Minimal Operator Set for Passive Depth from Defocus [*]

Masahiro Watanabe[†] **and Shree K. Nayar**[‡]

[†]Production Engineering Research Lab., Hitachi Ltd., Totsuka, Japan
[‡]Department of Computer Science, Columbia University, New York, USA

Abstract

A fundamental problem in depth from defocus is the measurement of relative defocus between images. We propose a class of broadband operators that, when used together, provide invariance to scene texture and produce accurate and dense depth maps. Since the operators are broadband, a small number of them are sufficient for depth estimation of scenes with complex textural properties. Experiments are conducted on both synthetic and real scenes to evaluate the performance of the proposed operators. The depth detection gain error is less than 1%, irrespective of texture frequency. Depth accuracy is found to be 0.5 ~ 1.2% of the distance of the object from the imaging optics.

1 Introduction

The depth from defocus method uses the relative defocus in two images taken with different optical settings to determine three-dimensional scene structure [Pentland-1987, Subbarao-1988, Ens and Lawrence-1991, Bove, Jr.-1993, Subbarao and Surya-1994, Nayar *et al.*-1995, Xiong and Shafer-1995]. The focus level in the two images can be varied by changing the focus setting of the lens, by moving the image sensor with respect to the lens, or by changing the aperture size. Depth from defocus is not confronted with the missing part and correspondence problems faced by stereo and structure from motion. This makes it an attractive prospect for structure estimation.

Despite these merits, at this point in time, fast, accurate, and dense depth from defocus has only been demonstrated using active illumination that constrains the dominant frequencies of the scene texture [Nayar *et al.*-1995, Watanabe *et al.*-1995]. Past investigations of *passive* depth from defocus indicate that it can prove computationally expensive to obtain a reliable depth map. This is because the frequency characteristics of scene textures are, to a large extent, unpredictable. Furthermore, the texture itself can vary dramatically over the image. Since the response of the defocus (blur) function varies with texture frequency, a single broadband filter that produces an aggregate estimate of defocus for an unknown texture cannot lead to accurate depth estimates. One solution is to use an large bank of narrow-band filters and compute depth in a least-squares sense using all dominant frequencies of the texture [Xiong and Shafer-1995, Gokstorp-1994]. However, this requires one to forego computational efficiency.

Subbarao and Surya [Subbarao and Surya-1994] proposed the *S-Transform* and applied it to depth from defocus. They modeled the image as a third-order polynomial in spatial domain, and arrived at a simple and elegant expression [Subbarao and Surya-1994]:

$$i_2(x,y)-i_1(x,y) \;=\; \frac{1}{4}({\sigma_2}^2-{\sigma_1}^2)\,\nabla^2\Big(\frac{i_2(x,y)+i_1(x,y)}{2}\Big) \tag{1}$$

where, i_1 and i_2 are the far and near focused images, respectively. The blur circle diameters in images i_1 and i_2 are expressed by their second central moments ${\sigma_2}^2$ and ${\sigma_1}^2$, respectively. Since an additional relation between σ_2 and σ_1 can be obtained from the focus settings used for the two images, σ_2 and σ_1 can be solved for and mapped to a depth estimate. As we see no terms that depend on scene frequency in equation (1), this can be considered to be a sort of texture-frequency invariant depth from defocus method. It produces reasonable depth estimates for large planar surfaces in the scene. However, it does not yield depth maps with high spatial resolution that are needed when depth variations in the scene are significant. We argue that this requires a more detailed analysis of image formation as well as the design of new filters based on frequency analysis.

In this paper, we propose a small set of filters, or operators, for passive depth from defocus. These operators, when used in conjunction, yield invariance to texture frequency while computing depth. The underlying idea is to precisely model relative image blur in frequency domain and express this model as a rational function of two linear combinations of basis functions. This rational expression leads us to a texture-invariant set of operators. The outputs of the operators are used as coefficients in a depth recovery equation that is solved to get a depth estimate. The attractive feature of this approach is that it uses only a small number of broadband linear operators with small kernel supports. Consequently, depth maps are computed not only with high efficiency and accuracy but also with high spatial resolution. Since our operators are derived using a rational expression to model relative image blur, they are referred to as *ratio-*

[*]This research was conducted at the Center for Research in Intelligent Systems, Department of Computer Science, Columbia University. It was supported in part by the Production Engineering Research Laboratory, Hitachi, and in part by the David and Lucile Packard Fellowship. The authors thank Yasuo Nakagawa of Hitachi Ltd. for his support and encouragement of this work.

1063-6919/96 $5.00 © 1996 IEEE

nal operators. Rational operators are general, in that, they can be derived for any blur model. In [Watanabe and Nayar-1995], we have shown how the outputs of rational operators can be used to derive a *depth confidence* measure that, in turn, can be used to enhance computed depth maps.

2 Depth From Defocus

Fundamental to depth from defocus is the relationship between focused and defocused images[Born and Wolf-1965]. Figure 1 shows the basic image formation geometry. All light rays that are radiated by object point P and pass the aperture A are refracted by the lens to converge at point Q on the image plane. The relationship between the object distance d, focal length of the lens f, and the image distance d_i is given by the lens law:

$$\frac{1}{d} + \frac{1}{d_i} = \frac{1}{f}. \qquad (2)$$

Each point on the object plane is projected onto a single point on the image plane, causing a clear or *focused* image i_f to be formed. If, however, the sensor plane does not coincide with the image plane and is displaced from it, the energy received from P by the lens is distributed over a patch on the sensor plane. The result is a blurred image of P.

It is clear that a single image does not include sufficient information for depth estimation, as two different scenes defocused to different degrees could produce identical images. A solution to the depth estimation problem is achieved by using two images, i_1 and i_2, separated by a known physical distance $2e$ [Ens and Lawrence-1991, Subbarao and Surya-1994]. The distance γ of the image i_1 from the lens should also be known. Given the above described setting, the problem is reduced to analyzing the relative blurring of each scene point in the two images and computing the position of its focused image. A restriction here is that the focused images of all of the scene points must lie between the *far-focused* sensor plane i_1 and the *near-focused* sensor plane i_2. For ease of description, we introduce the *normalized depth* α, which equals -1 at i_1 and 1 at i_2. Then, using $d_i = \gamma + (1+\alpha)e$ in the lens law (2), we obtain the depth d of the scene point.

2.1 Defocus Function

In Figure 1, $(1 \pm \alpha)e$ is the distance between the focused image of a scene point and its defocused image formed on the sensor plane. The light energy radiated by the scene point and collected by the imaging optics is uniformly distributed on the sensor plane over a circular patch with a radius of $(1 \pm \alpha)e\, a/d_i$[1]. This distribution,

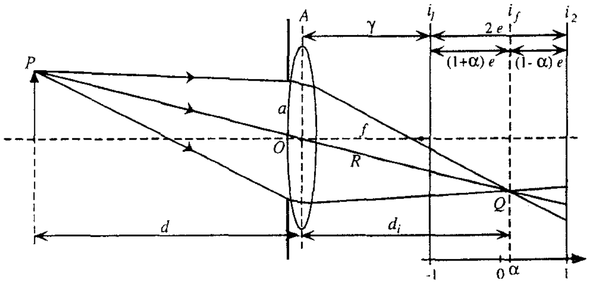


Figure 1: Image formation and depth from defocus. The two images, i_1 and i_2, include all the information required to recover scene structure between the focused planes in the scene corresponding to the two images.

also called the *pillbox*, is the defocus function:

$$h(x,y;(1\pm\alpha)e,F_e) = \frac{4{F_e}^2}{\pi(1\pm\alpha)^2e^2}\Pi\left(\frac{F_e}{(1\pm\alpha)e}\sqrt{x^2+y^2}\right) \qquad (3)$$

where, $+$ is used for image i_1, $-$ is used for image i_2, and $\Pi(r)$ is the rectangular function which takes the value 1 for $|r| < \frac{1}{2}$ and 0 otherwise. F_e is the effective *F-number* of the optics. In the optical system shown in Figure 1, F_e equals $d_i/2a$. In order to eliminate magnification differences between the near and far focused images, we have used *telecentric optics*, which is described in [Watanabe and Nayar-1996]. In the telecentric case, F_e equals $f/2a'$.

In Fourier domain, the defocus function in (3) is:

$$H(u,v;(1\pm\alpha)e,F_e) = \qquad (4)$$
$$\frac{2F_e}{\pi(1\pm\alpha)e\sqrt{u^2+v^2}} J_1\left(\frac{\pi(1\pm\alpha)e}{F_e}\sqrt{u^2+v^2}\right)$$

where, J_1 is the first-order Bessel function of the first kind, and u and v denote spatial frequency parameters in the x and y directions, respectively[2]. The effect of defocus in spatial and frequency domains can be written as:

$$\begin{aligned} i_2(x,y) &= i_f(x,y) * h(x,y;(1-\alpha)e,F_e)\,, \\ I_2(u,v) &= I_f(u,v)\cdot H(u,v;(1-\alpha)e,F_e)\,. \end{aligned} \qquad (5)$$

and

$$\begin{aligned} i_1(x,y) &= i_f(x,y) * h(x,y;(1+\alpha)e,F_e)\,, \\ I_1(u,v) &= I_f(u,v)\cdot H(u,v;(1+\alpha)e,F_e)\,. \end{aligned} \qquad (6)$$

[1] This geometric model is valid as far as the image is not exactly focused, in which case, a wave optics model is needed to describe the point spread function. Further, it is assumed that lens induced aberrations are small compared to the radius of the blur circle [Born and Wolf-1965].

[2] In the past, most investigators have used the Gaussian model instead of the pillbox model for the blur function. This is mainly to facilitate mathematical manipulations; the Fourier transform of a Gaussian function is also a Gaussian which can be converted into a quadratic function by using the logarithm. As we will see, in our approach, any form of blur function can be used.

Since α can vary from point to point in the image, strictly speaking, we have a *space-variant* system that cannot be expressed as a convolution. Therefore, equation (5) does not hold in a rigorous sense. However, if we assume that α is constant in a small patch around each pixel, equation (5) remains valid within the small patch.

2.2 Depth from Two Images

We now introduce the *normalized ratio*, $\frac{M}{P}(u,v;\alpha)$, where, $M(u,v) = I_2(u,v) - I_1(u,v)$ and $P(u,v) = I_2(u,v) + I_1(u,v)$. Equivalently, in the spatial domain, we have $m(x,y) = i_2(x,y) - i_1(x,y)$ and $p(x,y) = i_2(x,y) + i_1(x,y)$. Since the spectrum $I_f(u,v)$ of the focused image, which appears in equations (5) and (6), gets cancelled, the above normalized ratio is simply:

$$\frac{M(u,v;\alpha)}{P(u,v;\alpha)} = \frac{H(u,v;(1-\alpha)e,F_e) - H(u,v;(1+\alpha)e,F_e)}{H(u,v;(1-\alpha)e,F_e) + H(u,v;(1+\alpha)e,F_e)} \tag{7}$$

Figure 2 shows the relationship between the normalized image ratio M/P and the normalized depth α for several spatial frequencies. It is seen that $M/$P is a monotonic function of α for $-1 \le \alpha \le 1$, provided the radial frequency $f_r = \sqrt{u^2+v^2}$ is not too large. As a rule of thumb, this frequency range equals the width of the main lobe of the defocus function H when it is maximally defocused, i.e. when the distance between the focused image i_f and the sensor plane is $2e$. From the zero-crossing of the defocus function, the highest frequency below which the normalized image ratio $M/$P is monotonic is found to be:

$$f_r \le 0.61\,\frac{F_e}{e}\,. \tag{8}$$

For any given frequency within the above bound, since $M/$P is a monotonic function of α, $M/$P can be unambiguously mapped to a depth estimate β, as shown in Figure 2.

Besides serving a critical role in our development, Figure 2 also gives us new way of viewing previous approaches to depth from defocus: If one can by some method determine the amplitudes, I_1 and I_2 of the spectra of the two defocused images at a predefined radial frequency $f_{r0} = \sqrt{u_0{}^2 + v_0{}^2}$, a unique depth estimate can be obtained. This is the basic idea that most of the previous work is based on [Pentland-1987, Gokstorp-1994, Xiong and Shafer-1995], although the ratio used in the past is simply I_1/I_2 rather than the normalized ratio M/P introduced here.

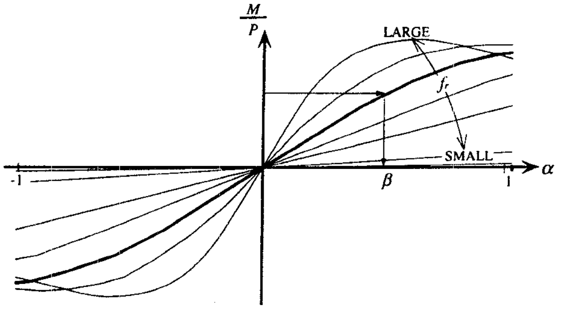


Figure 2: Relation between the normalized image ratio M/P and the defocus parameter α. An upper frequency bound can be determined, below which, M/P is a monotonic function of the defocus parameter α. For any given frequency within this bound, M/P can be unambiguously mapped to a depth estimate β.

3 Relative Defocus as a Rational Expression

We have established the monotonic response of the normalized image ratio M/P to the normalized depth (or defocus) α over all frequencies (see equation (7) and Figure 2). Our objective here is to model this relation in closed form. In doing so, we would like the model to be precise and yet lead us to a small number of linear operators for depth recovery. To this end, we model the function M/P by a rational expression of two linear combinations of basis functions:

$$\frac{M(u,v;\alpha)}{P(u,v;\alpha)} = \frac{\sum_{i=1}^{n_P} G_{Pi}(u,v)\,b_{Pi}(\alpha)}{\sum_{i=1}^{n_M} G_{Mi}(u,v)\,b_{Mi}(\alpha)} + \varepsilon(u,v,\alpha)\,, \tag{9}$$

where, $b_{Pi}(\alpha)$ ($i = 1..n_P$) and $b_{Mi}(\alpha)$ ($i = 1..n_M$) are the basis functions, $G_{Pi}(u,v)$ and $G_{Mi}(u,v)$ are the coefficients which are functions of frequency (u,v), and $\varepsilon(u,v,\alpha)$ is the residual error of the fit of the model to the function M/P. If the model is accurate, the residual error is negligible, and it becomes possible to use the model to map the normalized image ratio M/P to the normalized depth α. The above expression can be rewritten as:

$$\frac{M(u,v;\alpha)}{P(u,v;\alpha)} = \frac{\sum_{i=1}^{n_P} G_{Pi}(u,v)\,b_{Pi}(\beta)}{\sum_{i=1}^{n_M} G_{Mi}(u,v)\,b_{Mi}(\beta)} = R(\beta;u,v)\,. \tag{10}$$

Here, α on the left hand side represents the *actual depth* of the scene point while β on the right is the *estimated depth.* A difference between the two can arise only when the residual error is non-zero. If the normalized ratio on the left side is given to us for any frequency (u, v), we can obtain the depth estimate β by solving equation (10).

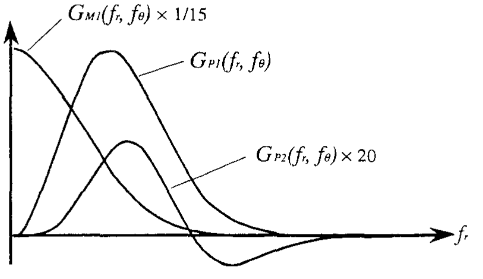


Figure 3: An example set of the coefficient functions obtained by fitting the polynomial model to the normalized image ratio M/P. Here, $G_{P1}(u, v)$ was chosen and the remaining two functions determined from the fit.

The above model for the normalized image ratio is general. In principle, any basis that captures the monotonicity and structure of the normalized ratio can be used. To be specific in our discussion, we use the basis we have chosen in our implementation. Since the response of M/P to α is odd-symmetric and is almost linear for small radial frequencies f_r (see Figure 2), we could model the response using three basis functions that are powers of β:

$$n_P = 2, \; n_M = 1, \; b_{P1}(\beta) = \beta, \; b_{P2}(\beta) = \beta^3, \; b_{M1}(\beta) = 1. \tag{11}$$

Then, equation (10) becomes[3]:

$$\frac{M(u, v; \alpha)}{P(u, v; \alpha)} = \frac{G_{P1}(u, v)}{G_{M1}(u, v)}\beta + \frac{G_{P2}(u, v)}{G_{M1}(u, v)}\beta^3 = R(\beta; u, v). \tag{12}$$

The term including β^3 can be seen as a small correction that compensates for the discrepancy of M/P from a linear model. From the previous section, we know that the blurring model completely determines M/P for any given depth α and frequency (u, v). The above polynomial model, $R(\beta; u, v)$, can therefore be fit to the theoretical M/P in equation (7) by assuming β to be α. This gives us the unknown ratios G_{P1}/G_{M1} and G_{P2}/G_{M1} as functions of frequency (u, v). In the case of a rotationally symmetric blurring model, such as the pillbox function, these ratios reduce to functions of just the radial frequency f_r.

Now, if we fix any one of the coefficient functions, say, $G_{P1}(u, v)$, all the other coefficients can be determined from the ratios[4]. Therefore, it is possible to determine all the coefficient functions that ensure that the above polynomial model accurately fits the normalized image ratio M/P given by equation (7). Figure 3 shows an example set (based on an arbitrary selection of $G_{P1}(u, v)$) of the coefficient functions, G_{P1}, G_{P2} and G_{M1}, for the case of the pillbox blur model.

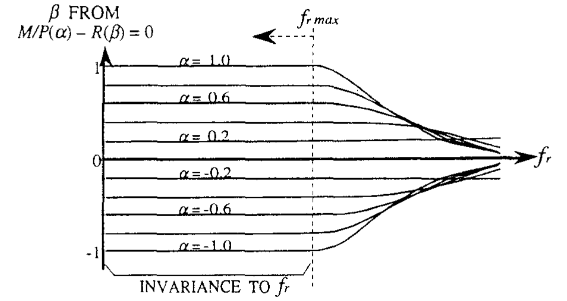


Figure 4: Depth β, estimated using the polynomial model in equation (12), is plotted as a function of spatial frequency for different values of actual depth α. We see that the estimated depth equals the actual depth and is invariant to frequencies within the upper bound $f_{r\,max}$ given by equation (13).

We now examine how well the polynomial model fits the plots in Figure 2 of the normalized ratio $\frac{M}{P}(u, v, \alpha)$. More precisely, we are interested in knowing how well the model can used to estimate depth. To this end, for each frequency, we select a "true" depth value α and find the corresponding ratio M/P using the analytical expression in (7). This ratio is then plugged into the polynomial model of (12) to calculate the depth estimate β using the Newton-Raphson method. This process is repeated for all frequencies.

Figure 4 shows that the estimated depth β is, for all practical purposes, equal to the actual depth α, indicating that the polynomial model is indeed accurate. Further, the estimated depth is invariant (insensitive) to texture frequency as far as the radial frequency f_r is below $f_{r\,max}$. Above this frequency limit $f_{r\,max}$, the response of $\frac{M}{P}(u, v; \alpha)$ to α shown in Figure 2, becomes non-monotonic within the region $-1 \leq \alpha \leq 1$ and hence an accurate depth estimate is not obtainable. In practice, any image can be convolved using a passband filter to ensure that all frequencies above $f_{r\,max}$ are removed. The rule of thumb used to determine $f_{r\,max}$ is given by equation (8). However, for the pillbox blur model, we have found via numerical simulation that $f_{r\,max}$ is in fact 1.2 times larger[5] than the limit given by equation (8).

$$f_{r\,max} = 1.2 \cdot 0.61 \frac{F_e}{e} = 0.73 \frac{F_e}{e}. \tag{13}$$

[3] We found that replacing $b_{P2}(\beta)$ by $(\beta - \frac{1}{a} \tanh a\beta)$ gives us a slightly better fit when the defocus model is the pillbox function. Yet, to reduce the computational cost of solving equation (10) for depth β, we have chosen this simple polynomial model.

[4] In practice, $G_{P1}(u, v)$ cannot be selected in an entirely arbitrary fashion. Other restrictions that need to be considered are detailed in [Watanabe and Nayar-1995].

[5] This number can be increased from 1.2 to 1.3 if a larger number of Newton-Raphson iterations are used. However, depth results in this additional range are not numerically stable in the presence of noise since the response curves of M/P tend to flatten out. Hence, we use only one iteration.

This is a valuable side-effect of introducing the normalized image ratio M/P; we can utilize 20% more frequency spectrum information than conventional methods which use the ratio I_1/I_2.

4 Rational Operators

We have introduced a rational expression model for the normalized ratio M/P and shown that the solution of equation (10) gives us robust depth estimates for all frequencies within a permissible range. Thus far, this robustness was demonstrated for individual frequencies. In this section, we show how the rational model can be used to design a small set of broadband operators that can handle arbitrary textures.

Taking cross-products in equation (10), we get:

$$\begin{aligned}&\sum_{i=1}^{n_M} M(u,v;\alpha)\,G_{Mi}(u,v)\,b_{Mi}(\beta) = \\ &\sum_{i=1}^{n_P} P(u,v;\alpha)\,G_{Pi}(u,v)\,b_{Pi}(\beta)\,. \qquad (14)\end{aligned}$$

By integrating over the entire frequency space, we get:

$$\sum_{i=1}^{n_M} c_{Mi}(\alpha)\,b_{Mi}(\beta) = \sum_{i=1}^{n_M} c_{Pi}(\alpha)\,b_{Pi}(\beta)\,, \qquad (15)$$

where:

$$\begin{aligned} c_{Mi}(\alpha) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} M(u,v;\alpha)\,G_{Mi}(u,v)\,du\,dv \\ c_{Pi}(\alpha) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(u,v;\alpha)\;G_{Pi}(u,v)\;du\,dv \qquad (16)\end{aligned}$$

Here, we invoke the power theorem [Bracewell-1965]:

$$\begin{aligned}&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(u,v)\,G(u,v)\,du\,dv = \\ &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,g(-x,-y)\,dx\,dy\,, \qquad (17)\end{aligned}$$

where, $F(u,v)$ and $G(u,v)$ are the Fourier transforms of functions $f(x,y)$ and $g(x,y)$, respectively. Since we are conducting a spatial-frequency analysis, that is, we are analyzing the frequency content in a small area centered around each pixel, the right hand side of equation (16) is nothing but a convolution. This implies that $c_{Mi}(\alpha)$ and $c_{Pi}(\alpha)$ are actually functions of (x,y) and can be determined by convolutions as:

$$\begin{aligned}&c_{Mi}(x,y;\alpha) \\ =\; &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} m(x',y';\alpha)\,g_{Mi}(x-x',y-y')\,dx'\,dy'\,, \\ &c_{Pi}(x,y;\alpha) \\ =\; &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x',y';\alpha)\;g_{Pi}(x-x',y-y')\;dx'\,dy' \qquad (18)\end{aligned}$$

where, $g_{Mi}(x,y)$ and $g_{Pi}(x,y)$ are the inverse Fourier transforms of $G_{Mi}(u,v)$ and $G_{Pi}(u,v)$, respectively. In short, all the coefficients needed to compute depth using the polynomial in equation (15) can be determined by convolving the difference image $m(x,y)$ and the summed image $p(x,y)$ with linear operators that are spatial domain equivalents of the coefficient functions. We refer these as *rational operators*. The outputs of these operators at each pixel (x,y) are plugged into equation (15) to determine depth $\beta(x,y)$.

As an example, if we use the model in equation (12), the depth recovery equation (15) becomes:

$$c_{M1}(x,y;\alpha) = c_{P1}(x,y;\alpha)\,\beta + c_{P2}(x,y;\alpha)\,\beta^3\,. \quad (19)$$

By substituting equation (18), we have:

$$\begin{aligned} g_{M1}(x,y) * m(x,y;\alpha) \quad &= \\ g_{P1}(x,y) * p(x,y;\alpha)\,\beta \quad + \quad & g_{P2}(x,y) * p(x,y;\alpha)\,\beta^3 \quad (20)\end{aligned}$$

Again, the above three rational operators are nothing but inverse Fourier transforms of the coefficient functions shown in Figure 3. We see that, though the operators are all broadband (see Figure 3), the above recovery equation is independent of scene texture and provides an efficient means of computing precise depth estimates.

4.1 Design of Discrete Rational Operators

As stated earlier, in principle, one of the three rational filters, say, $g_{P1}(x,y)$, can be chosen and then the remaining two filters $g_{M1}(x,y)$ and $g_{P2}(x,y)$ can be derived. However, since we are interested in high accuracy, several factors need to be considered during the design of the filters. These include: (a) the use of a prefilter to remove the DC component and high frequencies above $f_{r\,max}$ in the two images, (b) the best choice for $g_{P1}(x,y)$, and (c) the design of discrete filters with small support that have the desired frequency characteristics. For lack of space, we refer the interested reader to [Watanabe and Nayar-1995] for details of the design of discrete rational operator. Here, we present an example operator set that we have used in our experiments.

Figures 5 and 6 show the kernels and their frequency responses for the three rational operators and the prefilter, derived with kernel size set to 7×7 and $e/F_e =$ 2.307 pixels. Since the discrete Fourier transform of a kernel of size k_s has the minimum discrete frequency period of $1/k_s$, it is difficult to obtain precisely any response in the frequency region below $1/k_s$. Further, the spectrum in this region is going to be suppressed by the prefilter as it is close to the DC component. Therefore, the maximum frequency $f_{r\,max}$ must be well above $1/k_s$. We express this condition as $f_{r\,max} \geq 2\frac{1}{k_s}$. Using equation (13), we obtain:

$$\frac{2e}{F_e} \leq 0.73k_s\,. \qquad (21)$$

This condition can be interpreted as follows: The maximum blur circle diameter $2e/F_e$ must be smaller than 73% of the kernel size k_s. This is also intuitively reasonable as the kernel should be larger than the blur circle as it seeks to measure blur[6].

$$g_{M1} = \begin{pmatrix} -0.001 & 0.045 & 0.179 & 0.297 & 0.179 & 0.045 & -0.001 \\ 0.045 & 0.400 & 0.868 & 1.093 & 0.868 & 0.400 & 0.045 \\ 0.179 & 0.868 & 2.957 & 4.077 & 2.957 & 0.868 & 0.179 \\ 0.297 & 1.093 & 4.077 & 6.005 & 4.077 & 1.093 & 0.297 \\ 0.179 & 0.868 & 2.957 & 4.077 & 2.957 & 0.868 & 0.179 \\ 0.045 & 0.400 & 0.868 & 1.093 & 0.868 & 0.400 & 0.045 \\ -0.001 & 0.045 & 0.179 & 0.297 & 0.179 & 0.045 & -0.001 \end{pmatrix}$$

$$g_{P1} = \begin{pmatrix} -0.039 & -0.091 & -0.198 & -0.259 & -0.198 & -0.091 & -0.039 \\ -0.091 & -0.327 & -0.470 & -0.425 & -0.470 & -0.327 & -0.091 \\ -0.198 & -0.470 & 0.335 & 1.393 & 0.335 & -0.470 & -0.198 \\ -0.259 & -0.425 & 1.394 & 3.385 & 1.393 & -0.425 & -0.259 \\ -0.198 & -0.470 & 0.335 & 1.393 & 0.335 & -0.470 & -0.198 \\ -0.091 & -0.327 & -0.470 & -0.425 & -0.470 & -0.327 & -0.091 \\ -0.039 & -0.091 & -0.198 & -0.259 & -0.198 & -0.091 & -0.039 \end{pmatrix}$$

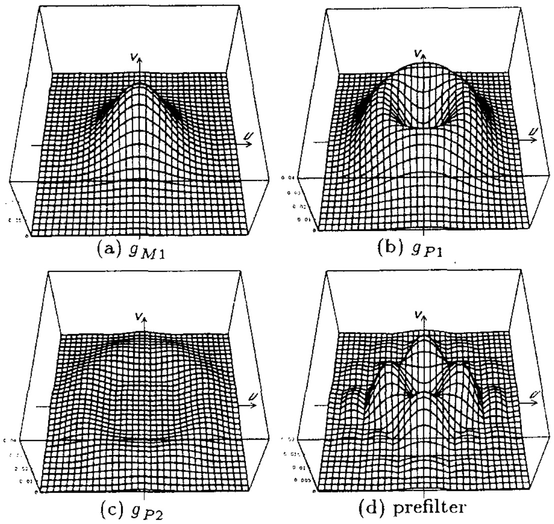
$$g_{P2} = \begin{pmatrix} 0.056 & -0.020 & -0.068 & -0.061 & -0.068 & -0.020 & 0.056 \\ -0.020 & -0.068 & 0.059 & 0.145 & 0.059 & -0.068 & -0.020 \\ -0.068 & 0.059 & 0.176 & -0.019 & 0.176 & 0.059 & -0.068 \\ -0.061 & 0.145 & -0.019 & -0.698 & -0.019 & 0.145 & -0.061 \\ -0.068 & 0.059 & 0.176 & -0.019 & 0.176 & 0.059 & -0.068 \\ -0.020 & -0.068 & 0.059 & 0.145 & 0.059 & -0.068 & -0.020 \\ 0.056 & -0.020 & -0.068 & -0.061 & -0.068 & -0.020 & 0.056 \end{pmatrix}$$

$$prefilt = \begin{pmatrix} -0.143 & -0.198 & -0.105 & -0.071 & -0.105 & -0.198 & -0.143 \\ -0.198 & -0.192 & 0.017 & 0.072 & 0.017 & -0.192 & -0.198 \\ -0.105 & 0.017 & 0.284 & 0.460 & 0.284 & 0.017 & -0.105 \\ -0.071 & 0.072 & 0.460 & 0.644 & 0.460 & 0.072 & -0.071 \\ -0.105 & 0.017 & 0.284 & 0.460 & 0.284 & 0.017 & -0.105 \\ -0.198 & -0.192 & 0.017 & 0.072 & 0.017 & -0.192 & -0.198 \\ -0.143 & -0.198 & -0.105 & -0.071 & -0.105 & -0.198 & -0.143 \end{pmatrix}$$

Figure 5: Rational operator kernels derived using kernel size of 7×7 and $e/F_e = 2.307$ pixels. Regardless of the nature of scene texture, passive depth from defocus can be accomplished using this small operator set. In general, the number of rational operators and their kernels depend on the order of the rational expression used to model the normalized ratio M/P, the selected kernel size k_s, and the imaging optics used (e/F_e).

5 Algorithm

Figure 7 illustrates the flow of the depth from defocus algorithm we have implemented. The far and near focused images are first added and subtracted to produce $p(x, y)$ and $m(x, y)$, respectively. Then they are convolved with the prefilter and subsequently with the three rational operators. The resulting coefficient images are then smoothed by local averaging. The final step is the computation of depth from the coefficients using a single iteration of the Newton-Raphson method [Watanabe and Nayar-1995]. Alternatively, depth computation can be achieved using a precomputed two-dimensional look-up table. The look-up table is configured to take $c'_{M1}(x, y)/c'_{P1}(x, y)$ and $c'_{P2}(x, y)/c'_{P1}(x, y)$ as inputs and provides depth $\beta(x, y)$ as output. In summary, a depth map is generated with as few as 5 two-dimensional convolutions, simple smoothing of the coefficient images, and a straightforward depth computation step. The above operations can be executed efficiently using a pipelined image processor. If one uses Datacube's MV200 pipeline processor, all the computations can be realized using as few as 10 pipelines. The entire depth from defocus algorithm can then be executed in 0.16 msec for an image of size 512×480.



(a) g_{M1} (b) g_{P1} (c) g_{P2} (d) prefilter

Figure 6: Frequency responses of the rational operators shown in Figure 5.

6 Experiments

We first illustrate the linearity of depth estimation and its invariance to texture frequency using synthetic images. The synthetic images shown in Figure 8 correspond to a planar surface that is inclined away from the sensor such that its normalized depth value is 0 at the top and 255 at the bottom. The plane includes 10 vertical strips with different textural properties. The left 7 strips have textures with narrow power spectra whose central frequencies are 0.015, 0.03, 0.08, 0.13, 0.18, 0.25 and 0.35, from left to right. The 8^{th} strip is white noise. The next two strips are fractals with dimensions of 3 and 2.5, respectively [Peitgen and Saupe-1988]. The near and far focused images were generated using the pillbox blur model. The defocus condition used was $e/F_e = 2.307$ pixels. In all our experiments, the digital images used are of size 640×480. The depth map estimated using the 7×7 rational operators and 5×5 coefficient smoothing is shown as a gray-coded image in Figure 8(c) and a wireframe in Figure 8(d). As is evident, the proposed algorithm produces high accuracy despite the significant texture variations between the vertical strips. A detailed error analysis can be found in

[6] Since the above conditions related to kernel size are rough, we suggest that the linearity of depth estimation be checked (using synthetic images) to find the best kernel size k_s. Such an evaluation is reported in [Watanabe and Nayar-1995].

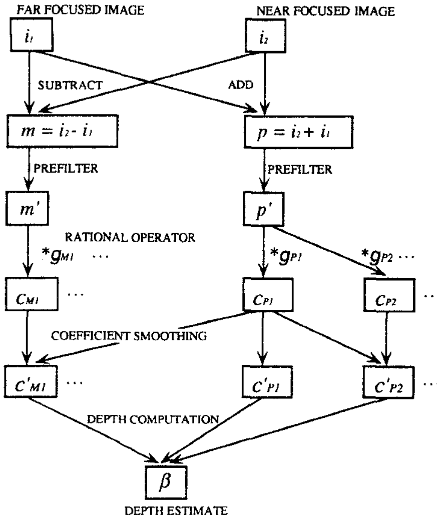


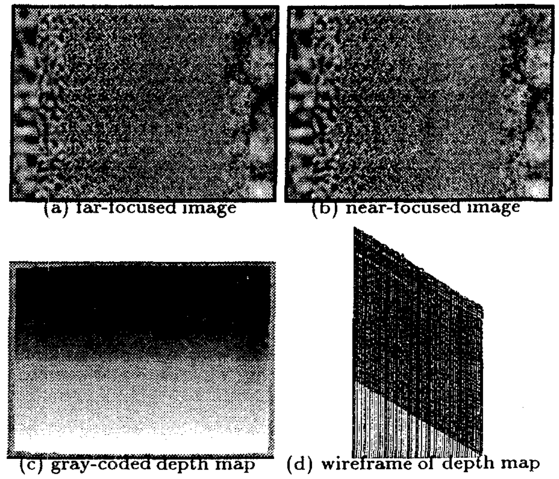
Figure 7: The flow of the depth from defocus algorithm. Using Datacube's MV200 pipeline processor, the entire algorithm can be executed in as little as 0.16 msec to obtain a 512×480 depth map.

[Watanabe and Nayar-1995].

Images of real scenes were taken using a SONY XC-77 monochrome camera. The lens used is a Cosmicar B1214D-2 with f=25mm. The lens was converted into a telecentric lens by using an additional aperture to make its magnification invariant to defocus (see [Watanabe and Nayar-1996]). As a result of telecentricity, image shifts between the far and near focused images are lower than 1/10 of a pixel. The lens aperture was set to F/8.3. The far-focused image i_1 was taken with the lens focused at 869mm from the camera, and the near-focused image i_2 with the lens focused at 529mm. These two distances were chosen so that all scene points lie between them. The above focus settings result in a maximum blur circle radius of $e/F_e = 2.307$ pixels. For each of the two focus settings, 256 images were averaged over 8.5 sec to get images with high signal-to-noise ratio.

Figure 9 shows results obtained for a scene that includes a variety of textures. Figure 9(a) and (b) are the far-focused and near-focused images, respectively. Figure 9(c) and (d) are the computed depth map and its wireframe plot. Depth maps of all the curved and planar surfaces are detected with high fidelity and high resolution without any post-filtering. After 9×9 median filtering, we get an even better depth map as shown in Figure 9(e).

The last experiment seeks to quantify the accuracy of depth estimation. The target used is a plane paper similar to the textured background in the scene in Figure 9. This plane is moved in steps of 25mm and a depth map of the plane .is computed for each position. The plot


(a) far-focused image (b) near-focused image

(c) gray-coded depth map (d) wireframe of depth map

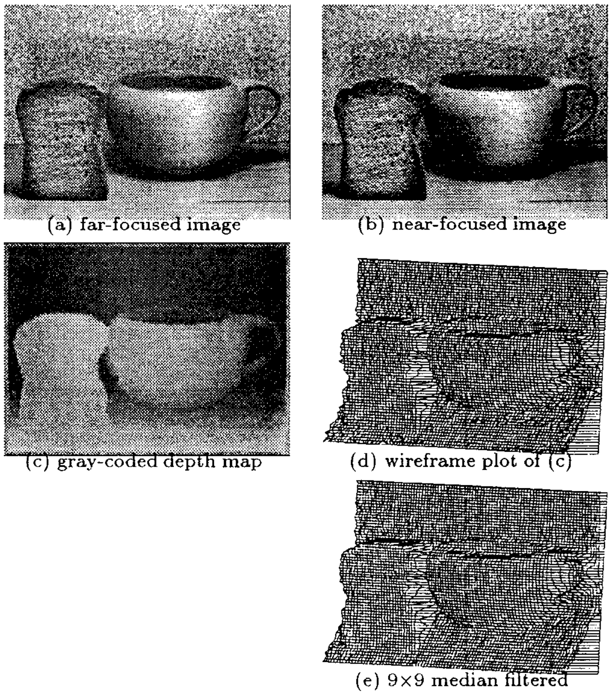
Figure 8: Depth from defocus applied to synthetic images of an the inclined plane is accurately recovered despite the significant texture variations.

in Figure 10 illustrates that the algorithm has excellent depth estimation linearity. The RMS error of a line fit to the measured depths is 4.2 mm. Depth values for a 50×50 area were used to estimate the RMS depth error for each position of the planar surface. In Figure 10 the RMS errors are plotted as $\pm\sigma$ error bars. The RMS error relative to object distance is seen to vary with object distance. It is 0.4% ~ 0.8% for close objects and 0.8% ~ 1.2% for objects farther than 880 mm. This is partly because of the mapping from the depth measured on the image side to depth on the object side.

7 Conclusions

We proposed the class of rational operators for passive depth from defocus. Though the operators are broadband, when used together, they provide invariance to scene texture. Since they are broadband, a small number of operators are sufficient to cover the entire frequency spectrum. Hence, rational operators can replace large filter banks that are expensive from a computational perspective. This advantage comes without the need to sacrifice depth estimation accuracy and resolution. We have detailed the procedure used to design rational operators. As an example, we constructed 7×7 operators using a polynomial model for the normalized image ratio. However, the notion of rational operators is more general and represents a complete class of filters. The design procedure described here can be used to construct operators based on other rational models for the normalized image ratio. Further, rational operators can be derived for any desired blur function.

The proposed depth from defocus algorithm requires



(a) far-focused image

(b) near-focused image

(c) gray-coded depth map

(d) wireframe plot of (c)

(e) 9×9 median filtered

Figure 9: The depth from defocus algorithm applied to a real scene with complex textures.

only a total of 5 convolutions. We tested the algorithm using both synthetic scenes and real scenes to evaluate performance. Depth accuracy was found to be 0.5 ~ 1.2% of object distance from the sensor. Given the efficiency of the algorithm, it is worth pursuing a real-time implementation using a pipeline image processing architecture such as the Datacube MV200. We estimate that such an algorithm would result in at least 6 depth maps per second of 512×480 resolution.

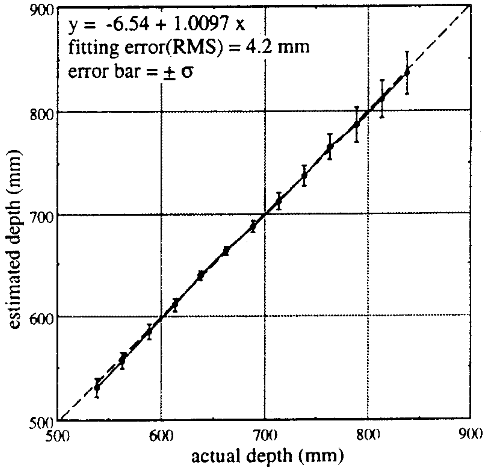


Figure 10: Depth estimation linearity for a textured plane. The plane is moved in increments of 25mm. away from the lens. All plotted distances are measured from the lens. The RMS error relative to object distance is 0.4% ~ 1.2%.

References

[Born and Wolf, 1965] M. Born and E. Wolf. *Principles of Optics.* London:Permagon, 1965.

[Bove, Jr., 1993] V. M. Bove, Jr. Entropy-based depth from focus. *Journal of Optical Society of America A*, 10:561–566, April 1993.

[Bracewell, 1965] R. N. Bracewell. *The Fourier Transform and Its Applications.* McGraw Hill, 1965.

[Ens and Lawrence, 1991] J. Ens and P. Lawrence. A matrix based method for determining depth from focus. *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 600–609, June 1991.

[Gokstorp, 1994] M. Gokstorp. Computing depth from out-of-focus blur using a local frequency representation. *Proc. on Intl. Conf. on Patt. Recog.*, October 1994.

[Nayar *et al.*, 1995] S. K. Nayar, M. Watanabe and M. Noguchi. Real-time focus range sensor. *Proc. of Intl. Conf. on Computer Vision*, pages 995–1001, June 1995.

[Peitgen and Saupe, 1988] H. O. Peitgen and D. Saupe, editors. *The Science of Fractal Images.* Springer-Verlag, New York, NY, 1988.

[Pentland, 1987] A. Pentland. A new sense for depth of field. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 9(4):523–531, July 1987.

[Subbarao and Surya, 1994] M. Subbarao and G. Surya. Depth from defocus: A spatial domain approach. *International Journal of Computer Vision*, 13(3):271–294, 1994.

[Subbarao, 1988] M. Subbarao. Parallel depth recovery by changing camera parameters. *Proc. of Intl. Conf. on Computer Vision*, pages 149–155, December 1988.

[Watanabe and Nayar, 1995] M. Watanabe and S. K. Nayar. Rational filters for passive depth from defocus. Technical Report CUCS-031-95, Dept. of Computer Science, Columbia University, New York, NY, USA, October 1995.

[Watanabe and Nayar, 1996] M. Watanabe and S. K. Nayar. Telecentric optics for computational vision. *Proc. of European Conference on Computer Vision (ECCV'96)*, April 1996.

[Watanabe *et al.*, 1995] M. Watanabe, S. K. Nayar and M. Noguchi. Real-time computation of depth from defocus. *Proc. of SPIE: Three-Dimensional and Unconventional Imaging for Industrial Inspection and Metrology*, 2599:A-03, November 1995.

[Xiong and Shafer, 1995] Y. Xiong and S. A. Shafer. Moment filters for high precision computation of focus and stereo. *Proc. of Intl. Conf. on Robotics and Automation*, pages 108–113, August 1995. Also, Technical Report CMU-RI-TR-94-28, Pittsburgh, PA, USA, September, 1994.