A SIMPLE, REAL-TIME RANGE CAMERA

A. Pentland, T. Darrell, M. Turk, and W. Huang

Vision Sciences Group, The Media Lab, M.I.T
20 Ames St., Cambridge MA 02138

ABSTRACT

We describe a simple imaging range sensor based on the measurement of *focal error*, as described in [Pentland 1982,1987]. The current implementation can produce range over a 1 cubic meter workspace with a measured standard error of 2.5% (4.5 significant bits of data). The system is implemented using relatively inexpensive commercial image processing equipment.

1 Introduction

It is often said that range information is lost during the process of image formation, so that vision is underconstrained. However this is not precisely true: it is only when using a pinhole camera that we lose all depth information. With real lenses (and in the human eye) the situation is as shown in Figure 1(a). We see that inside the camera, between the lens and the image plane, the 3-D shape of the world is copied along the surface where the image is exactly focused. Along each geometric ray between the image plane and the lens the image moves from being in relatively poor focus, to a point of best focus, and then back to being out of focus, as is illustrated by Figure 1(b). Thus if we could trace along the path of each incoming ray to find the point of exact focus then we could recover the shape of the 3-D world.

Autofocus methods actually function in this manner, by searching along the central ray to find the point of best focus for that particular point [1]. More recently several authors [2,3] have collected a series of images with different focal lengths in order to estimate the point of best focus for each image position. By collecting between eight and thirty images they have been able to reconstruct the scenes 3-D geometry with relatively good accuracy. However the need for many images, and to move the lens between each image, means that a second or more is required to compute range. During this time there must be no scene motion.

Further, changing the focal length (or any camera parameter other than aperture) introduces geometric distortion between the images, so that each image must be geometrically warped back to some standard geometry. Thus this approach also requires hardware capable of relatively sophisticated geometric warping.

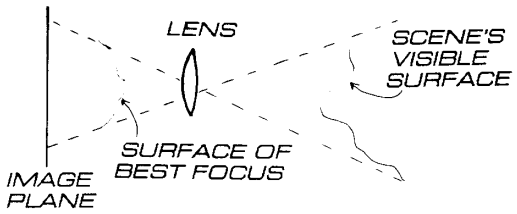


Figure 1: For most real lens systems range information is not lost during image formation.

In contrast, the method described here requires only one view: rather than search for the best focus, we simply measure the error in focus by comparing two geometrically identical images, one with a wide aperture, so that objects off the focal plane are blurred, with a small-aperture image where everything is sharply focused. The images are collected at the same time, so that scene motion is not a problem, and are collected along the same optical axis with the same focal length, so that there is no geometrical distortion.

Once the focal error has been measured we can extract depth immediately. The difference between autofocus techniques and this method, therefore, is analogous to the difference between convergence and stereopsis: both autofocus and depth-from-convergence change the camera parameters to measure depth at a single point, whereas both this method and stereopsis utilize the error signal (blur and disparity, respectively) to estimate depth.

Subbarao [8] has recently pointed out that there are many similar range techniques based on the idea of comparing two views that vary only by one camera parameter. These other methods, however, have the drawback that variation of any camera parameter other than aperture size will also introduce geometric distortion, and thus geometric warping will be required in order to estimate depth.

2 Measuring Range

Most real lens systems are exactly focused [1] at only one distance along each radius from the lens into the scene. The locus of exactly focused points forms a doubly curved, approximately spherical surface in three-dimensional space. Only when objects in the scene intersect this surface is their image exactly in focus; objects distant from this surface of exact focus are blurred, an effect familiar to photographers as depth of field.

The distance D to an imaged point is related to the parameters of the lens system and the amount of defocus by the following equation [7]:

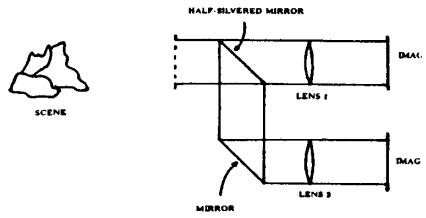
$$D = \frac{F v_0}{v_0 - F - \sigma f} \tag{1}$$

where v_0 is the distance between the lens and the image plane (e.g., the film location in a camera), f the f-number of the lens system, F the focal length of the lens system, and σ the spatial constant of the point spread function (i.e., the radius of the imaged point's "blur circle") which describes how an image point is blurred by the imaging optics. The point spread function may be usefully approximated by a two-dimensional Gaussian $G(r, \sigma)$ with a spatial constant σ and radial distance r. The validity of using a Gaussian to describe the point spread function is discussed in reference [7].

In most situations, the only unknown on the right-hand side of Equation (1) is σ, the point spread function's spatial parameter. Thus, we can use Equation 1 to solve for absolute distance given only that we can measure σ, the amount of blur at a particular image point. Measurement of σ presents a problem, however, for the image data is the result of both the characteristics of the scene and those of the lens system. To disentangle these factors, we can either look for places in the image with known characteristics (see [4-7,9,10]), or we can observe what happens when we change some aspect of the lens system. It is this second approach we have taken here.

2.1 Comparison Across Differing Apertures

Given two images of exactly the same scene, but with different depth of field, we can factor out the contribution of the scene to the two images (as the contribution is the same), and measure the focus directly. Figure 2(a) shows an optical system design for taking a single view of the scene and producing two images that are identical except for aperture size and therefore depth of field. This lens system uses a beam splitter to separate the original image into two identical images, which are then directed through lens systems with different aperture size. Because change in aperture does not affect the position of image features, the result is two images that are *identical*[2] except for focal error so there is no difficulty in matching points in one image to points in the other. Figure 2 (b) shows the device we have built to implement this optical design.



(a)

(b)

Figure 2: (a) A lens geometry for acquiring two images that are identical except for depth of field, (b) a picture of the device.

Because differing aperture size causes differing focal errors, the same point will be focused differently in the two images. The critical fact is that the magnitude of this difference is a simple function of only one variable: the distance between the viewer and the imaged point. To obtain an estimate of depth, therefore, we need only compare corresponding points in the two images and measure this change in focus.

2.1.1 Mathematical Background

We start by taking a patch $f_1(r, \theta)$ centered at (x_0, y_0) within the first image $I_1(x, y)$:

$$f_1(r, \theta) = I_1(x_0 + r\cos\theta, y_0 + r\sin\theta) \tag{2}$$

and calculate its two-dimensional Fourier transform $\mathcal{F}_1(t, \theta)$. The same is done for a patch $f_2(r, \theta)$ at the corresponding point in the second image, giving us $\mathcal{F}_2(t, \theta)$. Again,

[1] "Exact focus" is taken here to mean "has the minimum variance point spread function," the phrase "measurement of focus" is taken to mean "characterize the point spread function."

[2] Their overall brightness also differs, requiring the use of neutral density filters.

note that there is no matching problem, as the images are identical except for depth of field.

Now consider the relation of f_1 to f_2. Both cover the same region in the image, so that if there were no blurring both would be equal to the same intensity function $f_0(r,\theta)$. However, because there is blurring (with spatial constants σ_1 and σ_2), we have[3]

$$\frac{f_1(r,\theta)}{f_2(r,\theta)} = \frac{f_0(r,\theta) \otimes G(r,\sigma_1)}{f_0(r,\theta) \otimes G(r,\sigma_2)} \tag{3}$$

where $G(r,\sigma)$ is a two-dimensional Gaussian [4] with variance σ^2.

Noting that $f(r,\theta) = e^{-\pi r^2}$ and $\mathcal{F}(\lambda,\theta) = e^{-\pi\lambda^2}$ are a Fourier pair and that if $f(r,\theta)$ and $\mathcal{F}(\lambda,\theta)$ are a Fourier pair then so are $f(\alpha r,\theta)$ and $1/|\alpha|\mathcal{F}(\lambda/\alpha,\theta)$ we see that we may use Equation 3 to derive the following relationship between $\mathcal{F}_1$, $\mathcal{F}_2$ and $\mathcal{F}_0$

$$\mathcal{F}_1(\lambda,\theta) = \frac{\mathcal{F}_0(\lambda,\theta)G(\lambda,\frac{1}{\sqrt{2\pi}\sigma_1})}{\sqrt{2\pi}\sigma_1} \tag{4}$$

$$\mathcal{F}_2(\lambda,\theta) = \frac{\mathcal{F}_0(\lambda,\theta)G(\lambda,\frac{1}{\sqrt{2\pi}\sigma_2})}{\sqrt{2\pi}\sigma_2} \tag{5}$$

Thus

$$\frac{\mathcal{F}_1(\lambda)}{\mathcal{F}_2(\lambda)} = \frac{G(\lambda,\sigma_1)\sigma_2}{G(\lambda,\sigma_2)\sigma_1} = \frac{\sigma_2^2}{\sigma_1^2}\exp(\lambda^2 2\pi^2(\sigma_2^2 - \sigma_1^2)) \tag{6}$$

where $\mathcal{F}(\lambda) = \int_{-\pi}^{\pi}\mathcal{F}(\lambda,\theta)d\theta$. Thus, given $\mathcal{F}_1$ and $\mathcal{F}_2$ we can find σ_1 and σ_2, as follows. Taking the natural log of Equation 6 we obtain

$$\ln\frac{\sigma_2^2}{\sigma_1^2} + \lambda^2 2\pi^2(\sigma_2^2 - \sigma_1^2) = \ln\mathcal{F}_1(\lambda) - \ln\mathcal{F}_2(\lambda) \tag{7}$$

If make the first camera be a pinhole camera (so that $\sigma_1 = \epsilon$ for some small value ϵ), then we can derive the following relation:

$$k_1\sigma_2^2 + k_2\ln\sigma_2 + k_3 = \ln\mathcal{F}_1(\lambda) - \ln\mathcal{F}_2(\lambda) \tag{8}$$

where the k_i are constants. Thus the difference in localized Fourier power is a monotonic increasing function of the blur in the second image. Or, more importantly, by Equation 1, the distance to the imaged point is a monotonic decreasing function of the difference in the localized Fourier power.

[3] Equation 3 may be substantially in error in cases with a large amount of defocus, as points neighboring the patches f_1, f_2 will be "spread out" into the patches by differing amounts. This problem can be avoided by using patches whose edges trail off smoothly, e.g., $f_1(r,\theta) = I(x_0 + r\cos\theta, y_0 + r\sin\theta)G(r,\omega)$ for appropriate spatial parameter ω.

[4] The use of a Gaussian to model blur is discussed in reference [7]

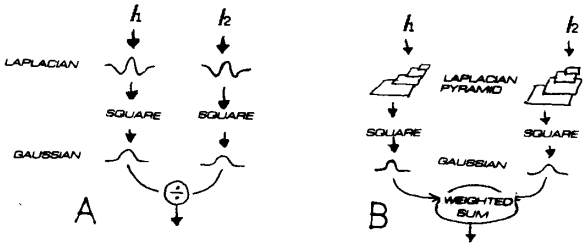


Figure 3: (a) Computation of range at one scale (b) Multiscale computation.

3 Practical Implementation

The calculation of Fourier transforms at each image point is too expensive to make for a practical technique. Because we need only the Fourier power, however, we may make use of Parseval's Theorem, which states that the integral of squared values over the spatial domain is equal to the integral of the squared Fourier components over the frequency domain (the Fourier power):

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|f(x,y)|^2\,dxdy = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|\mathcal{F}(r,\theta)|^2\,drd\theta \tag{9}$$

$$= \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathcal{P}(r,\theta)drd\theta \tag{10}$$

where $\mathcal{F}(r,\theta)$ is the Fourier transform of $f(x,y)$ and $\mathcal{P}(f_x,f_y)$ is the power spectrum. Convolution with a band-pass filter — such as a Laplacian — results in a signal which is restricted to a limited range of frequencies. Therefore, the integral of the square of the convolved signal is proportional to the integral of the power within the original signal over this range of frequencies.

This leads to the simple, single-scale implementation shown in Figure 3(a). Image data from each image is convolved with a 8 x 8 Laplacian filter, and the values squared. These values are then averaged using an 8 x 8 Gaussian filter, resulting in a "power image" for each camera; an estimate of the Fourier power at the center spatial frequency of the Laplacian filter for each image location. These two power images are then compared using a lookup table to produce an estimate of range. With our current setup we can obtain range data with a measured standard error of 6% at up to eight frames per second.

This single-scale processing scheme has been implemented using Datacube image processing equipment. Digitization requires two Digmax boards, convolution is accomplished by one VFir II board, temporary image storage is provided by a RoiStore board, and table lookup func-

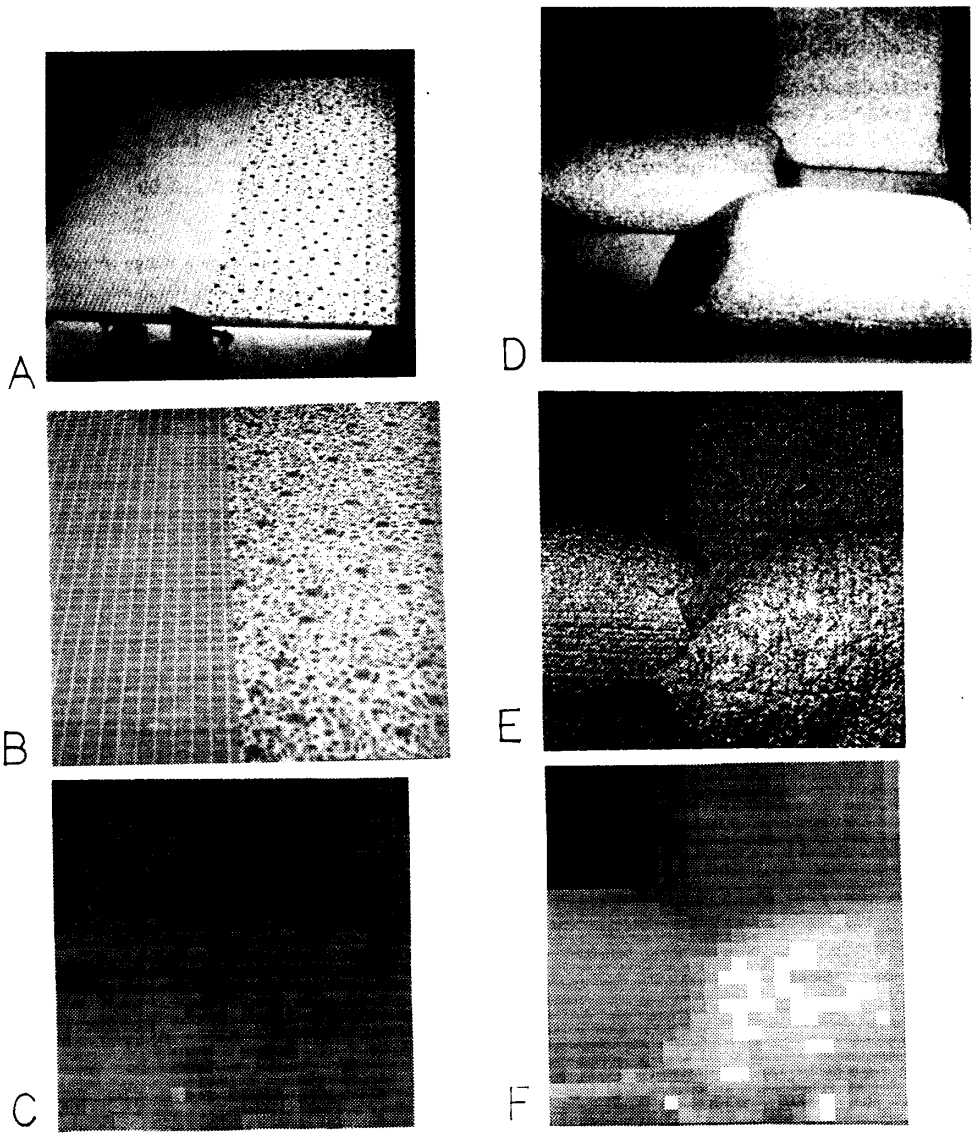


Figure 4: (a) A flat, textured plane, (b) its intensity image, (c) the range image, (d) a pile of pillows, (e) their intensity image, (f) the range image.

tions are provided by a MaxMux board. The total list price of this equipment (including power supplies, card cage, etc.) is approximately $23,500. The processing is controlled by a Sun 3/260. Cost of the optical elements was approximately $400.

3.1 Integrating Information over Scale

If we perform these calculations using a Laplacian pyramid rather than using the Laplacian of a single image (as shown in Figure 3(b)) then we will have measurements of Fourier power at several frequencies λ and can produce a more accurate and reliable estimate of range. The extra information can be utilized by reformulating Equation 7 as a regression equation in λ^2, e.g., as $A\lambda_i^2 + B = C_i$ where

$$A = 2\pi^2(\sigma_2^2 - \sigma_1^2) \quad B = \ln\frac{\sigma_2^2}{\sigma_1^2} \quad C_i = \ln \mathcal{F}_1(\lambda_i) - \ln \mathcal{F}_2(\lambda_i) \tag{11}$$

The solution to this quadratic regression equation is straightforward,

$$A = \frac{\sum_i(\lambda_i^2 - \bar{\lambda}^2)C_i}{\sum_i(\lambda_i^2 - \bar{\lambda}^2)^2} \tag{12}$$

where $\bar{\lambda}$ is the mean of the λ_i, giving a maximum-likelihood estimate of A. As $\sigma_1 \approx 0$, we have that

$$\sigma_2 = \sqrt{\frac{A}{2\pi^2}} \tag{13}$$

and thus, using Equation 1, absolute distance to the imaged surface patch:

$$D = \frac{Fv_0}{v_0 - F - \sigma_2 f_2} \tag{14}$$

where f_2 is the f-number of the second camera.

Our experience shows that this multiscale approach can more than double the range camera's accuracy, resulting in a measured standard error of 2.5%. Because processing in the Datacube is not tied to the frame rate, this multiscale approach to estimating range requires only about twice the processing time of the single scale implementation. No additional hardware is required to generalize the range camera to use multiscale processing.

3.1.1 Error Conditions

When there is insufficient high-frequency information in the image patch to enable the change in focus to be calculated this technique can produce errors. We currently solve this problem by introducing a threshold to remove all low-energy points, at the cost of missing some valid points.

3.2 Examples

Figures 4(a) and (b) show our calibration target (a flat, evenly textured plane), and Figure 4(c) shows the resulting range image. Figures 4(d) and (e) show a spatially-varying scene (an arrangement of pillows), and Figure 4(f) shows the resulting range image. It can be seen that reasonable accuracy achieved in both cases.

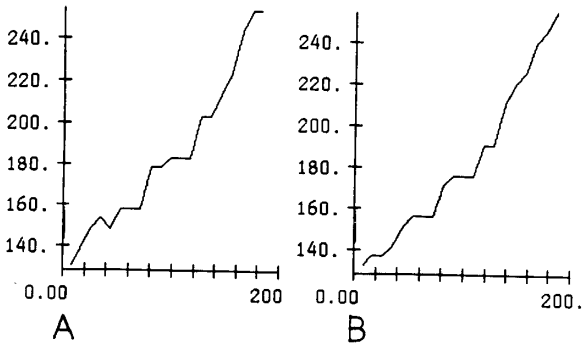


Figure 5: (a) A slice through a range image produced using the single scale technique, (b) A slice through a range image produced using the multiple scale technique.

Using the single-scale technique such range images can be produced up to eight times per second using our current equipment; with an additional VFir II convolution board range could be obtained thirty times per second. Using the multiscale technique processing time is almost doubled on our current hardware, but again with additional hardware range could be obtained thirty times per second.

Using the single scale technique we are currently achieving a standard error of 6% over a one cubic meter workspace, as measured using the calibration target shown in Figure 4(a). Because the single scale convolution kernel is 8 x 8, with most of the support in the central 4 x 4 region, there are roughly 128 x 128 completely independent measurements per image. An example of a vertical slice through a single-scale range image of the calibration target is shown in Figure 5(a); ideally the curve would be a straight line.

Using the multiscale technique we are currently achieving a standard error of 2.5 percent over a one cubic meter workspace, again measured using the calibration target shown in Figure 4(a). In the multiscale technique the convolution kernels become fairly large, so that there there are only roughly 32 x 32 completely independent range measurements per image. An example of a vertical slice through a multiple-scale range image of the calibration target is shown in Figure 5(b); ideally the curve would be a straight line.

4 Discussion

We have described an inexpensive implementation of the imaging range technique proposed in [7]. The major problems we have experienced in implementing the system were

mainly mechanical: aligning the cameras, setting the camera iris to equalize image brightness, and so forth. The major source of error was non-linearities in the cameras, primarily blooming due to specular reflections. Despite these problems, our experience shows that this ranging technique can be both economical and practical for tasks which require quick and reliable but coarse estimates of range. Examples of such tasks are initial target acquisition or obtaining the initial coarse estimate of stereo disparity in a coarse-to-fine stereo algorithm.

REFERENCES

[1] Jarvis, R. A., (1983) A perspective on range-finding techniques for computer vision, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, March 1983, pp. 122-139.

[2] Krotkov, E., (1986) Focusing, MS-CIS-86-22, Grasp Lab Technical Report No. 63, Dept. Computer and Information Science, University of Pennsylvania.

[3] Darrell, T. (1988) Pyramid based depth from focus, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 504-509, June 5-9, Ann Arbor, MI.

[4] Pentland, A., (1982) Depth of scene from depth of field, *Proceedings, Image Understanding Workshop*, September, 1982, Palo Alto, CA.

[5] Pentland, A., (1985) The Focal Gradient: Optics Ecologically Salient, *Investigative Opthomology and Visual Science*, Vol 26, No. 3, pp. 243, March 1985.

[6] Pentland, A., (1985) A new sense for depth of field, *Internaltional Joint Conference on Artificial Intelligence*, pp. 988-994, August, 1985, Los Angeles, CA.

[7] Pentland, A., (1987) A New Sense for Depth of Field, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 9, No. 4, July 1987, pp. 523-531.

[8] Subbarao, M. (1987) Direct recovery of depth-map, *IEEE Workshop on Computer Vision*, pp. 58-65, Miami Beach, FL.

[9] Grossman, P. (1987) Depth from focus, *Pattern Recognition Letters*, Vol. 5, No. 1, pp. 63-69.

[10] Subbarao, M. (1988) Depth recovery from blurred edges, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 498-503, June 5-9, Ann Arbor, MI.