by a range estimation method B is equivalent, in terms of precision, to a single measurement by another method A, i.e., $\sigma_B / \sqrt{n} = \sigma_A$. Applying the definition of sensitivity (of A with respect to B), it is obtained that $n = \sigma_B^2 / \sigma_A^2 = S_{A/B}^2$. Also, let a single measurement by A cost γ_A units, and that by B cost γ_B units. Then, in order to achieve the same precision, the ratio of the total computational costs of B to A or the *preferability* of A with respect to B is

$$P_{A/B} = \frac{\gamma_B n}{\gamma_A} = \frac{\gamma_B}{\gamma_A} S_{A/B}^2 = C_{B/A} S_{A/B}^2. \tag{22}$$

The term, $C_{B/A} = \gamma_B / \gamma_A$, is the *expendability* of B with respect to A. A value of $P_{A/B}$ greater than unity signifies that A is preferable to B.

A common measure of computational cost is the time required to obtain a measurement. Thus, (22) can be used to select a range estimation method that gives the best tradeoff between precision and time. To illustrate this point, we plot the preferability of focus to vergence in Fig. 4a and that of focus to parallel as well as general stereo in Fig. 4b using the sensitivity simulation results of Fig. 1. To the left of the point of intersection of any curve with $P_{F/.} = 1$, focus is the preferred depth cue. It is observed that although the precision of the general stereo method is marginally ($\sim$ 7%) better than that of focus (see Fig. 1), the cost of the former method is significantly ($\sim$ 70%) more than that of the latter when $Z = 1$ m. In this manner, the *computational merit* of an estimation method, measured by its cost, can be combined with its *technical merit*, measured by its sensitivity, to define an operational criterion for an estimation method.

IV. Conclusions

In this paper, we have presented performance analyses of three visual cues used in a typical active vision system for surface reconstruction: stereo, vergence, and focus, as sources of depth. The performance of each cue has been characterized by the derived expressions for the standard deviation of the relative uncertainty in the presence of random errors in imaging parameters. These statistics have been subsequently used to evaluate the relative performance of the cues. We have demonstrated the utility of the relative performance analysis by considering practical examples of 3D reconstruction. As an extension of the applicability of our analysis, we have proposed a model for combining the computational and reliability aspects of a cue that is useful for cue selection. Performance characterization of the visual cues in the manner described in this paper is likely to be useful in controlling an active vision system under time-constrained, reliability-demanding situations.

An integrated system provides the opportunity to reduce the uncertainties in range due to random errors. The use of sharp and well-localized features improves the reliability of stereo-based range estimates; focusing can ensure sharpness of image features for objects within the depth of field of the lens. Without any initial guess, the search for the optimal sensor plane position to register the sharpest image of an object point is slow because of the large range of sensor plane positions that needs to be examined. Also, the possibility of improper localization of the peak of the focus criterion is higher if this range is large. An estimate of the range from stereo or vergence can help in predicting the range of image plane positions by approximating the depth of focus, thereby improving the preferability of focus with respect to other cues. Finally, fusion of depth measurements obtained from different cues can lead to an estimate whose error is lower than those of the individual estimates. The fusion process is appropriate when the sensitivity of focus with respect to any other cue is not too different from unity. These are some of the many different ways the *tools* for performance evaluation of the depth cues developed in this paper will be beneficial for designing and operating systems that use these cues.

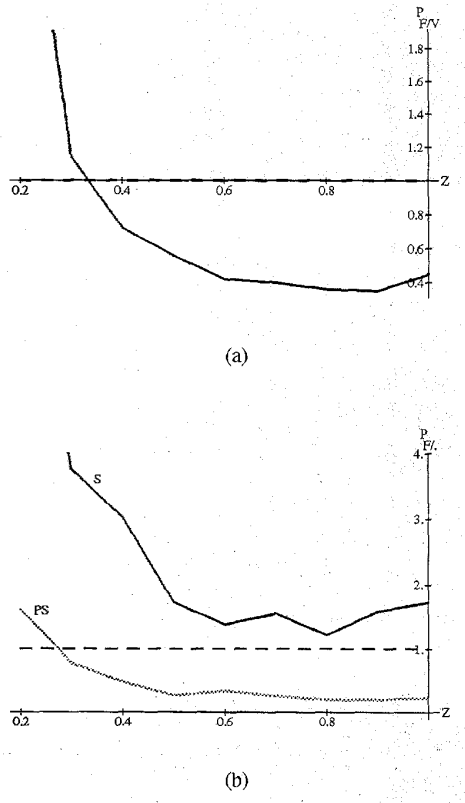


Fig. 4. Plots of cost ratio as a function of the range: (a) vergence to focus assuming $C_{V/F} = 1$, (b) parallel and general stereo (X) to focus assuming $C_{X/F} = 2$. The latter expendability is typical of most active vision systems. The range, Z, is in meters.

References

[1] N. Ahuja and A.L. Abbott, "Active stereo: Integrating disparity, vergence, focus, aperture, and calibration for surface estimation," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 15, no. 10, pp. 1,007–1,029, 1993.

[2] S.D. Blostein and T.S. Huang, "Error analysis in stereo determination of 3-d point positions," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 9, pp. 752–765, Nov. 1987.

[3] S. Das and N. Ahuja, "Multiresolution image acquisition and surface reconstruction," *Proc. Third Intl Conf. Computer Vision*, Osaka, Japan, pp. 485–488, Dec. 1990.

[4] A. Huertas and G. Medioni, "Detection of intensity changes with subpixel accuracy using laplacian-gaussian masks," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 8, no. 5, pp. 651–664, 1986.

[5] B. Kamgar-Parsi and B. Kamgar-Parsi, "Evaluation of quantization error in computer vision," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 9, pp. 929–940, Sept. 1989.
[6] E.P. Krotkov, *Active Computer Vision by Cooperative Focus and Stereo*. New York: Springer-Verlag, 1989.
[7] L. Matthies and S.A. Shafer, "Error modelling in stereo navigation," *IEEE J. Robotics and Automation*, vol. 3, pp. 239–248, June 1987.
[8] A.P. Pentland, "A new sense for depth of field," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 9, pp. 523–531, 1987.
[9] J.J. Rodríguez and J.K. Aggarwal, "Stochastic analysis of stereo quantization error," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 467–470, May 1990.
[10] M.A. Snyder, "Uncertainty analysis of image measurements," *DARPA Image Understanding Workshop*, pp. 681–693, Los Angeles, Calif., Feb. 1987.

On the Estimation of Rigid Body Rotation from Noisy Data

Daniel Goryn and Søren Hein, *Member, IEEE*

***Abstract*—We derive an exact solution to the problem of estimating the rotation of a rigid body from noisy 3D image data. Our approach is based on total least squares (TLS), but unlike previous work involving TLS, we include the constraint that the transformation matrix should be orthonormal. It turns out that the solution to the estimation problem has the same form as if the data are not noisy, and thus the solution to the standard Procrustes problem can be applied.**

***Index Terms*—Computer vision, rotation estimation, total least squares.**

I. Introduction

In computer scene analysis and computer vision applications, the problem of estimating motion parameters of objects often occurs. In the case of noise-free observations of object points, the problem is easily solved with zero error. However, in practice the observations are noisy. Arun, Huang, and Blostein [1] have proposed an algorithm which estimates the translation vector and unitary matrix that best map one point set into another. The algorithm is based on singular value decomposition (SVD), and assumes that noise is only present on one of the two point sets. The algorithm is not guaranteed to return a rotation matrix, and may instead return a reflection matrix. Subsequently, Umeyama [2] has improved upon the algorithm so that it always returns a rotation matrix. Faugeras and Hebert [3] have proposed an algorithm which directly returns a rotation matrix; these authors also assume that only one of the point sets is noisy.

In this note we modify the assumptions in [1], [2], and [3] by assuming that both point sets are noisy, and we show that the result of [1] remains valid under this more general assumption.

II. Constrained Total Least Squares Solution

We have available two sets of noisy observations of n points of a rigid body, that is, we have two $n \times 3$ matrices $\mathbf{A}$ and $\mathbf{B}$. We assume that the correspondence problem has been solved, so that corresponding points are in the same order in $\mathbf{A}$ and $\mathbf{B}$. The translation of the object can easily be taken into account as in [4], so we also assume that any translation has already been compensated for. The problem is to find a 3×3 matrix $\mathbf{X}$ and two perturbations $\Delta\mathbf{A}$ and $\Delta\mathbf{B}$ which satisfy

$$(\mathbf{A} + \Delta\mathbf{A})\mathbf{X} = \mathbf{B} + \Delta\mathbf{B}, \tag{1}$$

such that the perturbation size $E = \|\Delta\mathbf{A}\|^2 + \|\Delta\mathbf{B}\|^2$ is minimized and the matrix $\mathbf{X}$ is unitary, that is, $\mathbf{X}^T\mathbf{X} = \mathbf{I}$. We will use the Frobenius matrix norm.

To solve the problem, we first introduce the following variables for notational convenience:

$$\begin{aligned} \mathbf{D} &= (\mathbf{B}\mathbf{X}^T - \mathbf{A})/2 \\ \Delta\mathbf{A}' &= \Delta\mathbf{A} - \mathbf{D} \\ \Delta\mathbf{B}' &= \Delta\mathbf{B} + \mathbf{D}\mathbf{X}. \end{aligned} \tag{2}$$

Using the unitarity of $\mathbf{X}$, (1) can then be written

$$\Delta\mathbf{A}'\mathbf{X} = \Delta\mathbf{B}'. \tag{3}$$

Using (2), (3), and the unitarity of $\mathbf{X}$, the perturbation size becomes

$$\begin{aligned} E &= \|\Delta\mathbf{A}' + \mathbf{D}\|^2 + \|\Delta\mathbf{A}' - \mathbf{D}\|^2 \\ &= 2\left(\|\Delta\mathbf{A}'\|^2 + \|\mathbf{D}\|^2\right) \\ &\quad + 2\mathrm{tr}\left(\Delta\mathbf{A}'^T\mathbf{D}\right) - 2\mathrm{tr}\left(\Delta\mathbf{A}'^T\mathbf{D}\right) \\ &= 2\left(\|\Delta\mathbf{A}'\|^2 + \|\mathbf{D}\|^2\right) \\ &= 2\left(\left\|\Delta\mathbf{A} - \left(\mathbf{B}\mathbf{X}^T - \mathbf{A}\right)/2\right\|^2 + \left\|\mathbf{B}\mathbf{X}^T - \mathbf{A}\right\|^2/4\right) \end{aligned} \tag{4}$$

An obvious lower bound E_0 on E is obtained if the first term in the last line of (4) is zero, and $\mathbf{X}$ is chosen to be the value $\mathbf{X}_0$ which minimizes the second term. Furthermore, the value E_0 is attainable by setting $\mathbf{X} = \mathbf{X}_0$ and $\Delta\mathbf{A} = (\mathbf{B}\mathbf{X}_0^T - \mathbf{A})/2$, which in turn implies $\Delta\mathbf{B} = (\mathbf{A}\mathbf{X}_0 - \mathbf{B})/2$ due to (3). We have thus reduced the original problem to the problem of choosing a unitary matrix $\mathbf{X}$ to minimize $\|\mathbf{B}\mathbf{X}^T - \mathbf{A}\|^2$, which is recognized as the standard Procrustes problem [5]. The solution is [5]

$$\mathbf{X}_0 = \mathbf{U}\mathbf{V}^T,$$

where the orthogonal 3×3 matrices $\mathbf{U}$ and $\mathbf{V}$ are given in terms of the SVD of $\mathbf{A}^T\mathbf{B}$:

$$\mathbf{A}^T\mathbf{B} = \mathbf{U}\mathbf{S}\mathbf{V}^T.$$

This SVD method for determining $\mathbf{X}_0$ has the same form as the one presented in [1]. To always obtain a rotation matrix (and not a reflection matrix), the modification proposed in [2] should be used.

Acknowledgments

We would like to thank the reviewers and Profs. Jim Demmel and G. W. Stewart for comments that led to a clearer exposition.

References

[1] K.S. Arun, T.S. Huang, and S.D. Blostein, "Least-squares fitting of two 3-D point sets," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 9, no. 9, pp. 698–700, Sept. 1987.
[2] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 376–380, Apr. 1991.
[3] O.D. Faugeras and M. Hebert, "A 3D recognition and positioning algorithm using geometric primitive durfaces," *Proc. Int'l Joint Conf. Artificial Intelligence*, pp. 996–1,002, Aug. 1983.
[4] T.S. Huang, S.D. Blostein, and E.A. Margerum, "Least squares estimation of motion parameters from 3D point correspondences," *Proc. IEEE*

Manuscript received Sept. 2, 1993; revised May 16, 1995. Recommended for acceptance by S. Peleg.

The authors are with Siemens AG, 81730 München, Germany; e-mail: shein@hl.siemens.de.

To order reprints of this article, e-mail: transactions@computer.org, and reference IEEECS Log Number P95128.